

UNIVERSITÀ DELLA CALABRIA



DOCTORAL THESIS

***Transformer-based Language Models:
Understanding Language across Domains and Dimensions***

PhD in
Information and Communication Technologies

CYCLE
XXXVII

Coordinator:

Prof. Giancarlo Fortino

Supervisors:

Prof. Andrea Tagarelli

Prof. Ester Zumpano



ANDREA
TAGARELLI
17.01.2025
19:05:25
GMT+02:00



Ester Zumpano
21.01.2025
11:27:24
GMT+01:00

Author: Candida Maria Greco

CANDIDA
MARIA
GRECO
17.01.2025
10:14:33
GMT+01:00



GIANCARLO
FORTINO
22.01.2025
22:19:45
GMT+02:00



Language is a process of free creation; its laws and principles are fixed, but the manner in which the principles of generation are used is free and infinitely varied. Even the interpretation and use of words involves a process of free creation. –Noam Chomsky

Preface

Transformer-based language models (TLMs) have widely been recognized to be a cutting-edge technology for the successful development of deep-learning-based solutions to problems and applications that require natural language processing and understanding. Their ability to learn a contextual language understanding model allows them to comprehensively encapsulate the semantics of language and navigate intricate non-linear relationships between terms. This includes the ability to discern subtle and complex lexical patterns like sequential structures and long-term dependencies.

TLMs have set new standards in domain-specific fields such as legal domain. These domains present unique challenges, including complex syntactic structures, specialized terminology, and issues related to privacy and data security. The scarcity of domain-specific data further complicates the task of training robust models in these fields.

Beyond domain-specific applications, certain tasks remain challenging even within the general domain. For instance, lexical entailment—determining relationships between words in terms of implication—poses difficulties that extend across contexts.

TLMs must also deal with understanding more abstract aspects of language, such as the psycho-social dynamics expressed in the texts, which go beyond syntax and semantics to reflect the intentions and ethics underlying human actions and thoughts.

Throughout the Ph.D. course, my research has focused on the following objectives:

- To provide a systematic overview of TLM-based methods for AI-driven problems and tasks in the legal sphere;
- To employ TLM-based methods to analyze texts of European constitutions that share common topics;
- To evaluate the ability of TLMs to recognize lexical entailment relationships between verbs;
- To fill the gap in the literature by providing a thorough exploration of methods using TLMs for text-based identification of mental health issues;
- To explore the relationship between moral values and TLMs, advancing their ability to capture and predict moral dimensions in text.

Contents

1	Introduction	1
2	Background	5
2.1	Transformer	5
2.2	Transformer-based Language Models	7
2.2.1	Encoder-only models	13
2.2.2	Encoder-decoder models	28
2.2.3	Decoder-only models	34
2.3	Problems and Tasks	42
2.3.1	Problems and tasks for the general-domain	42
2.3.2	Problems and tasks for the legal-domain	44
2.3.3	Problems and tasks for the psychosocial-domain	52
2.4	Data and Benchmarks	57
2.4.1	Data and benchmarks for the general-domain	57
2.4.2	Data and benchmarks for the legal-domain	66
2.4.3	Data and benchmarks for psychosocial-domain	87
PART I: AI-BASED NLP METHODS AND APPROACHES FOR THE LEGAL DOMAIN		95
3	Bringing order into the realm of Transformer-based Language Models for Artificial Intelligence and Law	97
3.1	Related Works	97
3.2	Methods for Legal Understanding Tasks	98
3.2.1	Task-adaptive methods	99
3.2.2	Enhanced methods for COLIEE tasks	106
3.2.3	Domain-adaptive pre-training methods	115
3.2.4	Dealing with long legal documents	121
3.2.5	Methods for non-English legal languages	131
3.2.6	Multilingual and cross-lingual methods	141
3.2.7	Dealing with explainability and interpretability issues	145

3.3	Methods for Legal Generation Tasks	148
3.4	Chapter Review	151
4	Understanding and Summarizing Rights and Duties in European Constitutions through Transformer-based Language Models	161
4.1	Related Works	162
4.2	Data	164
4.3	Methods for Constitutional Text Understanding	166
4.3.1	Methodology	166
4.3.2	Experimental Evaluation	167
4.4	Methods for Constitutional Text Knowledge and Summarization . . .	176
4.4.1	Methodology	176
4.4.2	Experimental Evaluation	180
4.5	Chapter Review	189
PART II: AI-BASED NLP METHODS AND APPROACHES FOR LEXICAL ENTAILMENT		193
5	Transformer-based Language Models for Lexical Entailment Recognition	195
5.1	Related Works	196
5.2	Wordnet and Hyperlex Data	197
5.3	Methodology	198
5.4	Experimental Evaluation	200
5.4.1	Model settings and deployment	200
5.4.2	Evaluation criteria	200
5.4.3	Results	201
5.4.4	Further experimental results	206
5.5	Chapter Review	212
PART III: AI-BASED NLP METHODS AND APPROACHES FOR PSYCHOSOCIAL DIMENSIONS		217
6	Transformer-based Language Models for Mental-Health-related Tasks	219
6.1	Methods	220
6.1.1	Non-domain-adapted methods	220
6.1.2	Domain-adapted models	224
6.1.3	Methods for CLEF eRisk competition	227
6.2	Chapter Review	229
7	Transformer-based Language Models for Moral-Foundation- Theory-related Tasks	231
7.1	Related Works	232
7.2	Methods	234
7.2.1	Moral-driven methods	240

7.2.2	Moral-targeted methods	244
7.3	Chapter Review	247
8	E2MoCase: A Dataset for Emotional, Event and Moral Observations in News Articles on High-impact Legal Cases	253
8.1	Related Works	254
8.2	Data Collection	255
8.2.1	Case selection	255
8.2.2	News retrieval	256
8.3	Dataset Design and Annotation	256
8.4	Dataset Analysis	258
8.5	Dataset Validation	264
8.6	Chapter Review	268
9	ME²-BERT: Are Events and Emotions what you need for Moral Value Prediction?	269
9.1	Related Works	271
9.2	Preliminaries	272
9.3	Methodology	273
9.4	Computational Complexity aspects	277
9.5	Experimental Evaluation	278
9.5.1	Experimental setting	279
9.5.2	Results	280
9.6	Chapter Review	286
	Conclusions	289
	References	293
	Appendix A	345
A.1	RQ1 best matchings	345
A.2	RQ2 performances	346
A.3	Statistics on term anonymization.	353
A.4	Experimental environment	354
	Appendix B	357
B.1	Zero-shot Prompts on WordNet data	357
B.2	Zero-shot Prompts on HyperLex data	358
B.3	Few-shot prompt selection strategy	359
B.4	Experimental environment	366
	Appendix C	367
C.1	Prompt example for case selection in E2MoCase	367
C.2	SwissDox Query	367
C.3	Selected cases in E2MoCase	367
C.4	Correlation Event in E2MoCase	369

C.5	Plutchik's Wheel of Emotions	370
C.6	Outlier Removal and Cleaning process in E2MoCase	370
C.7	ME ² -BERT Notations	373
C.8	Experimental environment	374

Chapter 1

Introduction

Since the advent of BERT, Transformer-based language models (TLMs) have shown outstanding effectiveness in several tasks of Natural Language Processing (NLP). Their ability to learn a contextual language understanding model allows for capturing language semantics and non-linear relationships between terms, including subtle and complex lexical patterns such as sequential structure and long-term dependencies, thus obtaining comprehensive local and global feature representations of a text, without the need for feature engineering.

Over time, TLMs have expanded beyond general-domain applications to domain-specific fields like the *legal domain*, where they have made significant advancements in processing complex legal texts. Domain-specific texts often have intricate sentence structures, making parsing and generating such content challenging. Additionally, these texts frequently contain sensitive information, requiring careful handling to ensure privacy and security. In this regard, the use of TLMs was investigated in the legal domain, where they have become a cutting-edge technology for addressing many AI-driven legal tasks. The rapid progress of models like BERT has significantly enhanced the ability of AI systems to process and understand complex legal texts. Existing works discuss tasks like judgment prediction, legal question answering, and information retrieval, or provide a structural framework of legal information retrieval approaches, like natural language based, ontology-based, and deep-learning-based systems. However, these works focus primarily on early TLM applications or general AI techniques and do not offer a systematic analysis of TLM-based approaches for specific legal problems. Within this view, this thesis provides a systematic overview of TLM-based methods in the legal domain, focusing on how these models support legal processes and identifying both the advancements and the current limitations that present opportunities for further research.

While most existing studies in legal AI has been developed on law codes as well as case law documents, *constitutions* have remained relatively under-explored, presenting a new challenge for TLM-based methods. To fill this gap in the literature, we extended the exploration of TLMs in the legal domain to the analysis of European constitutions, with a focus on rights and duties. More specifically, our study resorts to TLMs to perform lexical and semantic similarity analysis of European constitutions,

particularly regarding the topics that define and protect citizens' rights and duties. Current literature applies text mining techniques to analyze similarities between constitutions but did not incorporate deep learning or TLMs. We also overcome this limitation by leveraging contextualized TLMs models for a deep analysis of constitutional texts across European countries. The research in this field was further advanced with the investigation of TLMs' internal knowledge of European constitutions and the application of TLMs to summarize constitutional texts. Recent advances on the use of TLMs for legal texts, particularly in European law, have shown progress in tasks like multi-label classification and legal document summarization. However, limited work has been done on summarizing constitutional texts. We assessed the awareness of such models regarding topics of rights and duties covered in European constitutional texts and their effectiveness in producing coherent and informative summaries that faithfully capture the essence of these foundational documents.

Beyond specific domains, certain linguistic tasks present significant challenges even in general settings. One such task is *lexical entailment*, which involves identifying hierarchical relationships between words, particularly verbs. This task is crucial for comprehending sentence meanings and verb dynamics. Verbs often have polysemous and context-dependent meanings, and capturing entailment relations among verbs involves addressing challenges like aspect, modality, tense, and fine-grained semantic differences. In this regard, the performance of various TLMs in recognizing lexical entailment is also explored in this thesis, through differently devised prompting strategies and zero-/few-shot settings over verb pairs from lexical databases.

In addition to syntactic and semantic challenges, TLMs are increasingly tasked with interpreting more abstract aspects of language that handle *psychosocial dimensions*, such as mental-health issues, moral values and emotions. *Mental-health* issues primarily concern with identifying and addressing psychological states, often requiring models to detect symptoms, assess risk factors, and offer diagnostic insights. TLMs have shown to be effective in detecting mental-health issues, yet existing studies mainly focus on specific tasks like detecting depression on social media. In this regard, we provided a comprehensive overview on TLM-based methods for detecting mental-health issues, highlighting their potential and limitations in clinical applications.

In contrast to mental-health issues, *moral values* relate more to the interpretative understanding of ethical norms, intentions, and affective states, which shape social interactions and cultural frameworks. Capturing and predicting these values requires models to go beyond traditional linguistic tasks, as this involves understanding the subtleties of human actions, thoughts and social norms. This adds complexity, shifting the focus from lexical or syntactic understanding to interpreting the moral dimensions in narratives. In this regard, a part of this thesis was devoted to the analysis of how TLMs capture moral values in text and whether they exhibit moral tendencies incorporated during training. This goal initially stemmed from an investigation of TLMs in the context of *Moral Foundation Theory* (MFT), a well-established framework identifying the core moral foundations shaping individual and social lives across cultures. Previous research has focused more broadly on ethics in AI, often neglecting fine-grained integration of MFT into TLMs. Existing sur-

veys typically evaluate TLM alignment with cultural or ethical values, but few have examined the moral foundations encoded in these models. Our study brings order to the MFT tasks, data sources, and methods, unveils main trends and discusses challenges and further perspectives. The research on this topic continued with the introduction of *E2MoCase*, a new dataset that integrates *emotions*, *events*, and *moral values* from media narratives about high-profile legal cases. Previous studies have focused on either emotions, moral reasoning, or event detection in isolation. Our work addresses the need for resources that jointly capture these three dimensions, enabling in-depth analysis of how emotional tone, moral values, and events shape public perception in legal contexts. Our constructed dataset supports a range of NLP tasks, such as moral value prediction and bias detection, offering a multi-dimensional view of media bias. Building on this dataset, *ME²-BERT* was proposed, which is a framework designed to integrate emotions, events, and moral values for accurate moral value prediction in *out-of-domain scenario*, where training and testing data come from different domains. The model uses a domain adaptation strategy that combines these three dimensions to better capture the complex interrelations shaping moral judgment in text. Prior models applied domain adaptation techniques but lacked integration of emotional and event-based information. *ME²-BERT* is a novel framework to simultaneously incorporate emotions and events, addressing the limitations of previous approaches and offering a novel way to understand how moral dimensions are embedded in narratives.

Contributions. The key contributions of this thesis are summarized as follows:

- We developed a comprehensive framework that systematically analyzes TLM-based methods for AI-driven tasks in the legal sphere, addressing challenges such as domain adaptation, handling long and complex legal texts, and processing multilingual legal corpora.
- We advanced the understanding of European constitutions performing a multi-faceted similarity analysis and exploring the knowledge and the summarization abilities of generative TLMs like GPT-3 on constitutional texts, focusing on rights and duties.
- We investigated the ability of state-of-the-art generative TLMs to recognize lexical entailment relationships, using diverse prompting strategies in zero- and few-shot settings.
- We conducted a thorough exploration of TLM-based methods for mental-health detection, analyzing how existing methods are able to effectively address a number of major mental-health issues.
- We pioneered the integration of emotions, events, and moral values through the creation of the *E2MoCase* dataset, which enables a psychosocial analysis of media narratives of high-impact legal cases.
- Building upon *E2MoCase*, we proposed *ME²-BERT*, a framework which combines emotional, event-based, and moral information for improved moral value prediction, particularly in out-of-domain scenarios.

Plan of the thesis. This thesis begins with a comprehensive background chapter (*Chapter 2*), which provides the fundamental concepts and knowledge required

to understand the research. In particular, it provides a detailed overview of main architectural aspects of the involved TLMs, tasks and problems addressed, data and benchmarks.

The remainder of the thesis is organized into three main parts, each focusing on a distinct domain of application for TLMs. The first part discusses TLMs-based NLP methods and approaches for the legal domain. It begins with a systematic survey of TLMs and their application to legal tasks (*Chapter 3*). The following chapter (*Chapter 4*) presents our work in analyzing European constitutions using TLMs. The second part investigates the ability of TLMs to address lexical entailment, evaluating TLMs in zero- and few-shot settings (*Chapter 5*). The third part explores the application of TLMs in addressing psychosocial challenges, including mental-health and moral dimensions. In particular, *Chapter 6* presents a systematic survey of TLMs-based methods for detecting mental-health conditions, such as depression and suicidal ideation, from textual data. *Chapter 7* and *Chapter 8* introduce, respectively, the E2MoCase dataset and the ME²-BERT framework.

It should be emphasized that each chapter of the thesis, and a large part of the background chapter, corresponds to a published work co-authored by the candidate.

Chapter 2

Background

2.1 Transformer

Since their debut in the NLP field, Transformer models (Vaswani et al., 2017) have become a standard de-facto in the development of revolutionary deep-learning models that pushed the state-of-the-art in many challenging NLP tasks. The key point in the Transformer architecture is the use of *attention* mechanisms (Bahdanau et al., 2015) as an alternative to the conventional recurrent neural networks (RNNs) and convolutional neural networks (CNNs). The attention takes into account all the hidden states of a neural network and chooses which ones is to ignore and which ones is to remember, giving the proper weight to words' dependencies regardless of the distance in the sequence. The weights are computed through an attention function, whose purpose is to value the relevance of an input word with respect to a target word. Nowadays, attention mechanisms are ubiquitous in NLP and have boosted performances on a wide range of tasks.

Transformer is the first model to rely solely on attention mechanisms, leaving out the use of RNNs and CNNs. The attention is modeled in two ways. Firstly, it estimates the importance of words in the same sequence, i.e., input words are treated as both sources and targets, with the goal of calculating appropriate representations of the input sequence that reflect syntactic-semantic connections. This mechanism is also referred to as *self-attention*. Secondly, the input representations are weighted through attention to predict target tokens in an auto-regressive manner. The Transformer's architecture consists of multiple stacked encoder-decoder structures (Fig. 2.1).

The attention function used in the Transformer is called *scaled dot-product attention* (Vaswani et al., 2017). For each token, three types of vectors are defined, namely query, key and value, with dimensions d_q , d_k and d_v , respectively. A weight is assigned to each value vector, depending on the compatibility of the query and the key. The query represents the token against which the importance of the other tokens, namely the keys, have to be evaluated. Formally, given n tokens in input, $\mathbf{Q} \in \mathbb{R}^{n \times d_q}$, $\mathbf{K} \in \mathbb{R}^{n \times d_k}$, $\mathbf{V} \in \mathbb{R}^{n \times d_v}$ are the representation matrices of queries, keys and values, respectively, the scaled dot-product attention $Att(\mathbf{Q}, \mathbf{K}, \mathbf{V}) \in \mathbb{R}^{n \times d_v}$ is

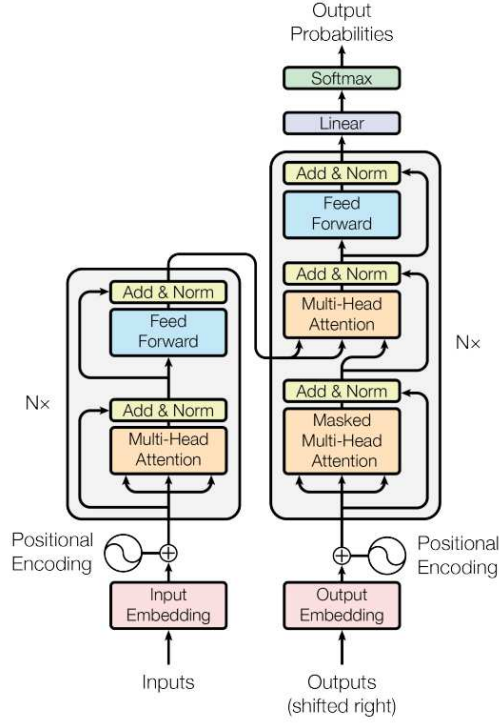


Fig. 2.1: Transformer architecture (Vaswani et al., 2017). On the left side, the encoder component with the multi-head self-attention mechanism. On the right side, the decoder component with two attention mechanisms, a masked multi-head self-attention for decoder outputs and a multi-head attention for encoder-decoder outputs.

defined as follows:

$$Att(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = softmax\left(\frac{\mathbf{QK}^T}{\sqrt{d_k}}\right) \cdot \mathbf{V}, \quad (2.1)$$

where term $\sqrt{d_k}$ is used as a scaling factor. By default, both d_k , d_q and d_v have a value of 64. All the layers in the model, as well as the embedding layers, produce outputs of dimension $d_{model} = 512$.

The encoders consist of multiple heterogeneous self-attention mechanisms, each capturing a different kind of word relation in order to reflect several syntactic-semantic nuances. A single type of self-attention is called *head*, so that the encoder has a *Multi-Head Self-Attention mechanism*:

$$\begin{aligned} MultiHead(\mathbf{Q}, \mathbf{K}, \mathbf{V}) &= concat(\mathbf{head}_1, \dots, \mathbf{head}_h) \cdot \mathbf{W}^{(O)} \\ \mathbf{head}_i &= Att(\mathbf{QW}_i^{(Q)}, \mathbf{KW}_i^{(K)}, \mathbf{VW}_i^{(V)}), \end{aligned} \quad (2.2)$$

where $\mathbf{W}_i^{(Q)} \in \mathbb{R}^{d_{model} \times d_q}$, $\mathbf{W}_i^{(K)} \in \mathbb{R}^{d_{model} \times d_k}$, $\mathbf{W}_i^{(V)} \in \mathbb{R}^{d_{model} \times d_v}$ are the projection matrices for queries, keys and values, respectively. $\mathbf{W}^{(O)} \in \mathbb{R}^{hd_v \times d_{model}}$ is a projection matrix for the output.

The decoders have a similar structure, but they require the use of *masked* multi-head self-attention mechanism in order to predict the current word without having access to subsequent words, i.e., in an auto-regressive manner. In addition, they also use an *Encoder-Decoder Attention mechanism*, i.e., a multi-head attention over the output of the encoder stack, in order to catch relevant information for the current target prediction.

Queries, keys and values may have different meanings depending on the type of attention. In the self-attention, they all come from the output of the previous layer in the encoder (as well as in the decoder). In the encoder-decoder attention, queries come from the previous decoder layer, while keys and values come from the encoder output.

2.2 Transformer-based Language Models

In the following, we describe main characteristics of the TLMs that have been used to address several problems and tasks undertaken throughout the Ph.D. course (cf. Section 2.3.2). Following Yang et al. (2024), we organize the presentation of TLMs by architecture type: Encoder-only, Encoder-Decoder and Decoder-only TLMs. Table 2.1 provides detailed information about these models, including their references, names, organizations, availability (open or closed), number of parameters, training approaches and the main NLP categories they are involved in. Regarding the last column (NLP category), the table indicates whether the models are applied to tasks of natural language understanding (NLU) and/or natural language generation (NLG). When relevant, it also specifies one or more particular aspects of NLP that the model is able to perform: conversation, instruction-following, math, coding, multi-modal.

Table 2.1: Details of the TLMs discussed in the background chapter. The columns include the reference, model name, organization, availability (open or closed), number of parameters, training approach (PT: Pre-training; KD: Knowledge Distillation; SFT: Supervised Fine-tuning; RLHF: Reinforcement Learning with Human Feedback), and NLP category (NLU: Natural Language Understanding; NLG: Natural Language Generation; conversation; reasoning; instruction-following; math; coding; multi-modal).

Reference	Name	Organization	Availability	# Params	Training	NLP category
ENCODER-ONLY TLMs						
Devlin et al. (2019)	BERT	Google	Open	110M (base) 340M (large)	PT+SFT	NLU
Liu et al. (2019)	RoBERTa	Meta	Open	125M (base) 355M (large)	PT+SFT	NLU
Reimers and Gurevych (2019)	S-BERT S-RoBERTa	-	Open	PT+SFT		NLU
Joshi et al. (2020)	SpanBERT	Meta	Open	340M	PT+SFT	NLU
Clark et al. (2020)	ELECTRA	Google	Open	13M (small) 109M (base) 334M (large)	PT+SFT	NLU
Conneau et al. (2020)	XLM-RoBERTa	Meta	Open	270M (base) 550M (large)		NLU
He et al. (2021)	DeBERTa	Microsoft	Open	138M (base) 405M (large) 1.5B (1.5B)	PT+SFT	NLU
Sanh et al. (2019)	DistilBERT	HuggingFace	Open	66M	KD+SFT	NLU
Lan et al. (2020)	ALBERT	Google	Open	12M (base) 18M (large)	PT+SFT	NLU

Continued on the next page

Reference	Name	Organization	Availability	# Params	Training	NLP category
				60M (xlarge) 235M (xxlarge)		
Iandola et al. (2020)	SqueezeBERT	-	Open	51M	PT+KD+SFT	NLU
Wang et al. (2020b)	MiniLM	Microsoft	Open	33M (mono-lingual) 118M (multi-lingual)	KD+SFT	NLU
Beltagy et al. (2020)	Longformer	AllenAI	Open	149M (base) 435M (large)	PT+SFT	NLU
Zaheer et al. (2020)	BigBird	Google	Open	128M (base) 359M (large)	PT+SFT	NLU
ENCODER-DECODER TLMs						
Beltagy et al. (2020)	LED	AllenAI	Open	162M (base) 460M (large)	SFT	NLU & NLG
Raffel et al. (2020)	T5	Google	Open	60M (small) 220M (base) 770M (large) 3B (3B) 11B (11B)	PT+SFT	NLU & NLG
Lewis et al. (2020a)	BART	Meta	Open	139M (base) 406M (large)	PT+SFT	NLU & NLG
Dong et al. (2019)	UniLM	Microsoft	Open	340M	PT+SFT	NLU & NLG
Song et al. (2019)	MASS	Microsoft	Open	base and middle sizes,	PT+SFT	NLG

Continued on the next page

Reference	Name	Organization	Availability	# Params	Training	NLP category
				#params unknown		
DECODER-ONLY TLMs						
Radford et al. (2018)	GPT	OpenAI	Open	117M	PT+SFT	NLG & NLU
Radford et al. (2019)	GPT-2	OpenAI	Open	1.5B	PT	NLG & NLU
Brown et al. (2020)	GPT-3	OpenAI	Closed	1.3B (babbage) 6.7B (curie) 175B (davinci)	PT	NLG & NLU conversation instruction-following
-	GPT-3.5	OpenAI	Closed	Undisclosed	PT+SFT+ RLHF	NLG & NLU coding, math, conversation
OpenAI (2023)	GPT-4	OpenAI	Closed	Undisclosed	PT+SFT+RLHF	NLG & NLU conversation, coding, math, instruction-following multi-modal
-	ChatGPT	OpenAI	Closed	based on the underlying model	PT+SFT+RLHF	NLG & NLU conversation, instruction-following
Black et al. (2022)	GPT-NeoX-20B	EleutherAI	Open	20B	PT	NLG & NLU math
Shliazhko et al. (2022)	mGPT	-	Open	1.3B; 13B	PT	NLG & NLU

Continued on the next page

Reference	Name	Organization	Availability	# Params	Training	NLP category
						conversation
Touvron et al. (2023a)	LLaMA	Meta	Open	6.7B; 13B 32B; 65B	PT	NLG & NLU coding, math
Touvron et al. (2023b)	LLaMA-2	Meta	Open	7B; 13B; 34B; 70B	PT	NLG & NLU coding, math
Touvron et al. (2023b)	LLaMA-2-Chat	Meta	Open	7B; 13B; 34B; 70B	PT+SFT+RLHF	NLG & NLU conversation, coding, math
Dubey et al. (2024)	LLaMA-3	Meta	Open	8B; 70B	PT+SFT+RLHF	NLG & NLU coding, math multi-modal
-	Vicuna	LMSYS Org	Open	7B; 13B; 33B	SFT	NLG & NLU conversation, coding, math, instruction-following;
Jiang et al. (2023)	Mistral	Mistral AI	Open	7B; 13B	PT	NLG & NLU coding, math
Jiang et al. (2023)	Mistral-Instruct	Mistral AI	Open	7B; 13B	PT+SFT	NLG & NLU instruction-following
Almazrouei et al. (2023)	Falcon	TII	Open	7B; 11B 40B; 180B	PT	NLG & NLU

Continued on the next page

Reference	Name	Organization	Availability	# Params	Training	NLP category
Zhang et al. (2022)	OPT	Meta	Open	125M; 350M; 1.3B 2.7B; 6.7B; 13B 30B; 66B	PT	NLG & NLU conversation
Köpf et al. (2023)	OpenAssistant	LAION AI	Open	based on the underlying model	SFT+RLHF	NLG & NLU conversation
Anthropic (2024)	Claude-3	Anthropic	Closed	Undisclosed	PT+SFT+RLHF	NLG & NLU coding, math multi-modal

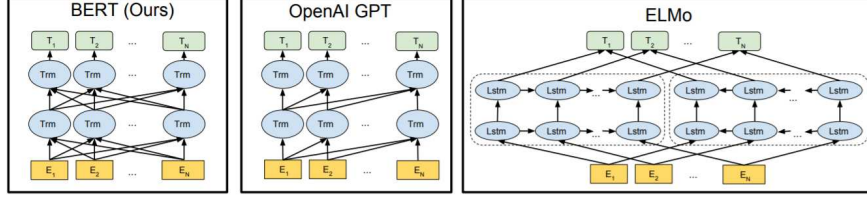


Fig. 2.2: Difference between deeply bidirectional BERT architecture, left-to-right GPT architecture and shallow bidirectional ELMo architecture (Devlin et al., 2019).

2.2.1 Encoder-only models

BERT. The first known TLM which has revolutionized the way to approach the natural language understanding challenges, pushing the state-of-the-art in many demanding NLP benchmarks, is Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019).¹ BERT has represented a breakthrough in several NLP benchmarks and it is still considered a must-have baseline (Rogers et al., 2020). It essentially consists of a stack of Transformer encoder layers (Fig. 2.1). The key aspects of BERT are the *bidirectional unsupervised pre-training* and the beneficial effect of having a unified architecture across different tasks. The framework uses the pre-training fine-tuning paradigm. In the pre-training stage, the model uses a masked language objective to get deep bidirectional representations from unlabeled text by jointly conditioning on left and right context-words in all encoder layers. A deep bidirectional model has shown to outperform a conventional unidirectional model or even a shallow left-to-right and right-to-left concatenated model (like ELMo (Peters et al., 2018)).

To obtain a deep bidirectional representation of unlabeled text in the pre-training phase, BERT employs a particular pre-training objective task called *Masked Language Modeling* (MLM). This task is to predict masked input words from unlabeled text by conjointly conditioning on left and right context-words in all encoder layers. In other words, BERT considers a deeply bidirectional context in the prediction of masked tokens, and as a result it builds more expressive and meaningful word embeddings than a conventional unidirectional or shallow bidirectional models (Fig. 2.2). In addition, BERT uses a second pre-training objective task called *Next Sentence Prediction* (NSP) to cover a variety of downstream tasks involving sentence pairs. Given two word sequences as input, the NSP task is to determine if the second sequence is subsequent to the first in the original document.

Each word is first tokenized into *word pieces* (Wu et al., 2016), which forms a vocabulary of 30,522 tokens.² The input representation is the sum of token em-

¹ <https://github.com/google-research/BERT>

² Using the WordPiece tokenization process, the vocabulary is obtained using a data-driven approach: given a training corpus and a number w of word pieces for the vocabulary, the task is

beddings with segment and position embeddings. Segment embeddings take into account the token location, i.e., the tokens of a sentence have all the same segment encoding. Position embeddings keep track of the token absolute position in the text. The first token of the sequence is always a special token called [CLS], and the output representation of this token is considered to be representative of the whole input sequence. In the NSP task, the two sentences in an input sequence are separated to each other by another special token called [SEP]. In the MLM task, words to be masked are replaced with the [MASK] special token. The introduction of [MASK] has the advantage to allow the model to rely jointly on left and right context. However, it also cause a mismatch between the pre-training and the subsequent, downstream fine-tuning phase, since the latter does not introduce any mask. Therefore, to mitigate this issue, not all the tokens are replaced with [MASK], but only a certain percentage; in particular, from all tokens, 15% are selected to be masked, using a random token in 10% of cases, [MASK] in 80% of cases and leave the original token in the remaining 10%. The model does not know which words have been randomly replaced, so it has to keep a distributional contextual representation for each token.

BERT is originally pre-trained using two unlabeled corpora, named Book Corpus and English Wikipedia, for a total of 3,300M words. The input sequence has a limit of 512 tokens. For NSP, the training examples are sampled so that in 50% of cases the second sentence is actually subsequent to the first, while in the remaining 50% the second sentence is randomly selected from the corpus.

The authors released BERT in two format sizes: BERT-*base* and BERT-*large*. BERT-*base* consists of 12 stacked encoder layers with 12 attention heads and a hidden size of 768, for a total of 110M parameters, while BERT-*large* has 24 stacked encoder layers, 16 attention heads and a hidden size of 1024 (340M total parameters).

Using BERT for downstream tasks is straightforward, as it just requires one additional task-specific layer on top of the model and a fine-tuning phase on parameters. The input representation is analogue to the pre-training one described above. Typically, the final hidden state corresponding to the [CLS] token is used as the input of the additional top layer for tasks such as classification and entailment; analogously, the final hidden states of the input tokens are fed into the additional top layer for token-level tasks, such as sequence-tagging and question-answering.

BERT contextual representations of words can also be used as word embeddings for feature-based models. The authors experimented using BERT embedding as input for a randomly initialized two-layer BiLSTM. They found that great performance can be achieved by combining the outputs of BERT layers, especially when the last four layers' outputs of BERT-*large* are concatenated.

Several studies have examined the encoded knowledge in the BERT weighted representations (Rogers et al., 2020). It has been shown that BERT naturally learns syntactical information, such as parts of speech and syntactic chunks, and its representations are characterized as being hierarchical. BERT has also semantic knowl-

to select w word pieces such that the segmented corpus contains as much unsegmented words as possible. WordPiece tokenization has shown to deal with the out-of-vocabulary words better than standard tokenization procedures.

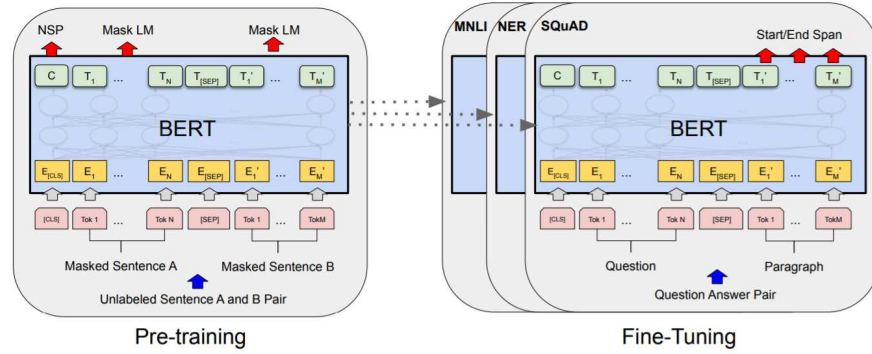


Fig. 2.3: BERT Pre-training and Fine-tuning (Devlin et al., 2019).

edge, as it encodes information like semantic roles, entity types, relations etc. Several studies stated that the most information about the order of words is in the lower layers of BERT, syntactic information is primarily in the middle layers, while the most task-specific information is in the final layers. Semantic information, instead, is all over the model since it pervades all the language. In the knowledge induction by filling in the blanks, a study has shown that, for some relation types, BERT can be competitive with methods supported by knowledge bases.

BERT was tested on a number of benchmarks, including GLUE, SWAG, SQuAD v1.1 and v2.0 (Section 2.4.1), getting the top results on eleven tasks. In fact, it advanced the state-of-the-art, obtaining 80.5 score (7.7% point improvement compared to previous models) on the official GLUE leaderboard,³ 86.3% test accuracy on SWAG (8.3% point improvement compared to GPT), 1.5 more points on SQuAD v1.1 Test F1 and 5.1 more points on SQuAD v2.0 Test F1 compared to previous models.

Several instances of BERT model have been released over the years, which mainly differ from each other in terms of model size (as above discussed) as well as training objective and case folding. The most common instances are the case-sensitive BERT-base-cased⁴ and the uncased BERT-base-uncased.⁵

There is also multi-lingual variants of BERT, such as *m-BERT*.⁶ Actually, only the base version of the m-BERT is available (12 layers, 12 attention heads, 768 hidden size, 110M parameters in total), trained for more than 100 languages and a specific version for Chinese.

Over the time, BERT has inspired extensive research in the development of different variants and enhancements of the model.

³ <https://gluebenchmark.com/leaderboard>

⁴ <https://huggingface.co/bert-base-cased>

⁵ <https://huggingface.co/google-bert/bert-base-uncased>

⁶ <https://github.com/google-research/bert/blob/master/multi-lingual.md>

RoBERTa. RoBERTa (Liu et al., 2019) is conceived to improve BERT based on the presumed under-training and non-optimal setting of the key hyperparameters of BERT.⁷ RoBERTa indeed develops a more robust and optimized approach through two main interventions. Firstly, the amount of training data is increased by adding three further corpora besides Book Corpus and English Wikipedia, namely CC-News, which consists of the English portion of CommonCrawl news dataset (76 GB), Stories (31 GB), and OpenWebText (38GB). Secondly, RoBERTa is pre-trained much longer, which is related to the finding that decreasing BERT’s number of steps and increasing the batch size leads to better results, with the same computational cost.

Substantial changes on the pre-training procedure have been made too. Pre-training sequences have a length of 512 tokens, discarding BERTshort-sequence strategies — in BERT, 90% of the pre-training steps involve sequences of length 128 tokens, and the remaining 10% use sequences of 512 tokens. The next sentence prediction task is removed, as it is not regarded as relevant to the model’s performance. Moreover, the masked language modeling task is modified, by introducing a *dynamic* masking that avoids the use of the same mask for each instance and for each epoch.

The text representation of RoBERTa is obtained through a byte-level vocabulary containing 50K subwords units, in contrast with the character-level vocabulary of 30K tokens used by BERT. The text segmentation method used in RoBERTa is the Byte-Pair Encoding (BPE) (Sennrich et al., 2016).

Like for BERT, two sizes of RoBERTa have been released: RoBERTa-*base* (12 layers, 12 attention heads, 768 hidden size) and RoBERTa-*large* (24 layers, 16 attention heads, 768 hidden size). RoBERTa has shown to overcome BERT in classic benchmarks such as GLUE, SQuAD v2.0 and RACE (Section 2.4.1).

S-BERT and S-RoBERTa. S-BERT and S-RoBERTa (Reimers and Gurevych, 2019) are modified versions of BERT and RoBERTa, respectively, especially designed for tasks such as semantic textual similarity, clustering, and information retrieval via semantic search.⁸ One problem is the massive computational overhead concerning the search for the most similar sentence-pairs in a collection of thousands of sentences. S-BERT and S-RoBERTa drastically reduce the computational effort of BERT for finding the most similar pair in a collection, while maintaining the same accuracy. S-BERT (resp. S-RoBERTa) fine-tunes BERT (resp. RoBERTa) based on a siamese network architecture which, given two pre-trained BERT (resp. RoBERTa) models, one for each input sentence, it ties the models’ weights which are updated during fine-tuning, to obtain semantically-expressive fixed-sized sentence embeddings based on a pooling operation. The two resulting sentence embeddings are concatenated with their element-wise difference, prior to the softmax layer for class prediction.

⁷ <https://github.com/facebookresearch/fairseq/blob/main/examples/roberta/README.md>

⁸ <https://github.com/UKPLab/sentence-transformers>

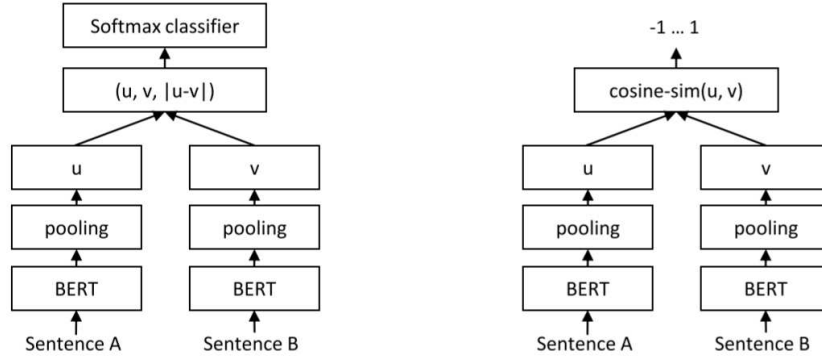


Fig. 2.4: S-BERT architectures for classification (on the left) and inference/regression (on the right) tasks (Reimers and Gurevych, 2019). In the first case, the concatenation of the sentence embeddings u and v and the element-wise difference $|u - v|$ is the input for a softmax classifier. In the second case, the cosine similarity between u and v is calculated.

Siamese and triplet networks are used to obtain semantically meaningful sentence embeddings, then these embeddings can be compared using similarity criteria (e.g., cosine similarity). The final layer structure depends on the task at hand to be optimized (Fig. 2.4). For classification purpose, the sentence embeddings are concatenated with their element-wise difference and fed into a softmax classifier, using an optimized cross-entropy loss. For regression purpose, cosine similarity between sentence embeddings and mean-squared-error loss are used. In addition, the authors experimented the model with a triplet objective function.

S-BERT has been trained on NLI data (SNLI and MultiNLI) and evaluated on semantic textual similarity (STS) tasks (Section 2.3.1), both unsupervised, i.e., without any task-specific training data, and supervised tasks, in which the STS benchmark was used for fine-tuning. S-BERT sentence embeddings has been evaluated using the *SentEval* toolkit, in which the embeddings are used as features for a logistic regression classifier. Using RoBERTa in place of BERT (S-RoBERTa), the model generates quite similar sentence embeddings and obtain similar performance. Although S-BERT embeddings are not designed for transfer-learning on other tasks, experiments reveal that they can achieve good performance in most *SentEval* transfer learning tasks.

Following S-BERT, a variety of models known as *Sentence-Transformers* has emerged over time.⁹ These models expand the original framework by incorporating various Transformer architectures, offering specialized solutions for tasks such as sentence embedding, semantic search, and clustering, and further improving efficiency and performance across diverse applications.

⁹ https://sbert.net/docs/sentence_transformer/pretrained_models.html

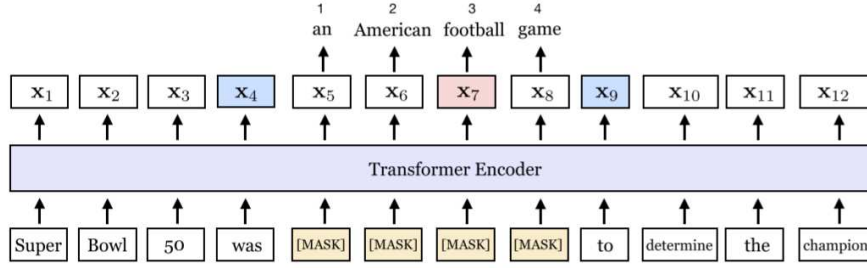


Fig. 2.5: SpanBERT training (Joshi et al., 2020). The masked span x_5, x_6, x_7, x_8 is predicted using the boundaries x_4 and x_9 .

SpanBERT. SpanBERT (Joshi et al., 2020) introduces a novel pre-training approach that differs from BERT in three aspects: the masking scheme, the training objective and the sequence training procedure.¹⁰

While maintaining the same percentage distribution of masking as BERT, SpanBERT masks adjacent random *spans* rather than random single tokens. The span masking is performed by sampling spans of text. The span length is picked from an unbalanced geometric distribution that favors short length selection, where the starting point of span masking is randomly selected too. Eventually, complete words only (instead of sub-words) are masked.

In addition, SpanBERT introduces a novel pre-training objective, called *span-boundary objective* (SBO). This task is to predict each token of the span by using only the observed tokens at the boundaries, i.e., the first token before the start and first token after the end of the span (Fig. 2.5). The idea is to encourage the model to record as much span information as possible in the boundary output encodings. Each token of the span is then represented using the boundary output encodings and relative position embedding of the target token. Like in MLM, SBO minimizes the cross-entropy loss. The overall loss is the sum of both MLM and SBO objectives for each token in the span:

$$\begin{aligned} L(x_i) &= L_{MLM}(x_i) + L_{SBO}(x_i) \\ &= -\log P(x_i | \mathbf{x}_i) - \log P(x_i | \mathbf{y}_i), \end{aligned} \quad (2.3)$$

where x_i is the i -th token in the span, $\mathbf{x}_i$ is the encoder output and $\mathbf{y}_i$ is the model representation of x_i , which corresponds to a function of the external boundaries and the relative token position in the span:

$$\mathbf{y}_i = f(x_{s-1}, x_{e+1}, p_{i-s+1}), \quad (2.4)$$

where s and e indicate the start and the end positions of the span, $\mathbf{x}_{s-1}$ is the boundary on the left of the span, $\mathbf{x}_{e+1}$ is the boundary on the right of the span, p_{i-s+1} is the

¹⁰ <https://github.com/facebookresearch/SpanBERT>

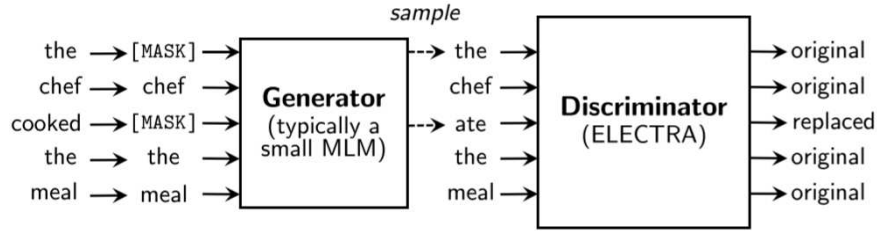


Fig. 2.6: Replaced token detection task in ELECTRA (Clark et al., 2020).

relative token position and $f(\cdot)$ is a function implemented as a 2-layer feed-forward network. The model gets rid of the NSP task, since the authors suggest that using only single sequences in the training process benefits from two aspects: the model can take advantage from longer contexts and the noise, coming from the context of other documents, is removed.

The implementation is the same as BERT-*large*, also in terms of corpora (i.e., Book Corpus, English Wikipedia) and tokenization method (i.e., WordPiece), but changing some configuration setting. Inspired by RoBERTa, in the training process the model uses different masks at each epoch and considers only sequences of 512 tokens till the end of the document, discarding BERT’s short-sequence strategies.

SpanBERT has been tested on several benchmarks regarding question answering, coreference resolution and relation extraction, as well as nine GLUE tasks. Experimentation reveals that SpanBERT can outperform BERT on most benchmarks.

ELECTRA. The ELECTRA model (Clark et al., 2020) originates from the observation that the use of a discriminative approach in the pre-training phase, which allows the model to learn from all input tokens, is computationally more efficient compared to a generative approach like MLM in BERT.¹¹ Within this view, a new pre-training task, called *Replaced Token Detection* (RTD), is proposed such that, instead of using a mask like in MLM, tokens are replaced with plausible generated tokens. In this way, the model is trained to discriminate real tokens from credible yet fake tokens. This type of corruption overcomes the BERT’s mismatch between pre-training phase, in which artificial [MASK] tokens are used, and fine-tuning phase (where there are no artificial tokens). The key aspect of the task is that the model can learn from all input tokens and not just from a small subset, thus achieving a significant speed-up in the training phase compared with BERT.

ELECTRA consists of a generator followed by a discriminator, both consisting of a Transformer encoder. The generator is trained to produce plausible tokens, replacing real tokens in a random set of positions. The training is performed in a MLM manner: given a random set of masked tokens, the generator learns to predict the original tokens. If the prediction is correct then it is regarded as “real”, otherwise as “fake”. The discriminator is trained to predict if the current token is real or fake

¹¹ <https://github.com/google-research/electra>

(Fig. 2.6). After pre-training, the generator is discarded so that only the discriminator is used in the fine-tuning phase.

ELECTRA is available in three model sizes: *ELECTRA-small* (12 layers, 4 attention heads, 256 hidden size), *ELECTRA-base* (12 layers, 12 attention heads, 768 hidden size), *ELECTRA-large* (24 layers, 16 attention heads, 1024 hidden size). In the token masking process, the 15% of tokens are masked out (like in BERT), except for *ELECTRA-large* in which the percentage is increased to 25%. *ELECTRA-large* is trained with 400 steps or 1.75M steps: in the first case, it achieved performance in GLUE comparably to RoBERTa, with much less computational cost (less than 1/4 pre-training cost), whereas in the second case, it outperforms RoBERTa still with less computational costs. *ELECTRA-large* outperforms RoBERTa also in SQuAD v2.0. Indeed, *ELECTRA-base* overcomes *BERT-large* on GLUE development set (85.1 against 84.0).

XLM-RoBERTa. XLM-RoBERTa (Conneau et al., 2020; Dadu and Pant, 2020) is a highly scalable cross-lingual model which was designed to improve cross-lingual language understanding (XLU).¹² It has indeed shown that low-resource language performance can be improved if scaled with similar high-resource language during pre-training, yet there is a trade-off between the positive transfer and the capacity dilution of the model (i.e., the number of parameters). More precisely, since model capacity is limited by time and memory constraints, the capacity reserved for a single language is diluted among all languages as the number of languages grows in a fixed-sized model, and once a certain level of dilution is exceeded, the overall performance is degraded. Adding more capacity to the model alleviates this condition, but it is unsuitable for models with modest sizes. The authors also deal with the appropriate allocation of the model capacity across high-resource and low-resource languages and the scaling of both model and vocabulary size with respect to the number of languages.

The architecture of XLM-RoBERTa emulates the XLM approach (Conneau and Lample, 2019) for training objective, languages, and data, with few refinements to increase performance at scale. The model is pre-trained with a multi-lingual MLM objective, using monolingual data and sampling pieces of text from each language. The training data is obtained from a filtered and cleaned version of CommonCrawl and consists of more than 2TB of data in 100 languages. Sub-word tokenization is directly applied on the raw text for all languages using a SentencePiece model. Following RoBERTa, the authors show that training the model longer and with a larger-scale corpus yields improvements in the model performance.

XLM-RoBERTa is available in two model sizes: *XLM-RoBERTa-base* (12 layers, 12 attention heads, 768 hidden size, 270M parameters) and *XLM-RoBERTa-large* (24 layers, 16 attention heads, 1024 hidden size, 550M parameters). In Goyal et al. (2021), two larger versions are proposed: *XLM-RoBERTa-xl* (36 layers, 32 attention heads, 2560 hidden size, 3.5B parameters) and *XLM-RoBERTa-xxl* (48 layers, 32 attention heads, 4096 hidden size, 10.7B parameters).

¹² <https://github.com/facebookresearch/xlm>

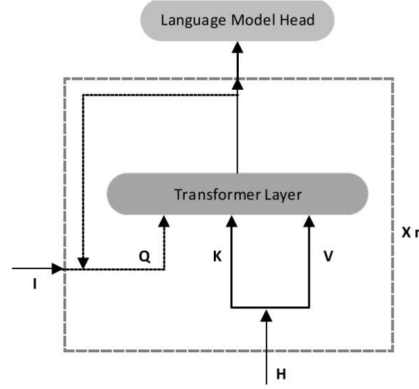


Fig. 2.7: DeBERTa’s enhanced mask decoder (EMD) (He et al., 2021). There are n stacked EMD layers and two inputs: H and We . H is the hidden state from previous Transformer layer, while We represents any additional information (H , absolute position or output from previous EMD layer). When $We = H$ and $n = 1$, EMD is equivalent to a BERT layer.

XLM-RoBERTa has been tested on cross-lingual understanding tasks as XNLI (Conneau et al., 2018), CoNLL (2002 (Sang, 2002) and 2003 (Sang and Meulder, 2003)) and MLQA. XLM-RoBERTa has shown to outperform previous models like mBERT and XLM in each of these cross-lingual benchmarks. Also, competitive results were obtained also in monolingual benchmarks like GLUE, proving that it is possible for a multi-lingual model to achieve on a single language the same performance than monolingual models.

DeBERTa. DeBERTa (He et al., 2021) introduces two novel techniques to improve upon BERT and RoBERTa: the *disentangled attention mechanism* and an *enhanced mask decoder*.¹³ The former consists in computing the attention weights among words using two disentangled matrices, which represent word contents and positions, respectively. In fact, unlike BERT, DeBERTa separates content and relative position encodings in two vectors, so that the attention score for a word pair is calculated by summing up the content-to-content, content-to-position and position-to-content attention:

$$\begin{aligned} A_{i,j} &= \{\mathbf{h}_i, \mathbf{p}_{i|j}\} \times \{\mathbf{h}_j, \mathbf{p}_{j|i}\}^\top \\ &= \mathbf{h}_i \mathbf{h}_j^\top + \mathbf{h}_i \mathbf{p}_{j|i}^\top + \mathbf{p}_{i|j} \mathbf{h}_j^\top, \end{aligned} \quad (2.5)$$

where $A_{i,j}$ is the attention score between tokens i and j , $\mathbf{h}_i$ (resp. $\mathbf{h}_j$) is the content representation of i (resp. j), $\mathbf{p}_{i|j}$ (resp. $\mathbf{p}_{j|i}$) is the relative position with token j (resp. i).

¹³ <https://github.com/microsoft/DeBERTa>

DeBERTa is pre-trained using the MLM task, but unlike BERT, it has no absolute positional embeddings. The absolute position is very important for the prediction of masked tokens, in order to discern syntactical nuances in a sentence. For this purpose, DeBERTa incorporates the absolute positional embeddings just before the *softmax* layer responsible of masked word predictions (Fig. 2.7). This approach is called *enhanced mask decoder* (EMD) and allows the model to consider relative position in all layers and absolute position only as an additional information when masked words have to be decoded. The pre-training settings of large DeBERTa models is analogue to BERT except for the use of the BPE vocabulary. Like RoBERTa, training data is composed by English Wikipedia and Book Corpus (already used in BERT) with the addition of OpenWebText and Stories, for a total of about 78Gb. Following RoBERTa, DeBERTa uses a dynamic batch data to handle documents shorter than 512 tokens. Inspired by SpanBERT, it includes also span masking strategies.

DeBERTa is implemented with three version sizes: DeBERTa-*base* (12 layers, 12 attention heads, 768 hidden size), DeBERTa-*large* (24 layers, 16 attention heads, 1024 hidden size) and DeBERTa-1.5B (48 layers, 24 attention heads, 1536 hidden size). DeBERTa-1.5B is fine-tuned on a new virtual adversarial training method, called *Scale-invariant-Fine-Tuning* (SiFT), that normalizes word embeddings into stochastic vectors, which reveals to be beneficial for improving training stability in downstream tasks.

In GLUE tasks, DeBERTa outperforms BERT, RoBERTa and ELECTRA, with much less pre-training data than RoBERTa and ELECTRA (78G against 160G). DeBERTa shows outstanding performance also in SQuAD, RACE and SWAG benchmarks. DeBERTa-1.5B was tested on SuperGLUE benchmarks (Section 2.4.1), where it surpassed for the first time human performance in terms of macro-average score.

In He et al. (2023), a novel version of DeBERTa, called DeBERTaV3, is proposed. It is based on a previous model update, DeBERTaV2, where the major changes are in the vocabulary, tokenizer, input encoding and shared projection matrices (position and content) in attention layers. DeBERTaV3 is pre-trained using the same settings as the original DeBERTa but following ELECTRA in the use of a generator-discriminator setting and replacing the MLM task with the ELECTRA's pre-training task, called *Replace Token Detection*. In eight GLUE tasks, DeBERTaV3-*large* (24 layers, 12 attention heads, 1024 hidden size) obtains an average score of 91.37 on development set, outperforming ELECTRA and RoBERTa. It was also tested on SQuAD v2.0, RACE and SWAG, outperforming previous models on development set.

In He et al. (2023), a multi-lingual version of DeBERTaV3-*base* (12 layers, 12 attention heads, 768 hidden size), called mDeBERTa-*base*, is also proposed. It has the same model structure as DeBERTaV3-*base*, and it was trained with the same corpus as XLM-RoBERTa (Conneau et al., 2020) and the same SentencePiece vocabulary as mT5 (Xue et al., 2021). The pre-training setting is similar to XLM-RoBERTa except for the number of steps. It was tested on XNLI and obtained higher results than previous base models (like mT5-*base* and XLM-RoBERTa-*base*) in all

languages and under both zero-shot cross-lingual-transfer and zero-shot translate-train-all settings.

DistilBERT. DistilBERT (Sanh et al., 2019) stands as a counterpart of the ever larger pre-trained language models, demonstrating that a smaller but still general-purpose language model can achieve comparable performance in much less time.¹⁴ This leads to significant benefits, since large-scale models often have hundred million parameters and their training phase have considerable computational costs and memory requirements.

The key aspect of DistilBERT corresponds to the *knowledge distillation*, a transfer learning technique whereby the knowledge of a larger pre-trained model, called “teacher”, is passed to a smaller model, the “student”, through a training phase wherein the student has to reproduce the same results of the teacher. In DistilBERT, the training objective is a linear combination of three loss functions: distillation loss, masked language modeling loss (same as BERT) and cosine embedding loss. The distillation loss is a cross-entropy loss function taking into account both the probabilities estimated by the teacher and the student:

$$L_d = \sum_i t_i * \log(s_i), \quad (2.6)$$

where t_i (resp. s_i) is the probability coming from the teacher (resp. from the student). Cosine embedding loss is used to measure the degree of similarity of two input vectors and aims to align the directions of student and teacher vectors.

The architecture of DistilBERT is the same as BERT except for the absence of the token embedding and the pooler layer. Moreover, the number of layers is halved compared to BERT. Following RoBERTa, the model was trained using dynamic masking, without the NSP task and with very large batches, but keeping the same corpora as the original BERT. As a result, DistilBERT has 40% fewer parameters than BERT-*base* and is 60% faster.

In the GLUE benchmark, DistilBERT obtains almost the same score as BERT in the development set. In SQuAD v1.1, it obtains good results, using a second knowledge distillation in the fine-tuning phase with BERT fine-tuned on SQuAD as teacher model.

AlBERT. AlBERT (Lan et al., 2020) stems from the same considerations of the DistilBERT’s authors regarding the excessive size of pre-trained language models, which raises concerns in terms of memory and time costs.¹⁵ In AlBERT, two parameter reduction techniques are proposed to mitigate the memory consumption and speed up the training time.

The first technique is a *factorized embedding parameterization*, which consists in the decomposition of word embeddings in smaller matrices. In BERT, the size E of the context-independent wordpiece embeddings is equivalent to the context-

¹⁴ https://github.com/huggingface/transformers/tree/main/examples/research_projects/distillation

¹⁵ <https://github.com/google-research/ALBERT>

dependent hidden layer size H ($E \equiv H$); however, it is assumed in ALBERT that these sizes should be untied, and in particular $H \gg E$ should hold. In the factorized embedding parameterization, vocabulary vectors of size V are first projected in a lower-dimensional space of size E and then projected to the hidden space of size H . Therefore, the embedding parameters are reduced from $O(V \times H)$ to $O(V \times E + E \times H)$, which is a significant reduction since $H \gg E$.

The second technique is the *cross-layer parameter sharing*. As default the model shares all parameters across layer, although other sharing techniques can be used. As a result, ALBERT has much smaller parameter size and the training is 1.7 times faster than BERT without severely affecting the performance. In addition, ALBERT configurations can be scaled up much larger than BERT. There are four sizes of ALBERT models: ALBERT-*base* (12 layers, 12 attention heads, 768 hidden size), ALBERT-*large* (24 layers, 16 attention heads, 1024 hidden size), ALBERT-*xlarge* (24 layers, 16 attention heads, 2048 hidden size), ALBERT-*xxlarge* (12 layers, 64 attention heads, 4096 hidden size). BERT-*base* has 108M total parameters, while ALBERT-*base* has only 12M parameters and ALBERT-*xlarge* has 60M parameters. ALBERT-*xxlarge* is around 70% of BERT-*large* parameters (235M against 334M), but is about 3 times slower because it has a larger structure. ALBERT-*large* has about 18 times fewer parameters compared to BERT-*large*.

To further improve the performance, the authors replaced the NSP task with another pre-training objective task, called *Sentence-Order Prediction* (SOP). Given two segments of text from the same document, the task is to predict if the second segment comes next the first segment in the document, i.e., they are consecutive (positive examples), or if the original order of the segments has been swapped (negative examples). In other words, the task is designed to train the model to understand the order of segments, focusing on *discourse coherence* predictions.

ALBERT is pre-trained on the same BERT corpora, with input length up to 512, vocabulary size of 30k. Text is tokenized using SentencePiece. ALBERT was fine-tuned under two settings: single-model and ensembles. It achieves better results than previous models like BERT in several downstream tasks.

SqueezeBERT. SqueezeBERT (Iandola et al., 2020) is a variant of BERT designed to be more efficient by reducing latency and computational requirements while maintaining comparable performance levels, particularly for use on mobile devices and other environments with limited computational resources. SqueezeBERT maintains the same number of layers, attention heads, and hidden dimensions as BERT-*base* but with approximately half the parameter count (51.1M total parameters). This reduction is achieved by introducing grouped convolutions (Fig. 2.8b), in which the input vector is split into separate sub-vectors to be independently processed in order to reduce the computational cost. A portion of these layers still contains the fully-connected layers used in BERT, which are considered a special case of 1D convolution (Fig. 2.8a). This architectural modification enables SqueezeBERT to process text sequences 4.3 times faster than BERT-*base* on a Google Pixel 3 smartphone, with comparable accuracy on the GLUE benchmark.

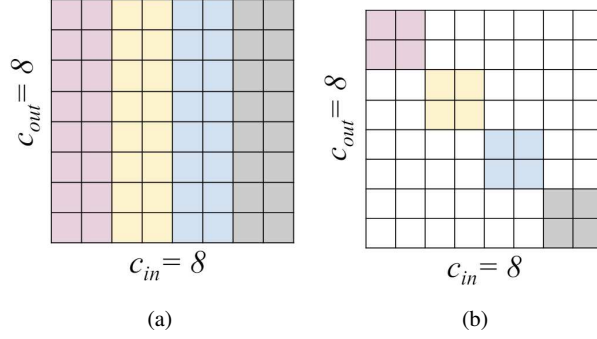


Fig. 2.8: (a) Convolutions and (b) grouped convolutions in SqueezeBERT (Iandola et al., 2020).

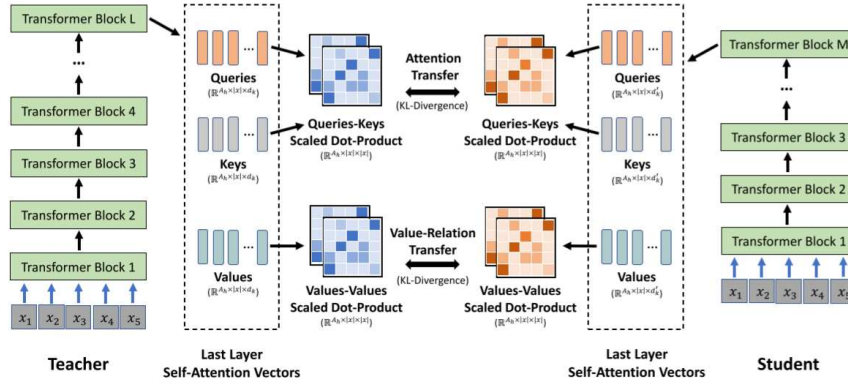


Fig. 2.9: Deep Self-Attention Distillation of MiniLM (Wang et al., 2020b).

SqueezeBERT also mirrors BERT’s pre-training on the BooksCorpus and English Wikipedia, utilizing MLM and SOP as objectives. To further enhance performance on GLUE, SqueezeBERT applies distillation during fine-tuning, focusing on capturing the final-layer logits of a BERT teacher model, thereby balancing efficiency with effectiveness on natural language understanding tasks.

MiniLM. MiniLM (Wang et al., 2020b) also adopts a knowledge distillation approach. By limiting the distillation process to the last teacher level, MiniLM aims to alleviate the difficulty in performing a layer mapping between the teacher and the student and to make the number of the student layers more flexible. Moreover, the model introduces the scaled dot-product between values, in addition to the scaled dot-product of queries and keys, to transfer the value relations (Fig. 2.9), thus achieving a deeper mimicking of the teacher as well as introducing more knowledge about word dependencies. Optionally, an intermediate-size student model (“the teacher

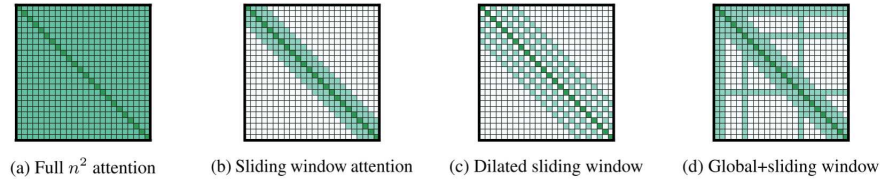


Fig. 2.10: Longformer attention pattern (Beltagy et al., 2020).

assistant”) can be interposed between the teacher and the student in the knowledge distillation process to alleviate the size gap between student and teacher and improve the model performance of smaller students. The knowledge of the teacher is hence distilled into the teacher assistant, which guides the training of the student.

MiniLM is trained for mono and multi-lingual tasks. In the mono-lingual setting, the teacher is the BERT-*base*, while for the multi-lingual setting the teacher is XLM-RoBERTa-*base*. The model proves to outperforms DistilBERT on SQUAD v2.0 and several tasks of GLUE, and achieves competitive performance on XNLI w.r.t. mBERT.

Longformer. Longformer (Beltagy et al., 2020) is specifically designed to handle long text sequences.¹⁶ Its attention mechanism combines local attention and global attention (Fig. 2.10), reducing the standard self-attention operations and scaling complexity from quadratic to linear with respect to the sequence length. Local attention is necessary to get contextual representations. In particular, the model uses a sliding window that can be dilated, i.e., it can have gaps of positions. For each attention head, different dilation configuration and different window size are provided. Different dilatation allows some heads to focus on local context through compact windows, while the others can focus on more distant contexts through enlarged windows. Global attention is added on few elected tokens and is crucial to get full sequence representations that are required for many NLP tasks (e.g., Question-Answering, document classification). How to select tokens for global attention is task-specific; for example it may consider the [CLS] token in classification tasks or all query tokens in question answering tasks. For auto-regressive tasks, the model has small non-dilated windows in lower layers, and increases the size and dilation (only on 2 heads) moving up to higher layers. The model is evaluated on the character-level text8 and enwik8 datasets.

Longformer is pre-trained with MLM objective, starting from a RoBERTa checkpoint with some minimal non-architectural changes to add the new attention pattern. Besides English Wikipedia and Bookcorpus, the pre-training includes a portion of RealNews, consisting of documents longer than 1200 tokens, and a portion of Stories corpus. The tokenizer is the same as RoBERTa, as well as the vocabulary, except for the introduction of special tokens for question answering data. The sliding window is set with a size of 512 and extra position embeddings are added (up to 4096,

¹⁶ <https://github.com/allenai/longformer>

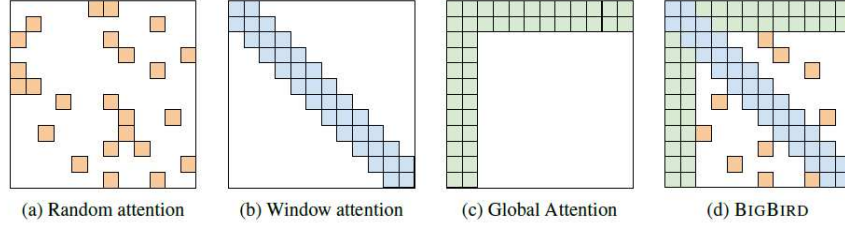


Fig. 2.11: BigBird attention pattern (Zaheer et al., 2020).

while RoBERTa has a maximum of 512) to support longer documents. Thereby, the model can handle up to 8 times longer documents than BERT and RoBERTa. This leads Longformer to outperform RoBERTa in tasks where long-document datasets are used, such as WikiHop and TriviaQA for question-answering (Section 2.4.1), OntoNotes (Pradhan et al., 2012) for co-reference resolution. Longformer is provided in two model sizes: Longformer-*base* (12 layers, 12 attention head, 768 hidden size) and Longformer-*large* (24 layers, 16 attention head, 1024 hidden size).

An encoder-decoder variant of Longformer, dubbed *Longformer-Encoder-Decoder* (LED), handles long text sequences in sequence-to-sequence tasks. The attention pattern of Longformer is applied to the Encoder part, while the Decoder part contains the full attention of standard Transformer. LED parameters are initialized following BART architecture in terms of number of layers and hidden sizes, with position embedding length extended to 16K tokens and enhanced initialization, in order to deal with longer texts. The model is released in two versions: LED-*base* (6 layers, 12 attention head, 768 hidden size) and LED-*large* (12 layers, 16 attention head, 1024 hidden size).

BigBird. BigBird (Zaheer et al., 2020) is a Transformer-based model that, like Longformer, can reduce the computational and memory cost by passing from a full-attention mechanism to a sparse-attention mechanism.¹⁷ In particular, the model uses a combination of three attention patterns: random attention, sliding window attention and global attention (Fig. 2.11). In the random attention, each query focuses on a set of random keys. In the sliding window attention, each query focuses on a set of their right and left neighboring tokens selected by a sliding window. In the global attention, some tokens (from the text or special tokens like [CLS]) attend to the whole sequence. Formally, given an input sequence $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n) \in \mathbb{R}^{n \times d}$, the generalized attention mechanism for $\mathbf{x}_i$ is as follows:

$$Att(\mathbf{X})_i = \mathbf{x}_i + \sum_{h=1}^H softmax((x_i \mathbf{W}_Q^{(h)}) \cdot (\mathbf{X}_{N(i)} \mathbf{W}_K^{(h)})^\top) \cdot (\mathbf{X}_{N(i)} \mathbf{W}_V^{(h)}), \quad (2.7)$$

¹⁷ <https://github.com/google-research/bigbird>

where H is the number of attention’s heads, $\mathbf{W}_Q^{(h)} \in \mathbb{R}^{d \times m}$, $\mathbf{W}_K^{(h)} \in \mathbb{R}^{d \times m}$ and $\mathbf{W}_V^{(h)} \in \mathbb{R}^{d \times d}$ are query, key and value matrices of the h -th head, respectively, $N(i)$ denotes the set of keys attended by the query i , while x_i and $\mathbf{X}_{N(i)}$ are the vector representations of query i and its keys, respectively. Like Longformer, this type of attention mechanism can reduce the Transformer complexity from quadratic to linear in the number of tokens, allowing the model to scale to much longer sequences in input, up to 8 times longer than previous Transformer-based models. The authors stated that sparse attention mechanisms have the same power and expressiveness as full-attention mechanisms, demonstrating that, on the one hand, when used in standalone encoders like BERT they are universal approximate methods of sequence-to-sequence functions and, on the other hand, that the sparse encoder-decoder models are Turing-complete under standard assumptions regarding precision.

In the pre-training settings, BigBird follows BERT and RoBERTa to create base and large models. It is pre-trained using the MLM objective and four corpora (Books, CC-News, Stories and Wikipedia), starting from a RoBERTa checkpoint. The model is available in two base versions: BigBird-ITC-*base* (in which only internal tokens in the text attend global attention) and BigBird-ETC-*base* (in which only the external token like [CLS] attends global attention). Both versions have 12 layers, 12 attention heads and hidden size of 768. Each version has its own setting regarding the number of tokens with global attention and random attention as well as the window size for local attention. Two large versions are also available: BigBird-ITC-*large* and BigBird-ETC-*large*, both with 24 layers, 16 attention heads and 1024 hidden size.

BigBird has been fine-tuned on Question-Answering challenging datasets (Natural Questions, HotpotQA-distractor, TriviaQA-wiki, WikiHop, cf. Section 2.4.1) outperforming competitors like Longformer and RoBERTa. In document classification and GLUE tasks, BigBird proves to be more advantageous than existing methods when longer documents and fewer training examples are being used (e.g., Arxiv, cf. Section 2.4.1). Moreover, like Longformer, an encoder-decoder setup of BigBird is also proposed, in which the sparse attention mechanism is introduced only in the Encoder side.

2.2.2 Encoder-decoder models

T5. Text-to-Text Transformer (T5) (Raffel et al., 2020) provides a unified approach in which one single model can be used for every task with the same objective, the same decoding process and the same training procedure.¹⁸ To specify the current task, a task-specific prefix is added to the input text before feeding it to the model. The architecture of T5 is quite similar to the original Transformer, with few changes regarding the normalization layer and the position embedding scheme. The baseline is similar to BERT-*base* in terms of size and configuration, for both the encoder and the decoder, but with 220 million parameters in total, which are twice BERT-*base*

¹⁸ <https://github.com/google-research/text-to-text-transfer-transformer>

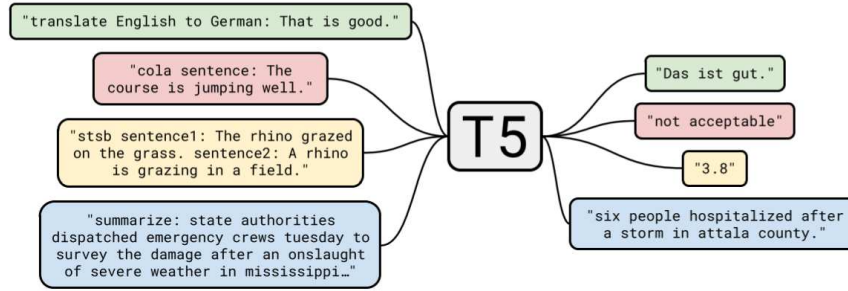


Fig. 2.12: T5 diagram (Raffel et al., 2020). Every task is considered as text. A task-specific prefix is added to the input text to specify the current task.

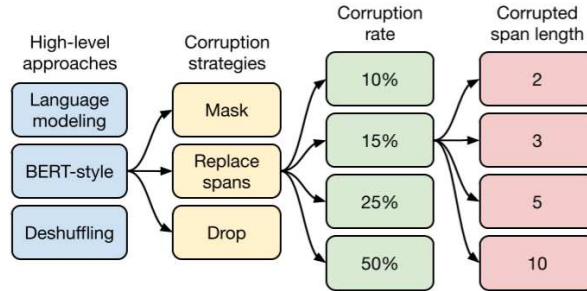


Fig. 2.13: A flowchart of the choices about unsupervised objective in T5 (Raffel et al., 2020). After accurate experimentation, the authors opted for a BERT-style approach, with spans replacing as corruption strategies, 15% corruption rate and average corrupted span length chosen in $\{2, 3, 5, 10\}$.

parameters, as T5 contains two layer stacks instead of one. The text encoding method is SentencePiece (Kudo and Richardson, 2018) with a multi-lingual vocabulary of 32K wordpieces.

The training objective for pre-training and fine-tuning is a maximum likelihood objective using teacher forcing and a cross-entropy loss. More specifically, T5 utilizes a BERT-style denoising objective that randomly samples and drops out spans of tokens (loosely inspired by SpanBERT), with corruption rate of 15% of tokens and varying the span length, and the aim is to predict the dropped-out tokens. The span of tokens is replaced with only one special token. Fig. 2.13 shows the experimentation choices for the pre-training objective.

For the model architecture, the authors reviewed and compared several Transformer variants (encoder-decoder, language model, prefix LM, see Fig. 2.14) and eventually stated that the original encoder-decoder is the most suitable form for the text-to-text framework. In addition, they found that a small domain-specific unla-

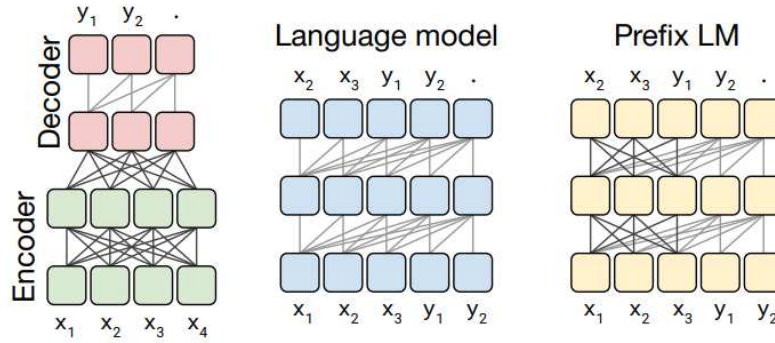


Fig. 2.14: Illustrations of the Transformer variants reviewed and compared in (Raffel et al., 2020). On the left, the encoder-decoder architecture which uses the fully-visible masking when the encoder is involved, while the causal attention is used inside the decoder. In the middle, a standard language model which uses a causal masking. On the right, the Prefix LM which involves fully-visible masking over the input and causal masking for the output.

beled dataset, repeated many times in the pre-training, can degrade performance in some downstream tasks. For this reason they adopted a new massive dataset called Colossal Clean Crawled Corpus (C4), consisting in hundreds of gigabytes of clean, high-quality text picked from the web. Using C4 in the pre-training phase led the model to be flexible enough to be fine-tuned to a variety of tasks while obtaining state-of-the-art results. In this respect, T5 is provided in five versions: T5-*base* (i.e., the baseline, 12 layers, 12 attention heads, 768 hidden size, 220M parameters), T5-*small* (6 layers, 8 attention heads, 512 hidden size, 60M parameters), T5-*large* (24 layers, 16 attention heads, 1024 hidden size, 770M parameters), T5-3B (24 layers, 32 attention heads, 1024 hidden size, 3B parameters), T5-11B (24 layers, 128 attention heads, 1024 hidden size, 11B parameters). It should be noted that C4 is regarded as big enough to allow the exploration of the effect of scaling up the amount of pre-training without overfitting, in particular when training on more data, with larger versions of T5 or ensemble of models. However, small models can still be useful when limited computational resources are available in fine-tuning.

T5 has been tested on various downstream tasks, such as classification, question-answering, translation and summarization. When fine-tuning on GLUE (as well as on SuperGLUE), the datasets of all the benchmarks are concatenated so that the model is fine-tuned just once, using small batch size in order to mitigate overfitting in low-resource tasks. For task with long output sequences, the performance improved with the use of beam search, while the other task are reported with greedy decoding. T5 achieves state-of-the-art result in GLUE and overcomes RoBERTa and AIBERT in SuperGLUE and SQuAD.

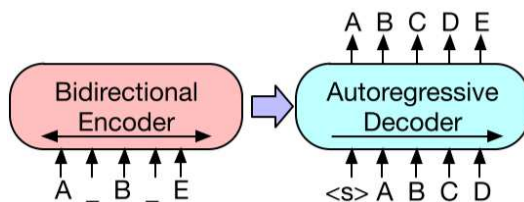


Fig. 2.15: BART architecture (Lewis et al., 2020a).

A multi-lingual variant of the model, called *mT5*, is also available (Xue et al., 2021).¹⁹ It was pre-trained on a Common Crawl dataset, called mC4, based on 101 languages. The architecture and training procedure is quite similar to T5. The model is available in four sizes: *mT5-small* (300M parameters), *mT5-base* (580M parameters), *mT5-large* (1.2B parameters), *mT5-xl* (3.7B parameters), *mT5-xxl* (13B parameters). It was tested on several XTREME (Hu et al., 2020) tasks, including XNLI and XQuAD (Artetxe et al., 2020). For each task, three fine-tuning settings are considered: zero-shot, translate-train, in-language multitask. mT5 achieved strong performance on each task, overcoming previous models like m-BERT.

BART. BART (Lewis et al., 2020a) is a denoising auto-encoder implemented as a sequence-to-sequence model, combining bidirectional and auto-regressive Transformers.²⁰ More precisely, it consists of a bidirectional encoder, following BERT, and a left-to-right auto-regressive decoder, following GPT (Fig. 2.15). The architecture follows the standard sequence-to-sequence Transformer, except for the activation function and the initialization of the parameters which are set up according to the GPT instructions. It is also related to BERT except for the decoder layers, in which there is an additional cross-attention with the final hidden layer of the encoder, and the absence of the additional feed-forward network before word prediction.

BART is available in a base size (6 layers, 12 attention heads,²¹ 768 hidden size) and a large size (12 layers, 16 attention heads,²² 1024 hidden size).

BART is pre-trained by first corrupting input text and then training the model to get back the original document. A key advantage of BART is the noising flexibility, i.e., it allows for arbitrary type of document corruption, including changing its length. This makes it possible to mask single tokens (like in the MLM task) as well as the entire input. The best results have been obtained shuffling the order of the original sentences and masking spans of text. This approach generalizes both MLM and NSP pre-training tasks. The masking method is similar to SpanBERT, but in this case the sampling is obtained using a different distribution (Poisson) and each span is replaced with a single [MASK] token.

¹⁹ <https://github.com/google-research/multi-lingual-t5>

²⁰ <https://github.com/pytorch/fairseq>

²¹ from <https://huggingface.co/facebook/bart-base/blob/main/config.json>

²² from <https://huggingface.co/facebook/bart-large/blob/main/config.json>

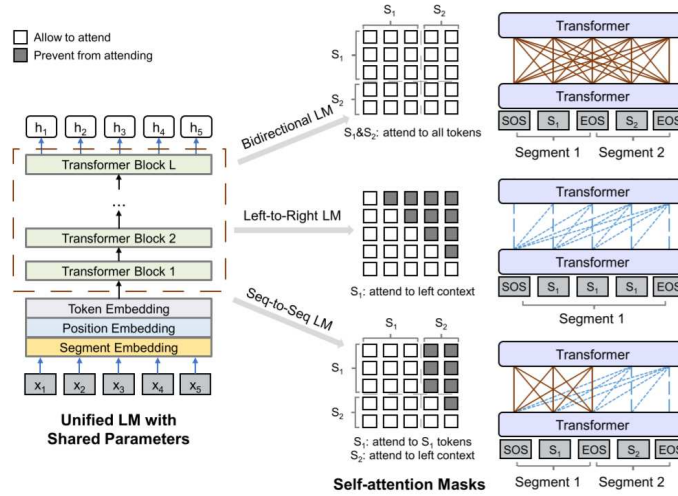


Fig. 2.16: UniLM Pre-training (Dong et al., 2019). A unique model with shared parameters across all the pre-training objectives is employed. Self-attention masks control the context for the prediction, based on the objective.

BART can be fine-tuned for a wide range of downstream tasks, such as sequence classification, token classification, sequence generation and machine translation. Documents are tokenized with BPE, the corruption rate is the 30% of tokens for each document, and all sentences are permuted. BART obtains similar results to RoBERTa in GLUE and SQuAD benchmarks, and achieves new state of the art compared to previous models in abstractive dialogue, summarization and abstractive Question-Answering.

In Liu et al. (2020b), a multi-lingual version of BART is provided, called *mBART*. It is the application of BART to large-scale monolingual corpora across many languages. mBART represents the first multi-language denoising pre-training method for a complete sequence-to-sequence model, which makes it possible a direct fine-tuning for supervised and unsupervised machine translation without task-specific modifications. mBART recognizes the language through a language id token ($\langle LID \rangle$) placed at the beginning of the sentence. The authors built several versions of the model such as mBART25 (pre-trained on 25 languages) and mBART06 (pre-trained on 6 European languages).

UniLM. UniLM (Dong et al., 2019) is a Transformer model pre-trained on uni-directional, bidirectional and sequence-to-sequence prediction tasks.²³ It uses self-attention masks to control the context of the prediction, thus enabling a single architecture for uni/bidirectional and sequence-to-sequence approaches. The parameter sharing across the different language modeling objectives allows the model to

²³ <https://github.com/microsoft/unilm>

learn more general text representations (Fig. 2.16). Different segment embeddings are used based on the pre-training task. Two special tokens, namely [SOS] and [EOS], are added at the beginning and at the end of each segment in the input, respectively. Such tokens play an important role for both natural language understanding and generation downstream tasks.

UniLM is initialized and pre-trained following BERT-*large*. For language understanding tasks, it is fine-tuned as a bidirectional encoder, where the contextual [SOS] embedding represents the input encoding. For language generation tasks, it is fine-tuned on a sequence-to-sequence task, where the [EOF] token can be used to learn when to stop the decoding process. UniLM is evaluated on several benchmarks, such as GLUE, SQuAD and CNN/DailyMail. Results demonstrate that UniLM is competitive against BERT on GLUE and better than BERT on SQuAD 2.0 development set, and outperforms previous state-of-the-art model on several generative tasks (CNN/DailyMail included).

Inspired by UniLM, Bao et al. (2020) propose *UniLMv2*, which is pre-trained on bidirectional and sequence-to-sequence tasks in a unified manner with a pseudo-masked language modeling (PMLM) procedure. The model parameters and the context encodings are shared across the two tasks. Given a corrupted input with masked tokens, UniLMv2 uses standard masks and an autoencoding approach to learn the relations between masked tokens and the context. In addition, it uses pseudo masks and a partially autoregressive modeling approach to learn the relations between masked spans or tokens (token-to-token, token-to-span, and span-to-span relations). The accessible context for each token in the partially autoregressive modeling is controlled according to a factorization order. Similarly to Yang et al. (2019), for each text in input a random factorization order is sampled. The model predicts one token or a span at each factorization step (for this reason it is partially autoregressive). The masks used in autoencoding pre-training provide global masking information, which can be used to access the position embeddings in the factorization steps. A standard mask, named [M], is employed to corrupt input tokens in the autoencoding pre-training. To handle factorization steps in the partially autoregressive pre-training, the [P] masks are appended to the input without replacing the original tokens. Such masks have the same position embedding of the corresponding tokens and their final embeddings are used for the predictions.

UniLMv2 size is similar to that of BERT-*base* (i.e., 12 layers with 12 attention heads and 768 hidden size). Fine-tuned and evaluated on several language understanding and generation tasks, UniLMv2 outperforms the base version of BERT, RoBERTa and XLNet on SQuAD and eight tasks of GLUE. On CNN/DailyMail, it proves to be more effective than other TLMs such as UniLM and MASS (base version).

MASS. MASS (Song et al., 2019) introduces a masked sequence-to-sequence pre-training, in which the encoder maps a sequence to several masked fragments, each composed of consecutive tokens, and the decoder predicts the masked fragments. The encoder and decoder are pre-trained simultaneously and in two stages: the encoder is forced to understand the meaning of the masked tokens to predict the tokens on

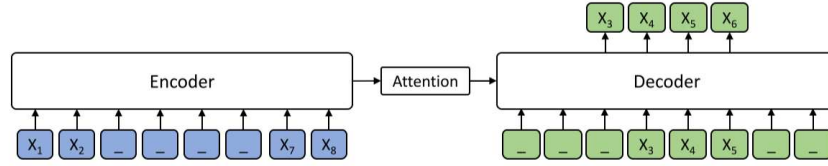


Fig. 2.17: MASS architecture (Song et al., 2019).

the decoder side, while the decoder has the input tokens completely masked, so it relies primarily on the representation of the input passed by the encoder. More precisely, the decoder predicts the first token of the masked fragment based solely on the encoder input representation, while subsequent tokens are predicted based on the encoder representation and the previously predicted fragment tokens (Fig. 2.17). This enables the model to learn generating sequences during pre-training.

Available in a base size (6 layers, 12 attention heads, 768 hidden size) and a middle size (6 layers, 16 attention heads, 1024 hidden size), MASS is pre-trained on the WMT monolingual corpus,²⁴ and fine-tuned on several benchmarks for machine translation (WMT14 English-French, WMT16 English-German and WMT16 English-Romanian), text summarization (Gigaword),²⁵ and conversational response generation (Cornell movie dialog (Danescu-Niculescu-Mizil and Lee, 2011)).²⁶

2.2.3 Decoder-only models

XLNet. XLNet (Yang et al., 2019) exploits the advantages of autoregressive language modeling and autoencoding denoising. Instead of masking the input, XLNet enables bidirectionality by permuting the order of context factorization, and is trained to predict a token in an auto-regressive manner, but the prediction order is randomized and not sequential. This means that the model predicts a token based on previously predicted tokens, whose positions in the text may be before or after the token to be predicted. In this way, the model is forced to learn predicting the target with a randomly ordered context, thus enabling the bidirectionality. XLNet can be employed for tasks that include text of any length. The architecture is inspired by Transformer-XL (Dai et al., 2019), from which it takes the segment-level recurrence mechanism and relative positional encoding scheme. The segment-level recurrence mechanism consists in caching the representations computed for the previous segment so that they can be used as extended context for processing the next segment, thus increasing the maximum perceivable distance between dependencies. The relative positional

²⁴ <https://www.statmt.org/wmt16/translation-task.html>

²⁵ <https://github.com/harvardnlp/sent-summary>

²⁶ <https://github.com/suriyadeepan/datasets/tree/master/seq2seq/>

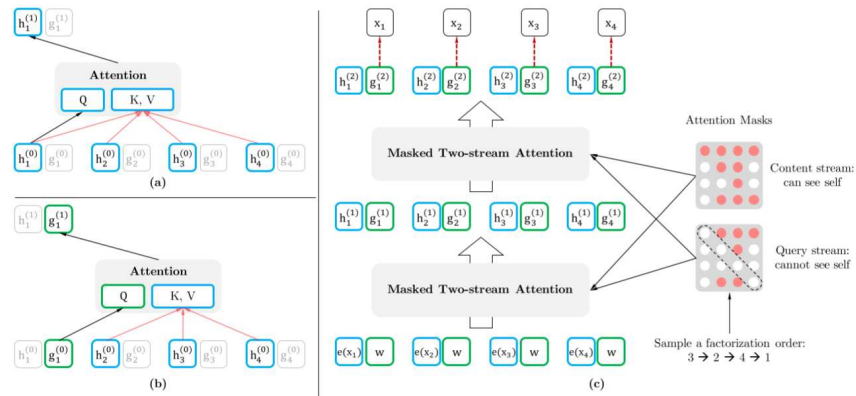


Fig. 2.18: XLNet (a) content stream attention, (b) query stream attention, and (c) two-stream self-attention (Yang et al., 2019).

encoding scheme adapts the input positional encodings to the segment-level recurrence mechanism. XLNet also introduces a new self-attention mechanism, called two-stream self-attention, which consists in a standard self-attention (content stream attention) and a query stream attention that mimics the masking behavior by obscuring the content of the token to be predicted but keeping its positional encoding visible. Fig. 2.18 shows the difference between content stream attention and query stream attention: the former has access to the representations of both the context and the query token, while the latter has access only to the contents of previous tokens and to the positional encoding of the query.

XLNet is pre-trained with various pre-training corpora, which also include the BERT pre-training corpora. Two sizes of the model are available: XLNet-*large* (24 layers, 16 heads and 1024 hidden size) and XLNet-*base* (12 layers, 12 heads and 768 hidden size). XLNet achieves substantial improvements over BERT and RoBERTa on various benchmarks, such as RACE, SQuAD v2.0 and GLUE.

GPT. GPT (Radford et al., 2018) is the first model of the Generative Pre-trained Transformer family developed by OpenAI.²⁷ Based on a Transformer decoder, with 12 layers, 12 attention heads, and 768 hidden layer size, GPT is an autoregressive model pre-trained on the BookCorpus dataset with a *causal language modeling objective* (CLM), i.e., to predict the next token given the left-side context. The model uses a BPE vocabulary size of 40K and maximum sequence length of 512 tokens. To address the downstream tasks in the fine-tuning phase, the inputs of the different tasks are converted into token sequences through the use of special tokens, in order to be processed by the model (Fig. 2.19).

The model achieved better performance than previous competitors at that time on several tasks, like text classification (SST-2 and COLA, cf. Section 2.4.1), textual

²⁷ <https://openai.com/>

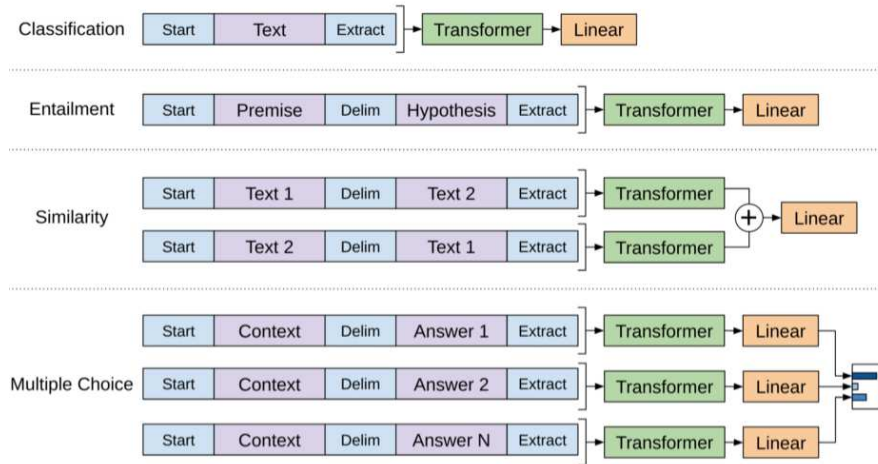


Fig. 2.19: Task-specific input conversion to token sequences in GPT (Radford et al., 2018).

entailment (such as SNLI, MultiNLI and QNLI, cf. Section 2.4.1), question answering (such as RACE, cf. Section 2.4.1), semantic similarity (such as STS Benchmark, cf. Section 2.4.1).

GPT-2. GPT-2 (Radford et al., 2019) is the successor of GPT, conceived to perform downstream tasks in a zero-shot learning setting. Like GPT, GPT-2 is pre-trained on a CLM task, but with WebText—a significantly larger dataset containing around 8 million documents (40GB of text) sourced from web pages that have been human-curated for quality. The architecture of GPT-2 is largely based on GPT, but it has more than one order of magnitude of additional parameters (1.5B parameters). The model’s context window is also expanded to 1024 tokens, with a vocabulary of 50k tokens. The model architecture incorporates pre-activation residual connections and layer normalization modifications for enhanced stability and generalization, with a larger batch size (512) than GPT to optimize learning efficiency.

To address downstream NLP tasks, the output of GPT-2 is conditioned on the input and task type. That is, in addition to the text, the model receives as input indications of the task to be performed. For example, to induce the summarization behavior, the text “TL;DR:” is added after the text to be summarized.

GPT-2 has shown to outperform previous state-of-the-art models in several language modeling datasets (such as LAMBADA (Paperno et al., 2016), enwik8 and text8).²⁸ On reading comprehension tasks— particularly machine translation (WMT-14 Fr-En), summarization (CNN/DailyMail, cf. Section 2.4.1) and question-answering (SQuAD, cf. Section 2.4.1)—the model proves to be comparable to supervised baselines in a zero-shot setting.

²⁸ enwik8 and text8 are available at <http://www.mattdmahoney.net/dc/text.html>

GPT-3, GPT-4 and ChatGPT. GPT-3 (Brown et al., 2020) significantly increases the size of GPT models to 175B parameters, about two orders of magnitude more than the predecessor GPT-2. The pre-training approach is similar to GPT-2, but the data are larger and more heterogeneous, as they contain the CommonCrawl dataset and high-quality reference corpora, including an expanded version of WebText, internet-based books corpora and the English Wikipedia. GPT-3 is available in different sizes, such as: *ada* (350M parameters), *babbage* (1.3B parameters), *curie* (6.7B parameters), *davinci* (175B parameters). The model is evaluated in zero-shot, one-shot and few-shot settings. In zero-shot setting, a description of the task to be performed is given to the model, whereas for one-shot and few-shot settings one or few examples of how the task is to be executed are provided to the model at inference time, after which the model can be prompted to execute the task on a new example.

GPT-3 is evaluated on several traditional language model benchmarks, e.g., LAMBADA, achieving better results compared to the previous fine-tuned state-of-the-art models in many cases. Results also show that scaling up the model greatly improves the performance in task-agnostic/few-shot setting, becoming even competitive with fine-tuned competitors in some NLP tasks. For example, on question-answering tasks, GPT-3 gets mixed results, with the zero-shot model outperforming the fine-tuned T5-11B on TriviaQA but being outperformed on the Natural Questions benchmark. In machine translation tasks, GPT-3 exceeds previous unsupervised competitors when translating to English, but underperforms when translating in the opposite direction. On SuperGLUE, the few-shot model obtains higher performance than the fine-tuned BERT-*large* on four tasks and almost reaches the state-of-the-art at that time (a fine-tuned 11B-parameter model) in two tasks. Regarding text summarization, some critical issues on the GPT-3 samples are found, e.g., the presence of semantically document-level repetition and loss of coherence on long passages.

GPT-3.5 is an improved version of GPT-3 which includes an enhanced fine-tuning with Reinforcement Learning from Human Feedback (RLHF), a training technique that uses feedback from human trainers as reward signals to guide the model toward preferred responses. It consists of a fine-tuning phase, in which human trainers review and score the outputs of the model, which is used to create a reward model. The model is then fine-tuned using Proximal Policy Optimization (PPO), a reinforcement learning algorithm which employs the reward model as an estimate for the true reward function (i.e., the human preference) and the model as the policy to optimize. (Schulman et al., 2017).

GPT-4 (OpenAI, 2023) is a large, multi-modal model capable of handling both text and image inputs. Unlike GPT-3, it was trained using a vast mix of public data and licensed sources, which allow for more extensive knowledge coverage, particularly in professional, academic, and multi-lingual contexts. It supports a significantly longer context window than GPT-3 (up to 32K tokens, whereas GPT-3 supports a maximum of 4,096 tokens). Similar to GPT-3.5, GPT-4 incorporates RLHF to align its responses with human expectations, but it uses a rigorous pipeline for emphasizing safety, factual accuracy, and content moderation to reduce harmful or biased outputs. On benchmarks like MMLU, GPT-4 consistently outperforms GPT-3 and GPT-3.5, demonstrating superior language understanding and multi-lingual

capabilities. Moreover, it reduces hallucinations by 19% points compared to GPT-3.5 and shows a stronger adherence to safety protocols and policies, with substantial reductions in producing disallowed or toxic content.

Perhaps the world’s most controversial AI language tool of our time, ChatGPT²⁹ is a GPT-based model trained to interact with humans, using either GPT-3.5 or GPT-4 as its core. In comparison to standalone GPT-3.5 and GPT-4, ChatGPT models are fine-tuned through an extended RLHF process, which emphasizes conversational alignment. ChatGPT shares the underlying architecture with InstructGPT (Ouyang et al., 2022), which is specifically designed and fine-tuned for generating detailed instructions given a prompt, but further adapts this framework for sustained, interactive dialogue. The result is a model that allows for human dialogue on a wide variety of topics, answering follow-up questions while maintaining a conversation flow.

GPT-NeoX-20B. GPT-NeoX-20B (Black et al., 2022) is an autoregressive language model with 20B parameters and trained on Pile (Gao et al., 2021).³⁰ The GPT-NeoX-20B architecture largely follows GPT-3 with some differences. It applies the rotary positional embeddings (Su et al., 2021) on the first 25% of embedding vector dimensions, which consist in rotating the embedding space in such a way that the attention between two tokens is linearly dependent on their distance in the text. Moreover, GPT-NeoX-20B computes attention and feed-forward layers in parallel, whose outcomes are subsequently summed. The BPE tokenizer and the vocabulary size is similar to GPT-2, but the tokenizer is trained on the Pile dataset and, unlike GPT-2, handles the presence of prefix spaces and repeated space tokens.

GPT-NeoX-20B is evaluated on several benchmarks (e.g., LAMBADA and TriviaQA), as well as mathematical and knowledge-based tasks. Results demonstrate that the model reaches higher performance than the similarly sized GPT-3 and fairseq models (Artetxe et al., 2022) when evaluated in five-shot setting.

mGPT. Shliazhko et al. (2022) propose a GPT-like model in two size versions (1.3B and 12B parameters) for multi-lingual tasks.³¹ mGPT aims to reproduce the architecture of GPT-3 starting from a GPT-2 implementation, based on the information described in Brown et al. (2020). The resulting models are trained on 60 languages through Wikipedia and Colossal Clean Crawled (C4) corpora, and evaluated under zero-shot and few-shot settings (as in Brown et al. (2020)) on several multi-lingual benchmarks, including text classification, text generation and sequence labeling. Overall, results show that larger models correspond to better generation abilities for all given languages. In addition, experimental evaluation on knowledge probing indicates a certain ability of mGPT to preserve factual knowledge.

LLaMA. LLaMA (Touvron et al., 2023a) is a collection of foundational language models designed for efficiency and accessibility. It is available in multiple versions, ranging from LLaMA-7B (7 billion parameters) to LLaMA-65B (65 billion parameters). Each version is pre-trained on a mixture of high-quality, publicly available

²⁹ <https://openai.com/blog/chatgpt>

³⁰ <https://github.com/EleutherAI/gpt-neox>

³¹ <https://github.com/ai-forever/mgpt>

datasets, including a significant proportion of English CommonCrawl (67%), C4, GitHub, Wikipedia, and Books3 (about 1.4 trillion tokens in total).

LLaMA incorporates architectural innovations to optimize efficiency. It uses *Swish-Gated Linear Unit* (SwiGLU) as its activation function, which captures complex patterns while minimizing computational demands. Instead of absolute positional embeddings, it uses the *Rotary Positional Embeddings* (RoPE), which encode relative positional information more flexibly and efficiently. For improving training stability, it employs *Root Mean Square Normalization* (RMSNorm), which normalizes inputs based on the root mean square of activations.

LLaMA was tested on question-answering task with Natural Questions and TriviaQA benchmarks. In both cases, the model demonstrated strong zero-shot and few-shot performance. The reading comprehension’s abilities of the model are evaluated on the RACE dataset, showing its effectiveness in zero-shot scenarios. In both question-answering and reading comprehension benchmarks, LLaMA-13B outperforms GPT-3 (175 billion parameters), despite having roughly one-tenth of its size. LLaMA’s mathematical problem-solving capabilities were assessed with the MATH (Hendrycks et al., 2021c) and GSM8k (Cobbe et al., 2021) datasets, covering middle and high school-level math problems, in which the largest model outperformed task-specific competitors, despite not being fine-tuned on mathematical data. In benchmarks such as MMLU (Hendrycks et al., 2021a), which tests a model’s zero-shot and few-shot performance across diverse domains including humanities, STEM, and social sciences, LLaMA achieves competitive results.

LLaMA-2. LLaMA-2 (Touvron et al., 2023b) introduces a series of models with significant advancements in performance, efficiency, and safety respect to LLaMA. Both its standard version (LLaMA-2) and its chat-optimized version (LLaMA-2-Chat, similarly to ChatGPT) range from 7 billion to 70 billion parameters.

In terms of architecture, it incorporates some improvements w.r.t. LLaMA, such as the support of a longer maximum context length (4096 tokens, while LLaMA can handle 2048 tokens maximum) and the introduction of *Grouped-Query Attention* (GQA), which enables more efficient attention computation, particularly for larger models.

LLaMA-2 was trained on an extensive, diverse corpus of publicly available data, significantly expanding on the data mix used for LLaMA. This dataset includes various high-quality sources across multiple domains, further enhanced by data filtering processes to minimize low-quality content such as highly repetitive or non-informative text. Additionally, personal or sensitive data have been discarded, aiming to improve the model’s reliability and adherence to privacy standards. LLaMA-2-Chat is obtained starting from LLaMA-2 and applying a supervised fine-tuning step followed several iteration of RLHF techniques, i.e. rejection sampling and PPO.

On popular benchmarks, LLaMA-2 achieves superior results compared to LLaMA across tasks like MMLU and TriviaQA, showing marked improvements in both language understanding and generation capabilities. These gains make it competitive with closed-source counterparts like GPT-4 across several domains, including code generation and complex question-answering tasks.

LLaMA-3. LLaMA-3 (Dubey et al., 2024) is currently the newest generation of LLaMA models, bringing substantial advancements in size, multi-lingual capabilities, extended context processing and overall versatility. Like LLaMA-2, LLaMA-3 uses a dense Transformer architecture but scales up to 405B parameters, significantly larger than LLaMA-2’s 70B model. It also adopts grouped-query attention, an expanded vocabulary for multi-lingual performance, and an extended RoPE base frequency for handling up to 128K context’s tokens.

Its training dataset includes 15.6 trillion tokens—over eight times LLaMA-2’s data volume—with comprehensive filtering and quality control to enhance performance. The data mix includes multi-lingual, mathematical, and reasoning content to support its broader capabilities, a strategy similar to GPT-4’s diverse training dataset but with a stronger emphasis on quality refinement.

The training of the model involved a combination of techniques, including initial pre-training, long-context pre-training, and a final annealing phase. This multi-phase approach helps the model learn to process and retain large amounts of information. The post-training includes supervised fine-tuning, rejection sampling, and Direct Preference Optimization (DPO) to align with human preferences, similar to GPT-4’s RLHF process.

LLaMA-3 demonstrates competitive results with GPT-4 on benchmarks, particularly in multi-lingual and long-context tasks.

Vicuna. Vicuna³² is an open-source chat-based model developed by fine-tuning LLaMA-13B on user-shared conversations from ShareGPT.³³ While it does not introduce architectural changes to the original LLaMA model, its fine-tuning is specifically designed for multi-turn conversational tasks.

Vicuna’s performance is assessed using GPT-4 as evaluator—scoring the responses based on helpfulness, accuracy, relevance, and detail—across multiple categories, such as roleplay, coding, and mathematics. It achieves approximately 90% of ChatGPT’s quality and is rated comparably to ChatGPT in 45% of cases, based on helpfulness, accuracy, and detail. In 90% of the questions, GPT-4 rated Vicuna higher than LLaMA-13B, with notable advantages in instruction-following, multi-turn dialogue handling, and accuracy in complex or technical tasks such as coding and math.

Mistral. Mistral (Jiang et al., 2023) is an open-source model based on a Transformer architecture with some key modifications. To enable effective handling of longer sequences (up to 131K tokens) with less memory usage, it employs a *Sliding Window Attention* (SWA) mechanism—which allows tokens to attend to a window of previous tokens—and *Rolling Buffer Cache*—which limits cache growing. Moreover, it uses pre-fill and chunking strategies to optimize prompt processing and reduce latency during generation. Although with fewer parameters (7B), it consistently outperforms LLaMA-2 13B on a wide range of benchmarks, including commonsense reasoning, world knowledge, reading comprehension, math, and code generation.

³² <https://lmsys.org/blog/2023-03-30-vicuna/>

³³ <https://sharegpt.com/>

An instruction version of the model, Mistral 7B-Instruct, is fine-tuned open, publicly available datasets specifically created for instruction-following tasks and with conventional fine-tuning techniques. Mistral 7B-Instruct outperforms other 7B-parameters competitors, such as LLaMA-2 7B and Vicuna 7B, on *Machine Translation Benchmark* (MT-Bench), which assesses models based on their performance in understanding and following instructions. Moreover, it matches the performance of some 13B chat models, such as LLaMA 2 13B-Chat, in dialogue scenarios.

Falcon. Falcon (Almazrouei et al., 2023) provide a series of open-source language models (from 7B to 180B parameters), based a Transformer architecture with some modifications. Specifically, it employs the *multigroup attention* mechanism, which groups heads into sets where each group has its own set of key-value pairs, achieving a high level of memory efficiency (similar to the grouped-query-attention in LLaMA-2 and LLaMA-3) but with increased flexibility and control within each group. Similar to LLaMA models, it utilizes RoPE embeddings to enhance sequence-based task performance. Additional architectural adjustments, such as parallelizing attention and MLP layers and removing biases from linear layers, increase memory efficiency and enable faster inference.

Falcon was trained on a custom dataset called RefinedWeb, a high-quality, curated dataset sourced from the web and filtered to reduce redundancy and enhance relevance. Its largest version consistently outperforms GPT-3.5 and LLaMA-2 models on most commonsense and reasoning tasks, although it does not reach GPT-4, especially on benchmarks like MMLU.

Based on Mistral 7B model, NeuralChat³⁴ is a fine-tuned adaptation designed specifically for chatbot applications.

OPT. OPT (Zhang et al., 2022) is an open-source alternative to GPT-3 models, replicating its performance and sizes. Available in sizes from 125 million to 175 billion parameters, OPT provides scalable options for various deployment needs, from lightweight applications to large-scale implementations. It was trained on publicly available datasets, using a large mixture of internet-based text corpora. The models have been tested on several benchmarks such as commonsense reasoning, physical commonsense reasoning, language understanding (e.g. SuperGLUE), dialog and chat examples, hate speech detection, achieving performance comparable to GPT-3, especially in zero-shot and few-shot settings. Overall, OPT’s performance closely aligns with that of GPT-3 in zero-shot settings, while generally shows minor differences in one-shot and few-shot evaluations.

OpenAssistant. OpenAssistant (Köpf et al., 2023) provides a series of open-source TLMs aligned with human preferences. It is fine-tuned with RLHF and supervised learning on question-answering and dialogue demonstrations. The fine-tuning is carried out on the *OpenAssistant Conversations* data, which is an assistant-style conversation corpus constructed with human assistance and annotations. It supports a wide range of model sizes depending on the underlying model chosen (e.g. Falcon-40B). Experimental evaluation shows that, on various conversational and natural

³⁴ <https://huggingface.co/Intel/neural-chat-7b-v3>

language understanding benchmarks, the models trained on OpenAssistant Conversations improve over their respective base models and, in some cases, closely match ChatGPT.

Claude-3. Claude-3 (Anthropic, 2024) is a family of three multi-modal TLMs: Claude-3 Opus (the most advanced), Claude-3 Sonnet (balancing speed and proficiency), and Claude-3 Haiku (optimized for efficiency). Claude-3 is designed for enhanced conversational AI capabilities, expert-level reasoning, complex content creation, multi-lingual capabilities and reduced biases, in order to ensure safer and more responsible AI interactions. The architecture varies across the three versions, with Opus being the largest and most sophisticated. The models utilize *Constitutional AI*, which involves rule-based reinforcement learning to align outputs with ethical guidelines, reducing biases and avoid generating harmful or unethical content. The models are trained using a proprietary mixture of publicly available datasets, curated data from third-party sources, and in-house labeled data. The model family competes closely with models like GPT-4 and achieves state-of-the-art results in several benchmarks, such as reasoning and math (e.g. GSM8K (Cobbe et al., 2021), MATH (Hendrycks et al., 2021c)), multi-lingual reasoning (e.g. MMLU), coding (e.g. HumanEval) and vision capabilities (e.g. ChartQA) benchmarks.

2.3 Problems and Tasks

In this section, we provide a brief overview of key NLP problems and tasks categorized by application domain (general, legal, and psychosocial). Please note that this is not an exhaustive list of all NLP tasks, rather the focus is specifically on tasks and problems addressed during the Ph.D. course (cf. chapters 3-7) or mentioned in Section 2.2.

2.3.1 Problems and tasks for the general-domain

Text Classification. Text classification is a pivotal NLU task aimed at the automatic categorization of texts into predefined classes (or categories). Formally, a text sequence, represented as $x = [x_1, x_2, \dots, x_n]$ is assigned a label from a set $C = [c_1, c_2, \dots, c_m]$ corresponding to specific domains or topics, depending on the intended classification objective. This task can be approached in both supervised and unsupervised learning settings. In supervised classification, each input example is labeled, i.e. it is associated with one or more predefined classes. In contrast, unsupervised classification seeks to group examples without predefined labels, often through clustering techniques.

Text classification encompasses both *single-label* and *multi-label* paradigms. In single-label classification, each instance is associated with one exclusive label, while

in multi-label classification, an instance may belong to multiple categories. This task has been extensively studied across the scientific community, driven by its applications in sentiment analysis and topic categorization.

Question-answering. Question-answering concerns on automatically providing answers to questions expressed in natural language. In formal terms, given a question $q = [q_1, q_2, \dots, q_n]$, the aim is to select, retrieve or generate an answer a that satisfies the informational need of q .

Question-answering tasks can be broadly classified into *open-domain* and *closed-domain* systems. In open-domain question-answering, systems aim to answer questions across a wide range of topics using unstructured sources like large text corpora, which necessitates sophisticated retrieval and comprehension capabilities. In contrast, closed-domain question-answering focuses on a narrower, specialized subject area, often leveraging structured knowledge sources such as domain-specific databases, ontologies, or curated texts.

Question-answering systems operate in various paradigms, including *extractive*, where relevant spans of text are identified and extracted from documents, and *generative*, where the model composes original answers. Extractive question-answering typically suits fact-based, closed-domain tasks, while generative question-answering is often applied to open-domain tasks requiring complex language generation. Additionally, *multi-hop reasoning* enables question-answering systems to process information from multiple sources to address complex questions.

Language Modeling. Language modeling involves predicting the probability of sequences of words, thereby capturing patterns and structures in language. Formally, given a sequence of words $w_1, w_2, \dots, w_{n-1}$, the model's objective is to estimate the probability distribution of the next word w_n , or in some cases, to predict the entire sequence's likelihood. Language modeling is essential for various NLG applications, including machine translation, text generation, and speech recognition.

Semantic Textual Similarity. Semantic textual similarity determine the degree of semantic similarity relations between pairs of texts. Given two text sequences, $x = [x_1, x_2, \dots, x_n]$ and $y = [y_1, y_2, \dots, y_m]$, the aim is to measure how closely the meanings of these sequences align on a continuous similarity scale. It is widely applied in areas such as information retrieval, machine translation evaluation, and paraphrase detection (Agirre et al., 2012).

Semantic textual similarity tasks can be *supervised* or *unsupervised*. In supervised semantic textual similarity, labeled datasets provide similarity scores. For unsupervised Semantic textual similarity, no training data is provided and the similarity is estimated through embedding-based approaches that project texts into high-dimensional spaces where semantically similar texts are closely positioned (Reimers and Gurevych, 2019).

Abstractive Summarization. Abstractive summarization is to generate a concise summary of a given text, based only on the reading comprehension of the model and using words from its own vocabulary. Formally, given a source document $x = x_1, x_1, \dots, x_n$, the goal is to map the input sequence into a new sequence

$y = y_1, y_2, \dots, y_m, m \ll n$ that summarizes the content of the source and it is not a subsequence of the document. The abstractive summarization requires a broad understanding of the natural language, at the document level rather than at the word or sentence level. In contrast, the *extractive* summarization consists of providing a summary by selecting the most salient sentences of the text, so it can essentially be considered a text classification task and is significantly less difficult.

Language Inference. This encompasses a range of tasks requiring models to identify logical relationships, implicit knowledge, and contextual dependencies within text. A prominent example is *Natural Language Inference (NLI)*, or *Textual Entailment*, which aims to determine whether a “hypothesis” text can logically be inferred from a “premise” text. Formally, given a pair of sentences (p, h) , where p represents the premise and h represents the hypothesis, the task is to classify their relationship into one of three categories: entailment (if h logically follows from p), contradiction (if h contradicts p), or neutral (if h is unrelated to p).

Another key aspect of language inference is *Commonsense Reasoning*, which requires models to leverage implicit knowledge and everyday logic to deduce conclusions that are not explicitly stated in the text. It involves understanding cause-and-effect relationships, temporal ordering, and other forms of pragmatic reasoning rooted in real-world experience.

Lexical Entailment. Lexical Entailment explores the semantic relationships between verbs, to determine if the occurrence of one verb implies the occurrence of another (Fellbaum, 1990; Geffet and Dagan, 2005). Analogous to logical entailment, which applies to propositions, lexical entailment is to determine if two verbs v_1 and v_2 have an entailment relationship, i.e. if the sentence $\langle \text{subject} \rangle v_1$ logically entails the sentence $\langle \text{subject} \rangle v_2$.

Lexical entailment encompasses not only binary relationships but also *graded lexical entailment (GLE)* tasks, i.e. assigning a degree of the lexical entailment relation between two concepts (Vulić et al., 2017).

Lexical entailment can contribute to broader applications, such as hypernymy detection and semantic similarity, where understanding entailment is essential for capturing the degree of relatedness among words.

2.3.2 Problems and tasks for the legal-domain

We organize our discussion on the legal problems into three broad areas, namely *search*, *review*, and *prediction* (Greco and Tagarelli, 2024a). Through our discussion, we attempt to organize the flow of presentation by distinguishing tasks involving *codes* (e.g., statutes, regulations, contracts) from those concerning *case law*; however, it is often the case that a task can be regarded as relevant for any type of legal document. Please note that the three macro categories are actually interleaved and interrelated in many practical scenarios, therefore our purpose of classification should be taken with a grain of salt, mainly for the sake of presentation.

Legal Search. Legal search corresponds to a need for legal information, and hence requires the detection and retrieval of documents potentially relevant to support legal decision-making. For instance, lawyers may search for laws enacted by parliaments or civil codes (similar to legislation in a civil law jurisdiction), but also for documents in litigation, patents, and several other documents that can support a law firm (Locke and Zuccon, 2022).

The searched documents are also called legal authorities in Dadgostari et al. (2021), which points out how the legal search is driven by a notion of *relevance*, which should be *'determined functionally with respect to norms and practices concerning legal reasoning and argumentation within a legal community'*. Thus, a document is regarded as *'legally relevant exactly when it is understood by the dominant legal community as containing information that bears on a legal question of concern'* (Dadgostari et al., 2021).

Legal search has been addressed in Dadgostari et al. (2021) as a citation recommendation problem: given a citation-free legal text (CFLT), to find the most suitable set of opinions, from a reference legal corpus, to be cited for the input CFLT. Then, if the CFLT is opinions from the corpus where all citation information is deleted, the search results can be compared to the actual citation information. More in general, legal search tasks are mainly addressed from two perspectives, namely Information Retrieval and Textual Entailment. While the former is intuitively seen as an essential part of any legal search task, the latter actually corresponds to Natural Language Inference, since it aims to determine, given any two textual fragments (e.g., two sentences), whether the one can be inferred from the other one; the entailment is said “positive”, resp. “negative” when the first text can be used to prove that the second text is true, resp. false, otherwise (i.e., if the two texts have no correlation) the entailment is regarded as “neutral” (Kim et al., 2021).

Since 2014, the Competition on Legal Information Extraction/Entailment (COLIEE) has served as an international forum to discuss issues related to legal information retrieval and entailment.³⁵ The COLIEE editions from 2014 to 2017 focus on a two-phase legal question answering task: given a legal bar exam question q , the first phase is to retrieve a set of articles from a target Civil Code corpus (i.e., Japanese civil code) that are deemed as appropriate for answering q , whereas the second phase is to determine if the (gold) relevant articles entail q or *not* q . Since the 2018 edition, both the retrieval and entailment tasks are also applied to *case law* texts, which are relatively long documents consisting of the facts (i.e., factual statements) in a case. Searching for case law is a peculiarity of common-law jurisdictions, which comes about from the principle of “stare decisis” (doctrine of precedent), and has unique challenges that have emerged in law research (Locke and Zuccon, 2022); conversely, for the civil-law jurisdictions, statutes are applied in the decision-making for a given legal issue in a *mutatis mutandis* approach, i.e., when asserting the substantial identity of two facts, we want to ignore the circumstances of a contingent nature, which

³⁵ webdocs.cs.ualberta.ca/~miyoung2/jurisin_task/index.html;
webdocs.cs.ualberta.ca/~miyoung2/COLIEE201i ($i \in \{5, 6, 7\}$);
sites.ualberta.ca/~miyoung2/COLIEE2018/; sites.ualberta.ca/~rabelo/COLIEE20i
($i \in \{19, 20, 21\}$)

are naturally different. The most recent edition at the time of the development of this research, COLIEE-2021 (Rabelo et al., 2022), proposes five tasks:

- Legal Case Retrieval (Task 1) – the goal is to identify the cases from a court case corpus that support the decision of a query case; such cases are also called “noticed” with respect to the query case, i.e., precedent cases that are referenced by the query case. Formally, given a set of candidate cases $C = \{c_1, \dots, c_n\}$ and a query case q , the task is to identify the supporting cases $C_q = \{c \mid c \in C \wedge \text{noticed}(q, c)\}$, where $\text{noticed}(q, c)$ denotes that c should be noticed given the query case q .
- Legal Case Entailment (Task 2) – given a query case, the goal is to identify one or more paragraphs from a case relevant to the query that entail(s) the decision of the query. Formally, given a query case q and a case c_i relevant for q represented by its paragraphs $\{c_{i1}, \dots, c_{in_i}\}$, the task is to identify the set of paragraphs $\{c_{ij} \mid c_{ij} \in c_i \wedge \text{entails}(c_{ij}, q)\}$, where $\text{entails}(c_{ij}, q)$ is true if the paragraph c_{ij} entails q .
- Statute Law Retrieval (Task 3) – this is the former phase-1 in COLIEE-2014, i.e., given a civil code S and a legal bar exam question q , to retrieve the set of articles S_q from S such that $\text{entails}(S_q, q)$ or $\text{entails}(S_q, \text{not } q)$.
- Statute Law Entailment (Task 4) – this is the former phase-2 in COLIEE-2014, i.e., given a legal bar exam question q and relevant articles S_q , to determine if it holds that $\text{entails}(S_q, q)$ or $\text{entails}(S_q, \text{not } q)$.
- Legal Question Answering (Task 5) – this is regarded as a combination of Task 3 and 4 (although, in the COLIEE competition, any knowledge source other than the results of Task 3 can be used).

Training data are pairs $\langle \text{query}, \text{noticed case(s)} \rangle$ for Task 1, triplets $\langle \text{query}, \text{noticed case(s)}, \text{entailing paragraph IDs of the case(s)} \rangle$ for Task 2, pairs $\langle \text{query}, \text{relevant article(s)} \rangle$ for Task 3, triplets $\langle \text{query}, \text{relevant article(s)}, Y/N \text{ answer} \rangle$ for Task 4 and Task 5; the test data are only queries for Tasks 1, 3, and 5, whereas they include queries and relevant texts for Tasks 2 and 4.

It is worth noticing that supporting cases are relevant factors in court decision-making and are actually used in the attorney’s ligation (Nguyen et al., 2021a). Legal case entailment is also useful in practice, since a decision for a new case can be predicted by implication of previous cases; it can also be treated in combination with case retrieval, as developed in Vuong et al. (2023), where a supporting model is introduced to describe the case–case supporting relations and to define paragraph–paragraph and decision–paragraph matching strategies. Analogous considerations hold for the statute law tasks. Moreover, the latter are particularly challenging since, besides the need for addressing the long articles, legal bar exam questions describe specific legal cases, while the language used in statute law tends to be more general.

A further perspective on legal question answering is taken in Zheng et al. (2021), where a *multiple choice question answering* task, dubbed CaseHOLD, is defined from legal citations in judicial rulings. The citing context from the judicial decision serves as the prompt for the question, whereas the answer choices are holding statements derived from citations following text in a legal decision. Holdings are

central to the common law system, as they represent the predominating, precedential legal rule when the law is applied to a particular set of facts. Analogously, in Xiao et al. (2021), legal question answering is addressed on the JEC-QA dataset, which consists of multiple-choice questions from the Chinese national bar exam, where the questions and candidate choices are concatenated together to form the inputs of the models.

Legal Document Review. Document review is another critical process for law practitioners and lawyers, as it usually involves document sets that are unmanageable for a team of humans given their amount, the cost of reviewers, and deadlines in the context of legal proceedings. The purpose of legal document review is for the parties to a case to organize and analyze the available documents so as to determine which are sensitive or otherwise relevant to the litigation. For instance, document review can be intended to negotiate or revise an agreement, ensure that the filings of an attorney’s client comply with appropriate regulations, modify a brief for a trial motion, inspect a contract to avoid potential risks, or review client tax documents. Relevance, responsiveness to a discovery request, privilege, and confidentiality are essential criteria for any document in the review, but also in the analysis of the information to relate key documents to alleged facts or key legal issues in the case.

Shaghaghian et al. (2020) recognizes four main tasks of document review, namely information, fact, comparative, and rule navigation, which are primarily characterized in terms of the following problems:

- *Passage retrieval* – Navigating a user to answers for non-factoid questions is in fact seen as equivalent to retrieving relevant text passages during the document review process. Passage Retrieval to answer non-factoid questions can be modeled as a binary text classification task, i.e., given a set of queries $\{q_i\}_{i=1..Q}$ and a set of candidate texts (e.g., sentences, snippets, paragraphs) $\{s_j\}_{j=1..N}$, each pair question-snippet (q_i, s_j) is assigned label 1 if s_j contains the answer to q_i , and label 0 otherwise.
- *Named entity recognition* – Examining factoid questions whereby the user is searching for specific facts or entities is instead modeled as named entity recognition (e.g., extraction of facts from a court decision document, such as Date of Argument, Date of Decision, Petitioner, Judge, Sought Damages and Damages Awarded Monetary Values). Named Entity Recognition to extract facts or elements of factoid questions can be modeled as a sequence labeling, multi-class classification task, i.e., given a set of fact-related classes $\{c_i\}_{i=1..C}$, each token is assigned a class (or a distribution over the classes).
- *Text similarity* – Computing text similarity is essential to identify matching texts according to different aspects; for instance, to identify the differences between a regulation and its amended version, or to discover the discrepancies of regulations in different jurisdictions. Text similarity to identify matching texts at various, pre-determined levels can be modeled as a binary, resp. multi-class, text classification task, i.e., given a set of matching levels $\{m_i\}_{i=1..M}$ and a set of texts $\{s_j\}_{j=1..N}$, each pair of texts (s_j, s_k) is assigned a class m_i depending on the degree of matching between s_j and s_k .

- *Sentiment analysis* – This can be addressed to identify the polarity, or the mood, associate with certain legal statements, with the purpose of, e.g., identifying rules imposed by deontic modalities, which are of the form of obligations, prohibition and permission statements. This can be modeled as a binary, resp. multi-class, text classification task, i.e., given a set of texts $\{s_j\}_{j=1..N}$, each text is assigned a class depending on the polarity or sentiment expressed in the text.

In Xiao et al. (2021), *legal reading comprehension* is addressed to predict the start positions and end positions given question-answer pairs with corresponding supporting sentences. Legal document review is also related to *document recommendation*. As discussed in Ostendorff et al. (2021), a typical recommendation scenario occurs during the preparation of a litigation strategy, when the involved legal professionals are provided with recommended other decisions that possibly cover the same topic or provides essential background information (e.g., they overrule the target decision). Also, *text segmentation*, i.e., the task of dividing a document into multi-paragraph discourse units that are topically coherent, can be useful to one or more of the above tasks, especially when the existing logical boundaries imposed to the document might not be sufficient to detect fine-grain topic changes. In Aumiller et al. (2021), text segmentation is used to solve a *topical change detection* problem (also called same topic prediction): Given two chunks of text of the same type (e.g., paragraphs, sections) and binary labels, to determine if the two chunks belong to the same topic, otherwise a change in topic is detected and so the beginning of a new chunk of text. Also, Savelka et al. (2021) introduce the task of *automatic functional segmentation*, which is to segment adjudicatory decisions of cases according to the functional role of the parts.

Contracts, in various forms, are major target of interest for document review tasks. Zheng et al. (2021) focus on contract documents such as *Terms-of-Services*, for the detection of potentially unfair contractual terms. A contractual term (clause) is regarded as unfair if it has not been individually negotiated, and it corresponds to an evident imbalance in the rights and obligations of the parties, to the detriment of the consumer (Zheng et al., 2021). A binary classification task can hence be defined, whereby positive examples are the potentially unfair contractual terms. The Terms-of-Service task can help consumers better understand the terms they agree to when signing a contract and ease access to legal advice about unfair contracts. Hendrycks et al. (2021b) address the *legal contract review* task, which is to analyze a contract to understand rights and obligations of the signatories as well as to evaluate the associated impact. This task can be seen as similar to extractive question answering, where each question is the description of a label category and language models have to detect the spans of the contract is related to the label. Leivaditi et al. (2020) specialize the legal contract review task to lease agreements and address this task from two perspectives: detection of sentences expressing a potential risk to one or more signatories (binary classification) and extraction on important entities for the domain (entity recognition). Unlike Hendrycks et al. (2021b) and Leivaditi et al. (2020), which aim to find what kinds of terms are present, Koreeda and Manning (2021) focus on knowing what exactly each of these terms states. Given a set of hypotheses and a contract, the task is to decide if the contract entails, contradicts or

is neutral to each hypothesis (three-class classification) and detect the evidence, i.e., spans, in the contract that determine the decision (multi-label binary classification). On privacy policies, Ahmad et al. (2021) define the *intent classification* task, which is to predict sentences explaining privacy practices, along with a slot filling task to detect text spans within a sentence expressing specific details. A slot extraction task is also performed in Bui et al. (2021) to detect spans in the text expressing different types of user data.

Another legal context that can be included in the document review category to a broader extent concerns a special case of retrieval task, namely *regulatory information retrieval* (Chalkidis et al., 2021b), i.e., to ensure a regulatory compliance regime regarding an organization’s processes/controls. A compliance regime includes corrective, detective and preventive measures such that either, given a control/process, to retrieve relevant laws in order to apply corrective measures or, given a new law, to retrieve all the affected controls/processes in order to apply corrective or preventive measures. Regulatory information retrieval is defined as a special case of document-to-document information retrieval, since the query is an entire document — unlike traditional information retrieval, whereby queries are usually short texts.

More tasks concern case law documents. Legal cases are lengthy and unstructured, although they are actually characterized by an implicit thematic structure into sections such as “facts of the case”, “arguments given by the parties”, etc. These sections are often called as *rhetorical roles*. Identifying such semantic roles is essential for improving the roles readability of the documents but also helps in downstream tasks such as classification and summarization. The task is challenging since in most cases legal documents can vary in structure and rhetorical labels can be subjective. Bhattacharya et al. (2019b) introduce the *rhetorical role labeling* task, which is to label sentences of a legal case with the corresponding rhetorical role. This task was also introduced in the context of the Artificial Intelligence for Legal Assistance (AILA) 2020 competition (Task 2), whereby the predefined labels are “Facts”, “Ruling by Lower Court”, “Argument”, “Statute cited”, “Precedent cited”, “Ratio of the decision”, and “Ruling by Present Court”.³⁶

To support legal document review, special cases of retrieval are also involved. For instance, Martino et al. (2022) deal with the identification of *paragraph regularities* in legal cases, which is addressed by using a nearest-neighbor search method to efficiently select the most similar paragraphs appearing in a set of reference documents. *Explanatory sentence retrieval* (Savelka and Ashley, 2021) is instead to retrieve useful sentences to explain predetermined legal concepts. Explanations of legal concepts can be inferred looking at how they have been applied in previous cases, allowing a lawyer to elaborate supporting or contrary arguments related to particular accounts of meaning. Searching through legal documents, a lawyer can find sentences mentioning a particular concept, but not all of them could be useful for explaining that concept. Therefore, the aim is to automatically rank sentences in order to assign higher scores to explanatory sentences.

³⁶ <https://sites.google.com/view/aila-2020/task-2-rhetorical-role-labeling-for-legal-judgements>

It is also highly desirable for legal professionals dealing with cases to access to their summaries, also known as *headnotes*. However, creating headnotes is certainly time-consuming, therefore automatic *summarization of legal judgments* is another meaningful problem in the legal domain. Two related tasks have been introduced in the Artificial Intelligence for Legal Assistance (AILA) 2021 competition, namely to identify “summary-worthy” sentences in a court judgment (Task 2a) and to generate a summary from a court judgment (Task 2b).³⁷ The former can be seen as a sentence classification task, whereas the latter can be addressed either by collecting the detected summary-worthy sentences so as to form *extractive summaries* or by using generative models to produce *abstractive summaries*.

Legal Outcome Prediction. Legal relevance is related to the well-known predictive theory of the law first introduced in Oliver Wendell Holmes (1897). In contrast to previous definitions of the law, Holmes formulated the law as a prediction, particularly the behavior of a court, so as to build a more useful approach in practice when dealing with those individuals who care little for ethics or lofty conceptions of natural law (i.e., the “bad men”). Besides the Holmes’ theory, predictive tasks in law are more generally concerned with judicial opinions. For instance, as discussed in Dadgostari et al. (2021), given the content of a source judicial opinion, one task is to predict the other opinions that are cited in the source document; or, given a source document and a set of related opinions identified by law professionals, to predict their answers.

The primary predictive task in law is commonly referred to as *legal judgment prediction* (LJP), i.e., to predict the outcome of a judicial decision based on the relevant facts and laws (Aletras et al., 2016; Zhong et al., 2018; Chalkidis et al., 2019a). For instance, Aletras et al. (2016) define the problem of case prediction as a binary classification task, which is to predict whether one of a predetermined, small set of articles of the ECtHR Convention has been violated, given textual description of a case, which includes the facts, the relevant applicable law and the legal arguments.³⁸ In Xiao et al. (2021), the LJP task is addressed on both criminal and civil cases from the CAIL-Long dataset. Fact descriptions are taken as input whereas the judgment annotations are extracted via regular expressions; each criminal case is annotated with the charges, the relevant laws, and the term of penalty, and each civil case is annotated with the causes of actions and the relevant laws. For criminal cases, the charge prediction and the relevant law prediction are formalized as multi-label classification tasks, whereas the term of penalty prediction task is formalized as a regression task. For civil cases, the cause of actions prediction

³⁷ <https://sites.google.com/view/aila-2021/task-2-summarization-of-legal-judgements>

³⁸ The above view has been recognized not only as one of the most challenging by the legal community, but it has also raised controversial debate on the role of AI applied to law. Adversarial opinion is in fact based on the evidence that, in real-life scenarios, judges are unlikely to defer to AI to decide the outcome of a case. Nonetheless, the authors adopt an opinion that is commonly shared with most researchers and practitioners of AI in law, whereby it should be seen as a powerful tool to aid legal professionals to increase their access to justice, and ultimately address unmet needs of the legal community.

is formalized as a single-label classification task, and the relevant law prediction is formalized as a multi-label classification task. In Dong and Niu (2021), the three types of prediction are addressed in a context of graph node classification, where a Transformer model is combined with a graph neural network model. Malik et al. (2021) propose the *court judgment prediction and explanation (CJPE)* task, which requires to predict the decision of a case and to provide explanations for the final decision, where explanations correspond to portions in the case description that best justify the outcome.

A related axis of prediction is the one introduced in Mahari (2021), dubbed as *legal precedent prediction*, which is to predict passages of precedential court decisions that are relevant to a given legal argument posed in the context of a judicial opinion or a legal brief. Both judicial opinions and legal briefs usually contain a number of independent legal arguments, each citing its own set of precedent, where the precedent depends on the context of the entire case as well as on the specific legal argument being made (Mahari, 2021). Clearly, in common law jurisdictions, this is particularly useful as legal professionals build their arguments by drawing on judicial precedent from prior opinions.

Another critical task is *overruling prediction*, i.e., to determine if a statement is an overruling, i.e., a sentence that nullifies a previous case decision as a precedent, by a constitutionally valid statute or a decision by the same or higher ranking court (which establishes a different rule on the point of law involved). In Zheng et al. (2021); Limsopatham (2021), the overruling prediction is modeled as a binary classification task, where positive examples are overruling sentences and negative examples are nonoverruling sentences from the law. The overruling task is clearly important for legal professionals, since verifying whether cases remain valid and have not been overruled is essential to ensuring the validity of legal arguments.

Case importance and *article violation* are also considered (Chalkidis et al., 2019a), (Limsopatham, 2021). Predicting the importance of a case can be seen as a regression task, e.g., to measure on a scale from lower scores for key cases, to higher scores for unimportant cases. Given the facts of a case, article violation is to predict if any human rights article or protocol has been violated by the case (binary classification), or which human rights articles and/or protocols have been violated (if any) by the case (multi-label classification). A special case of the above task is the *alleged violation prediction* introduced in Chalkidis et al. (2021c), whose aim is to predict the allegations made by applicants given the facts of each case. This can be useful to identify alleged violations for plaintiffs, facts supporting alleged violations for judges but also for legal experts to identify previous cases related to the allegations. The task is treated in Chalkidis et al. (2021c) as a multi-label text classification, since the model might select multiple articles that were allegedly violated (according to the applicants).

Employment notice prediction (Lam et al., 2020) is to predict the number of months awarded for *reasonable notices* in employment termination cases. If the employer does not comply with the obligation to provide an appropriate employment notice or payment in lieu of notice, judges determine the compensation that an employer owes to an employee at the time of termination. Courts might rely on factors

such as length of service, employee's age, character of employment, aggravated damages to establish what constitutes reasonable notices, but it is not clear how to weigh each individual factor and how they should be used. As a result, the case law on employment notice turns out to be inherently inconsistent and subjective. Lam et al. (2020) define this problem as a text classification task, in order to obtain a similar decision-making process of a judge who would rely allegedly on past cases and differences of fact to decide the amount of reasonable notice.

2.3.3 Problems and tasks for the psychosocial-domain

The tasks in this section are organized according to two key psychosocial aspects: mental-health and morality. It should be noted that the discussion on mental-health is significantly shorter than that on morality. This is clearly not due to a lack of significance of the mental-health topic by itself—we are aware that mental-health is a highly critical area—but, rather, because the work conducted during the Ph.D. course on mental-health was limited to a relatively brief period (Greco et al., 2023).

Mental-health tasks. NLP systems may help identify, monitor, and analyze mental-health disorders by processing text from a variety of sources, including social media, clinical notes, and electronic health data. These tasks are essential for identifying at-risk individuals and symptom patterns. Key mental-health tasks include:

- **Mental-health disorder detection:** Identifying symptoms of depression, anxiety, bipolar disorder, post traumatic stress disorder (PTSD), attention deficit hyperactivity disorder (ADHD), and other mental-health conditions.
- **Suicidal ideation and self-harm detection:** Identifying early signs of suicidal thoughts or self-harm.
- **Emotion detection:** Detecting emotions such as sadness, worry, and joy within text data, often used to infer mental-health status.
- **Early risk detection:** Detecting early signs of various mental-health issues, which includes predicting pathological gambling and self-harm risk based on social media activity.
- **Severity assessment:** Determining the severity of mental-health issues, such as the level of depression symptoms based on standardized assessments or model predictions.

Morality-related tasks. Morality-related tasks in NLP involve analyzing and understanding how moral principles are expressed, debated, and perceived in text. These tasks include detecting moral values, assessing the moral framing of arguments, and examining how different moral foundations influence opinions and behaviors.

Moral Foundations Theory (MFT), formulated by Haidt and Joseph (2004a), provides a psychological framework that underlies many of morality-related tasks in NLP literature. It claims that multiple psychological systems form the basis of

Foundation	Care/harm	Fairness/cheating	Loyalty/betrayal	Authority/subversion	Sanctity/degradation
Adaptive challenge	Protect and care for children	Reap benefits of two-way partnerships	Form cohesive coalitions	Forge beneficial relationships within hierarchies	Avoid communicable diseases
Original triggers	Suffering, distress, or neediness expressed by one's child	Cheating, cooperation, deception	Threat or challenge to group	Signs of high and low rank	Waste products, diseased people
Current triggers	Baby seals, cute cartoon characters	Marital fidelity, broken vending machines	Sports teams, nations	Bosses, respected professionals	Immigration, deviant sexuality
Characteristic emotions	Compassion for victim; anger at perpetrator	Anger, gratitude, guilt	Group pride, rage at traitors	Respect, fear	Disgust
Relevant virtues	Caring, kindness	Fairness, justice, trustworthiness	Loyalty, patriotism, self-sacrifice	Obedience, deference	Temperance, chastity, piety, cleanliness

Fig. 2.20: The original five moral foundations (Graham et al., 2013). The Foundation column refers to the basic building blocks or core principles of human morality.

our intuitive ethics, which are then shaped by cultures' virtues, narratives, and institutions.

MFT's initial framework established five fundamental principled dichotomies within this evolutionary context, or *moral foundation dimensions*. Each of these dimensions represents a critical aspect of social cooperation and conflict resolution that would have been essential for our ancestors' survival and flourishing:

- **Care (vs. Harm):** making pressure to care for others and repel hurting others as the basis of this foundation. It prompts us to care and be kind to people who are suffering in vulnerable positions.
- **Fairness (vs. Cheating):** centering on mutual altruism, it correlates with the rule of fairness and treatment of social justice, lauding virtues like honesty and integrity.
- **Loyalty (vs. Betrayal):** bringing to light the key aspect of group or tribe loyalty which always resulted in survival advantages when organizing strong relational bonds.
- **Authority (vs. Subversion):** the organization's underlying principle, i.e. the conservatism, emphasizing the importance of respect and the legitimacy of authorities and traditions.
- **Sanctity (vs. Degradation):** based on the concern with purity and pollution and it is related to the 'behavioral immune system,' which serves as a guide to our reactions to potential sources of infection or moral corruption.

Care, Fairness, Loyalty, Authority and Sanctity³⁹ are categorized as *Virtues*, while their counterparts are referred to as *Vices*. Fig. 2.20 summarizes the key aspects behind each moral dimension of the MFT. Each foundation is connected with 5 basic

³⁹ Note that the value of Sanctity is also referred to as Purity.

principles of human morality (Graham et al., 2013): (i) Adaptive challenge: The basic survival need or challenge that each foundation addresses. (ii) Original triggers: the initial stimuli that originally activated each moral foundation. (iii) Current triggers: Modern stimuli that currently activate each moral foundation. (iv) Characteristic emotions: the typical emotions associated with each moral foundation. (v) Relevant virtues: the virtues or positive character traits linked to each moral foundation.

MFT offers a strong framework for examining how innate moral intuitions, shaped by evolution and culture, influence behavior and social interactions. It has been widely used in social psychology and studies exploring cultural differences and political ideologies (Kobbe et al., 2020).

We discuss and formalize the tasks related to the integration of TLMs and MFT (Zangari et al., 2024). Our objective is to present a formal definitions of the primary tasks addressed by the investigated works, serving as a valuable foundation for future research endeavors. We categorize the identified tasks as illustrated in Fig. 2.21. We emphasize that, to the best of our knowledge, this is the first work to provide a taxonomy of the main MFT tasks faced by current research on TLMs.

We define T_D as the MFT theory and D as the set of *moral foundation values* associated to T_D . Without loss of generalizability, we assume that D does not strictly represent the original MFT foundations with virtues and vices, but rather moral values in which virtues and vices can be treated separately, i.e.,

$$D = \left\{ \begin{array}{ll} \text{Care,} & \text{Harm,} \\ \text{Fairness,} & \text{Cheating,} \\ \text{Loyalty,} & \text{Betrayal,} \\ \text{Authority,} & \text{Subversion,} \\ \text{Purity,} & \text{Degradation} \end{array} \right\}.$$

Following this approach, D could be easily integrated with other moral values, i.e. $\{\text{non-moral}\} \cup D$ (Hoover et al., 2020a), which can be useful for aligning with the continuous refinements in MFT (Iyer et al., 2012; Atari et al., 2020; Atari and Haidt, 2023; Karami et al., 2023)

- **Moral values prediction.**

Moral values prediction involves learning a function M_Θ that infer the moral values D from the input data. Formally, given the input space X , the objective of the moral values prediction task is to learn $M_\Theta : X \mapsto Y$, where $Y \subseteq \mathcal{P}(D)$, and $\mathcal{P}(D)$ indicates the powerset of D . Note that M_Θ could be a TLM that is freezed, fine-tuned or further pre-trained for performing the final downstream task.

In order to train the TLM for the moral values prediction task, we assume the availability of a labeled dataset $L = (x_i, d_i)_{i=1}^n$, where $x_i \in X$ is an input instance, $d_i \subseteq D$ is the corresponding label(s), and n is the size of the dataset. The goal is to minimize the loss function $\mathcal{L}(M_\Theta(x_i), d_i)$ over the labeled dataset, where the choice of the loss function depends on the particular setting of the problem (e.g., single-label or multi-label classification).

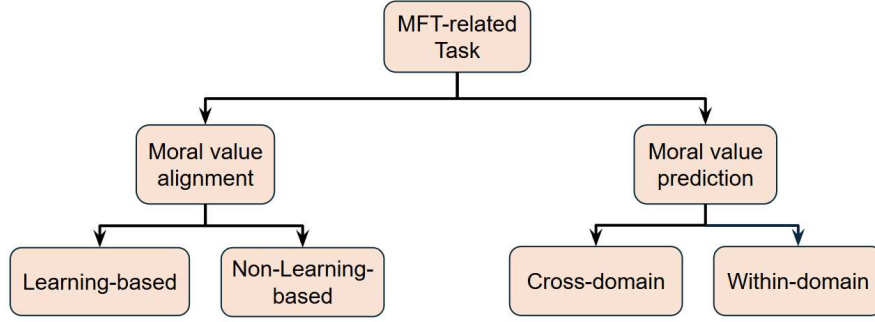


Fig. 2.21: Categorization of the main MFT-related tasks addressed by existing literature.

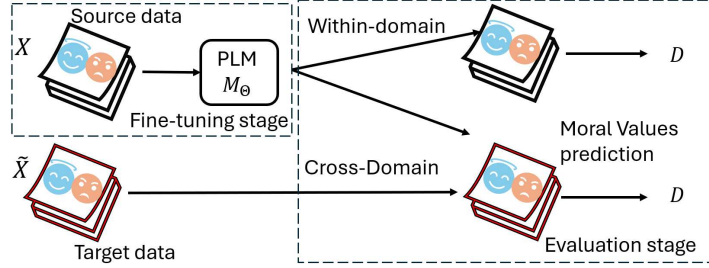


Fig. 2.22: Within-domain and Cross-domain moral values prediction task. In the within-domain scenario, the model is trained and tested on data originating from the same distribution (source). In the Cross-domain scenario, the model is trained and tested on data originating from different distributions (source and target).

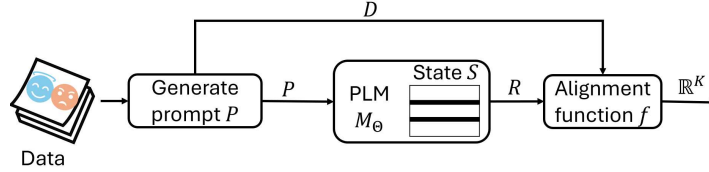


Fig. 2.23: Moral values alignment problem. Given data with moral contents, a prompt expressing specific moral values D is generated and input to a TLM, which has a state S . The moral values D and the TLM's response are then fed into the alignment function f to evaluate how well the TLM aligns with D . If the TLM is only used for inference, we are dealing with a Non-learning-based moral value alignment problem. If the TLM includes components that are trained or fine-tuned to align with f , we are dealing with the Learning-based moral value alignment problem.

Cross-domain and within-domain moral values prediction: methods that address the moral values prediction task can be framed into cross-domain and within-

domain scenarios. Note that in the current literature these scenarios can be referred to as out-of-domain and in-domain (Preniqi et al., 2024; Guo et al., 2023a; Liscio et al., 2022). Specifically, since all the works we analyzed intend the same concept for cross- (within-) and out-of- (in-) domain, we use the term *domain* to refer to a set of data characterized by a distinct distribution of textual features and labels, which can differ either due to variations in topical content (e.g., different social movements within the same dataset) or due to differences in data sources (e.g., social media posts vs. news articles). Under this definition, all cross-domain moral values prediction methods assume that training and test data are sampled from different domain distributions, both in terms of labels and textual features. Indeed, cross-corpus or cross-topic data are typically collected in different time period, and this difference can potentially impact the topics, the sentiments expressed about the topics, and the type of justification and reasoning used for the expressed sentiments (Trager et al., 2022).

We will use the terms cross-domain and within-domain to create various setups that enrich the previously defined tasks, such as cross-domain moral values prediction and within-domain moral values prediction. Within-domain methods focus on optimizing performance within the same domain of the training data, that is, train and test data are drawn from the same distribution. Cross-domain approaches focus on transferring knowledge from the source domain(s) to the target domain. They involve adapting a model trained on one or multiple domains to perform well in a different, often less-resourced, domain. These methods operate under two key assumptions to improve generalization capabilities across distinct contexts: the training and evaluation data are drawn from distinct distributions (i.e., differing in terms of labels and textual features), and the existence of latent data representations that are invariant across different environments. Figure 2.22 schematizes how a moral values prediction task is typically addressed.

- **Moral values alignment.** In the moral value alignment problem the objective is to assess whether TLM response is aligned with specific moral values. In this context, alignment refers to the evaluation of whether the response of the TLM is in line with the moral values expressed in the input text X and denoted by D . Hereinafter, we will assume that the alignment function f needs to be minimized, indicating that a lower value means a higher degree of moral alignment for the TLM. The operational framework for addressing this task is shown in Fig. 2.23. In our formulation, the problem can be approached as follows.

Learning-based Moral Value Alignment: we provide a similar setting to the one proposed by Dognin et al. (2024). Let us denote with S the current conversation history of the TLM, P the set of possible prompts given to the model, R the set of responses that can be generated, and $f : R \times D \mapsto \mathbb{R}^k$ an alignment function. The objective is to learn a function $M_\Theta : S \times P \mapsto R$ that provides responses $r_i \in R$ for a given prompt in a given state such that $f(r_i, d_i)$ is minimized over the dataset $L = (r_i, d_i)_{i=1}^n$, with $d_i \in D$.

For example, if a TLM is prompted with moral dilemmas, the aligned responses would be those that consider the moral dimensions relevant to the scenario.

Suppose a scenario that involve choosing between honesty and loyalty, such as deciding whether to tell the truth about a friend’s mistake or to keep the secret. The responses chosen by the TLM should reflect an understanding of the moral weight of fairness versus loyalty and make a decision that appropriately consider both values.

Non-learning-based Moral Value Alignment: This approach involves analyzing the responses of TLMs without the need for learning or adaptation. The predefined function M_{Θ} is combined with a module that maps the responses of the TLMs to moral foundations based on the principles of MFT. This approach entails analyzing TLM responses without the need for learning or adaptation. The predefined function M_{Θ} is combined with a module that maps the responses of the TLMs to moral foundations based on the principles of MFT. For instance, TLMs can be treated as questionnaire respondents and their response can be compared with human responses, hence providing insight into the personality of TLMs.

2.4 Data and Benchmarks

2.4.1 Data and benchmarks for the general-domain

In this section, we briefly describe key NLP benchmarks, categorized by task or problem. Similarly to Section 2.3, we emphasize that this is not an exhaustive list of benchmarks but rather a selection of those referenced in Section 2.2 and those that were directly addressed in the research conducted during the Ph.D. period.

Text Classification benchmarks. Exemplary textual classification benchmarks are SST-2, COLA, IMDB and Yelp. SST-2 (Socher et al., 2013) is a widely used benchmark for sentiment classification, focusing on binary sentiment analysis (positive or negative) of movie review sentences. COLA assesses the models’ ability to predict linguistic acceptability. Each sentence in COLA is labeled as either grammatically acceptable or unacceptable. IMDB (Maas et al., 2011) is a popular benchmark dataset for sentiment classification, containing 50k highly polarized movie reviews, split evenly between positive and negative sentiment. Yelp⁴⁰ datasets focus on sentiment analysis in user-generated reviews. Yelp-5 is used for star prediction, where the task is to predict a rating from 1 to 5 stars based on the text of a review. Yelp-2 is a polarity classification dataset, where reviews are labeled as either positive or negative, similar to IMDB but based on shorter, more concise text.

Question-Answering benchmarks. One of the most popular Question-Answering benchmarks is SQuAD (Rajpurkar et al., 2016), which includes 100k crowdsourced questions on Wikipedia articles. Given a question and a Wikipedia passage, the task is to extract the answer span in the passage. There are two versions of the benchmark: SQuAD v1.1, where questions have always an answer in the corresponding passage,

⁴⁰ <https://paperswithcode.com/dataset/yelp>

and SQuAD v2.0, which admits unanswerable questions. NewsQA (Trischler et al., 2017) consists of 119k crowdsourced questions on a set of 10k CNN news articles. Given a question and an article, the task is to extract the answer span from the article. The span length is arbitrary and some questions are unanswerable. In SearchQA (Dunn et al., 2017), 140k question-answer pairs are crawled from the J! Archive and augmented through web page snippets from Google. HotpotQA (Yang et al., 2018) involves multi-hop reasoning on 113k Wikipedia-based question-answer pairs, and a competing model is required to output also the supporting facts used to derive the answer. Natural Questions (Kwiatkowski et al., 2019) contains about 300k examples of real questions issued to the Google search engine; given a question and a Wikipedia page, it is required to return a long answer and a short answer. SWAG (Zellers et al., 2018; Devlin et al., 2019) contains 113k sentence-pair completion examples; the goal is to choose the most credible prosecution of a sentence, given four possible choices. RACE (Lai et al., 2017) consists of 100k questions and 28k passages, created by human experts; given a question and a passage, the task is to choose the correct answer among four possible candidates. In Wikihop (Welbl et al., 2018), multi-hop reading comprehension of Wikipedia articles is needed to answer multiple-choice questions. MLQA (Lewis et al., 2020b) is a cross-lingual Question-Answering dataset with 12k questions in English and around 5,000 in each of six other languages (Arabic, German, Spanish, Hindi, Vietnamese, and Simplified Chinese). Each instance is often parallel across four languages, allowing for both direct transfer and generalized cross-lingual Question-Answering tasks. TriviaQA (Joshi et al., 2017) consists of 650k question-answer pairs from trivia websites. Each question is paired with evidences from either Wikipedia or web search results. The task is to find the answer span within a large pool of documents, making it challenging for retrieval-based models.

Language Modeling benchmarks. One prominent benchmark in language modeling is LAMBADA (Paperno et al., 2016). LAMBADA focuses on testing a model’s ability to predict the final word in a sentence, where doing so requires understanding the entire context rather than relying only on the most recent few words. The dataset contains passages from novels, with the target word often relying on commonsense reasoning and long-range dependencies.

Semantic Textual Similarity benchmarks. An important benchmark for semantic textual similarity is STS-2016 (Agirre et al., 2016), where the similarity degree between any two sentences is assigned a real-valued score from 0 (indicating semantic independence) to 5 (indicating semantic equivalence). STS-2016 includes sentence pairs from diverse domains, such as news headlines, question-answer pairs, and image captions, offering a comprehensive test of models’ semantic understanding across different contexts. STS-Benchmark (Cer et al., 2017) is another widely used dataset in STS evaluation, introduced at SemEval-2017. It contains 8,628 sentence pairs from sources like news headlines, image captions, and forums, with human-annotated similarity scores between 0 and 5. The STS-Benchmark serves as a standard evaluation resource for STS models, providing a balanced mix of formal and informal language.

Abstractive Summarization benchmarks. CNN/DailyMail (Hermann et al., 2015; Nallapati et al., 2016; See et al., 2017) is one of the representative datasets for AS, which consists of almost 300k examples of news articles from CNN and Daily Mail websites, paired with multi-sentence summaries. Unlike CNN/Daily-Mail, where summaries tend to be similar to the article sentences, XSum (Narayan et al., 2018), created from BBC articles, is highly abstractive and focuses on the *extreme summarization* task, i.e., to generate a single-sentence summary. Arxiv (Cohan et al., 2018) builds on 215K scientific articles from the homonymous repository, where the abstracts of the papers are used as gold summaries. Analogously, the abstracts of patents are used as gold summaries in BigPatent (Sharma et al., 2019), which includes 1.3 million of U.S. patents retrieved from the Google Patents Public Datasets.

Language Inference benchmarks. Some of the most prominent NLI benchmarks include SNLI, MultiNLI, XNLI, QNLI and SWAG. SNLI (Bowman et al., 2015) contains 570k sentence pairs manually labeled with entailment, contradiction, or neutral relationships. It provides high-quality annotations with examples covering a broad range of topics. MultiNLI (Williams et al., 2018) extends SNLI by introducing 433k sentence pairs from multiple topics, such as fiction, government reports, and spoken dialogue. XNLI (Conneau et al., 2018) is an extension of MultiNLI to 15 languages. It includes the same sentence pairs as the MultiNLI validation and test sets, translated into multiple languages. QNLI (Wang et al., 2018) is a reformulation of the SQuAD dataset into a binary entailment task. In QNLI, the task is to determine if a given sentence from a context paragraph entails the answer to a question.

SWAG (Zellers et al., 2018) is designed for sentence completion with a focus on commonsense reasoning. It contains 113,000 sentence-pair examples where the task is to choose the most plausible continuation from four options. SWAG requires models to understand situational plausibility, making it relevant for testing inference skills in sentence comprehension.

Lexical Entailment benchmarks. Two widely recognized and openly accessible lexical resources for studying lexical entailment are WordNet and HyperLex, which serve both as sources of lexical knowledge and as benchmarks for evaluating entailment tasks.

WordNet (Miller et al., 1990) is a well-known large lexical database of English, providing features for different uses as online dictionary, thesaurus, and lexical ontology. WordNet stores terms into lexical source files by syntactic categories, i.e., nouns, verbs, adjectives, adverbs, which are grouped into sets of cognitive synonyms, called *synsets*, each expressing a distinct (lexicalized) concept. Focusing on verbs, they are categorized according to semantic fields corresponding to 15 lexicographers' files: motion, perception, communication, competition, change, cognitive, consumption, interaction, creation, emotion, possession, body care, social behavior, weather, stative functions. Note that verbs can appear in various forms, including monadic, phrasal verbs, and compound verbs. Entailment relations between verb synsets can in principle be distinguished as hyponyms or troponyms (and their reverse form, i.e., hypernyms), antonyms, and (other kinds of) entailments. In particular, *troponymy* is

a special case of entailment, since a verb v_1 is a troponym of verb v_2 if the activity (corresponding to) v_1 is doing v_2 in some manner; moreover, antonyms can also be troponyms (e.g., fail/succeed entails try, forget entails know). In practice, however, WordNet verb entailments can be accessed via either *hyponym* or *entailment set function*, as follows: if v_1 entails v_2 , then in WordNet either v_1 belongs to the set of v_2 's hyponyms or v_2 belongs to the set of v_1 's entailments.

HyperLex (Vulić et al., 2017) was built as a gold standard resource for measuring and evaluating how well semantic models capture *graded* or soft lexical entailment. To this aim, HyperLex data contain 2616 word pairs, of which 380 are verb pairs, associated with asymmetric scores on a scale 0-6 that were annotated by humans according to the question “To what degree is X a type of Y?”.

WordNet and HyperLex have been widely recognized as a valuable support for entailment tasks, both as sources of knowledge and benchmarks. An example of this complementary contribution is the study in Renner et al. (2023), which showcases the utility of WordNet in supporting GLE tasks, demonstrating how leveraging hierarchical synset structures can improve performance, as also assessed by considering the HyperLex dataset as a benchmark in experimental evaluation.

Multitask benchmarks. The well-known GLUE is a collection of nine NLU tasks having different domain, difficulty level, and data size. The tasks comprise both single-sentence and sentence-pair input modes and cover linguistic acceptability (CoLA), sentiment classification (SST-2), paraphrase identification (MRPC, QQP), sentence similarity (STS-B) and inference tasks (MNLI, QNLI, RTE, WNLI). SuperGLUE (Wang et al., 2019) is a GLUE-styled benchmark with additional challenging tasks on question answering (BoolQ, COPA, MultiRC, ReCoRD), coreference resolution (WSC), word sense disambiguation (WiC), and natural language inference (CB, RTE). Both benchmarks assess a model by averaging its scores over the various tasks. While GLUE benchmarks concern natural language understanding tasks, GLGE (Liu et al., 2021a) focuses on 8 generative tasks, covering abstractive summarization, answer-aware question generation, conversational question answering, and personalizing dialogue, each with three difficulty levels (i.e., easy, medium, hard). MMLU (Hendrycks et al., 2021a) covers 57 diverse tasks across domains such as history, mathematics, and medical sciences. The task involves answering multiple-choice questions designed to measure a model’s broad knowledge and reasoning skills, simulating real-world academic and professional assessments. XTREME (Hu et al., 2020) is a multi-lingual multi-task benchmark designed to evaluate the cross-lingual generalization capabilities of multi-lingual models across 40 languages and 9 diverse tasks. The tasks include classification (XNLI and PAWS-X), structured prediction (POS and NER), question answering (XQuAD, MLQA, and TyDiQA-GoldP), and sentence retrieval (BUCC and Tatoeba). XTREME emphasizes zero-shot cross-lingual transfer, where models are trained on English data and evaluated in other languages without additional training in the target language.

A comparison of TLMs on NLP benchmarks. We present here our analysis of TLMs performance on NLP benchmarks (Greco et al., 2022). This study is not an exhaustive review of all TLMs across all benchmarks, but rather it focuses on key

Table 2.2: NLP benchmarks and corresponding leaderboards

Benchmarks	Leaderboards
GLUE	https://gluebenchmark.com/leaderboard
SuperGLUE	https://super.gluebenchmark.com/leaderboard
SQuAD v1.1	https://paperswithcode.com/sota/question-answering-on-squad11-dev
SQuAD v2.0	https://rajpurkar.github.io/SQuAD-explorer/
SWAG	https://leaderboard.allenai.org/swag/submissions/public
RACE	https://paperswithcode.com/sota/reading-comprehension-on-race
GLGE	https://microsoft.github.io/glge/
Yelp-2	https://paperswithcode.com/sota/text-classification-on-yelp-2
Yelp-5	https://paperswithcode.com/sota/text-classification-on-yelp-5
IMDB	https://paperswithcode.com/sota/text-classification-on-imdb
TACRED	https://paperswithcode.com/sota/sentiment-analysis-on-imdb
Wikihop	https://paperswithcode.com/sota/relation-extraction-on-tacred
HotpotQA	https://paperswithcode.com/sota/question-answering-on-wikihop
Natural Questions	https://hotpotqa.github.io/
STS-2016	https://ai.google.com/research/NaturalQuestions/leaderboard
CNN/DM	https://paperswithcode.com/sota/semantic-textual-similarity-on-sts16
XSum	https://paperswithcode.com/sota/abstractive-text-summarization-on-cnn-daily
Arxiv	https://paperswithcode.com/sota/text-summarization-on-x-sum
BigPatent	https://paperswithcode.com/sota/text-summarization-on-arxiv
	https://paperswithcode.com/sota/text-summarization-on-bigpatent

Table 2.3: NLP benchmarks and involved TLMs

Benchmarks	Methods
	Text Classification
IMDB	BERT, RoBERTa, DistilBERT, Longformer, BigBird, XLNet
Yelp	Bigbird, RoBERTa, BERT, XLNet
	Question-Answering
SQuAD v1.1	BERT, LUKE, SpanBERT, RoBERTa, DeBERTa, DistilBERT, AIBERT, T5, ELECTRA, BART, DPR
SQuAD v2.0	BERT, SpanBERT, RoBERTa, DeBERTa, AIBERT, ELECTRA, GPT-3, XLNet
NewsQA	BERT, SpanBERT, RoBERTa, AIBERT, BART
SearchQA	BERT, SpanBERT
HotpotQA	BERT, SpanBERT, RoBERTa, DPR, Longformer, BigBird
Natural Questions	BERT, SpanBERT, RoBERTa, DPR, Longformer, BigBird, GPT-3
Wikihop	Longformer, BigBird
RACE	BERT, RoBERTa, DeBERTa, AIBERT, GPT-3, XLNet
	Commonsense reasoning
SWAG	BERT, RoBERTa, DeBERTa, AIBERT, T5, BART, BigBird, XLNet, Prophetnet
	Relation Extraction
TACRED	BERT, SpanBERT, ERNIE, LUKE
	Semantic textual Similarity
STS-2016	S-BERT, S-RoBERTa, BERT
	Abstractive Summarization
CNN/DM	BART, T5, GPT-2, Pegasus, ProphetNet
XSUM	BART, Pegasus, BigBird
Arxiv	LED, Pegasus, BigBird
BigPatent	Pegasus, BigBird
	Multitask
GLUE	BERT, SpanBERT, RoBERTa, DeBERTa, DistilBERT, AIBERT, T5, ELECTRA, BART, BigBird, ERNIE
SuperGLUE	BERT, RoBERTa, DeBERTa, AIBERT, T5, ERNIE, GPT-3
GLGE	ProphetNet, BART

tasks within NLU: document classification, question-answering, relation extraction, semantic textual similarity. Furthermore, this work marked the beginning of our exploration into abstractive summarization, a task that would later become a central focus of the Ph.D. research. *For each benchmark, we analyzed models that, at the time of the study (2021-2022), represented the state of the art.*

To retrieve the performance scores, we referred to the original works of each TLM, with the complementary support of publicly available benchmark leaderboards (Table 2.2) that we accessed on February-March 2022.

We begin with getting an overview of the popularity of the benchmarks in terms of TLMs involved. As shown in Table 2.3, we notice a higher participation in Question-Answering benchmarks and GLUE benchmarks, which is not surprising as they have been active for longer than other benchmarks; nonetheless, more recent benchmarks, such as GLGE, have been attracting attention from more advanced TLMs. In Fig. 2.24, we report the % scores of the top-3 TLM performers, also with the addition of BERT when available, for selected benchmarks.

In GLUE, the best TLM is ERNIE, which is also ranked third in absolute position par with a combination of DeBERTa with CLEVER fine-tuning according to the current leaderboard. While BERT was the first TLM to advance GLUE state-of-the-art, ERNIE, DeBERTa and T5 have now pushed the benchmark to the next level achieving impressive results. A similar scenario can be found in SuperGLUE, where ERNIE 3.0, DeBERTa and T5 represent the top-3 TLMs, with even a larger score-gap to BERT than in GLUE. In absolute terms, ERNIE 3.0 obtains the 3rd position among all NLP methods in the leaderboard. Notably, DeBERTa is able to beat the SuperGLUE human baselines, thus marking a milestone toward the general-purpose language models performance (He et al., 2021).

The GLGE benchmarks provide scores of BART and ProphetNet as strong baselines and available on the leaderboard. In GLGE-Easy, ProphetNet outperforms BART and is ranked absolute 3rd, while in GLGE-Medium and GLGE-Hard settings, ProphetNet and BART are respectively ranked first.

In SQuAD v1.1, the majority of TLMs have been evaluated on development set (for short, dev set), since the test set was not released to the public. Largest T5 (11B parameters) achieves the best F1 score, improved only by a model based on XLNet (Li et al., 2020). The original papers of RoBERTa, AlBERT, DeBERTa and ELECTRA present dev F1 scores on the benchmark not shown in the SQuAD leaderboard, while the T5 paper shows a higher score than that reported in the leaderboard, which may be due to an earlier version of the paper reporting a score of 95.64. According to the scores reported in the papers, the leaderboard on dev F1 would change to the following order: T5 (96.22), DeBERTa/AlBERT (95.5), and XLNet (95.1), but recent versions of DeBERTa⁴¹ enhance performance on SQuAD reaching a dev F1 score of 96.1. Test set of SQuAD v2.0 is not public too, but various TLMs have been submitted for evaluation on test set with scores available on the leaderboard. Also in this case, recent versions of DeBERTa reach AlBERT in the first position

⁴¹ <https://github.com/microsoft/DeBERTa>

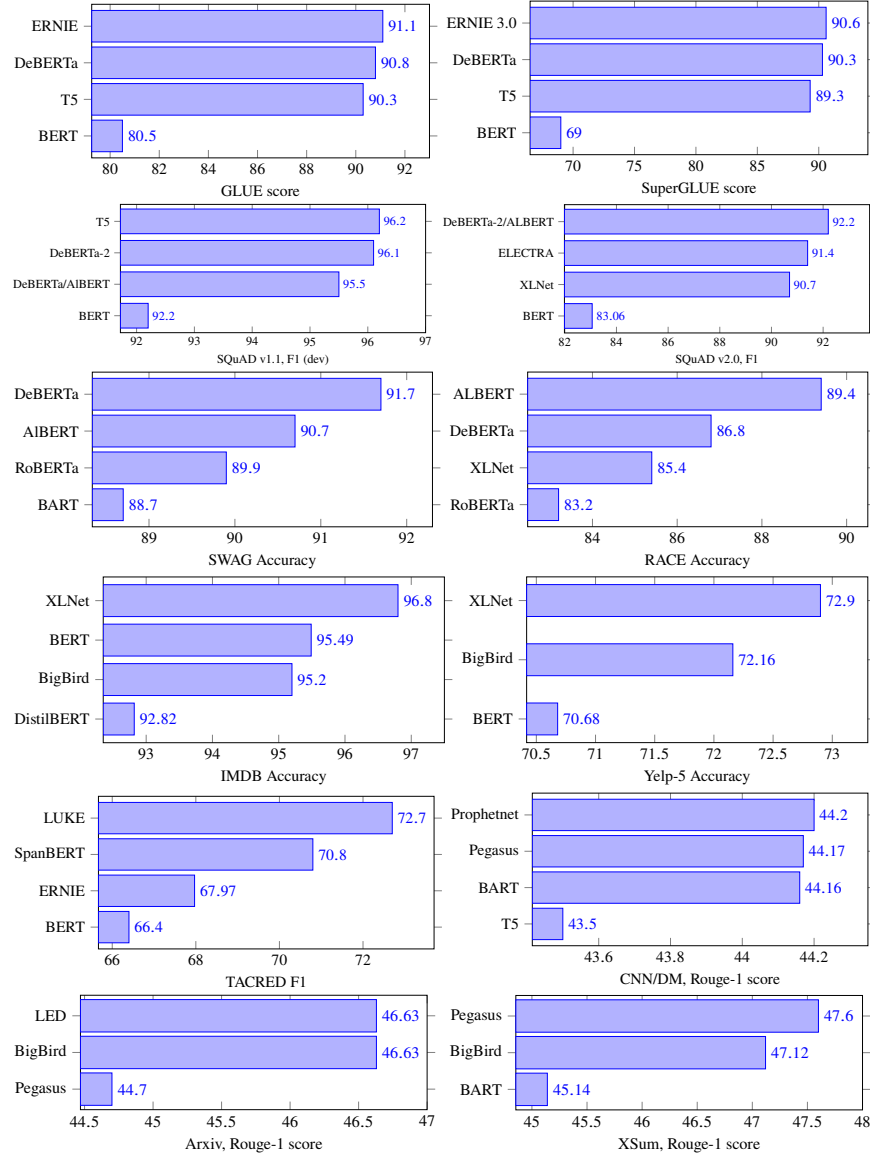


Fig. 2.24: Best-performing TLMs on selected NLP benchmarks

(92.2), followed by ELECTRA and XLNet. In the leaderboard, we notice a large use of ensemble models for this task.

The SWAG leaderboard presents scores for TLMs that are not included in the original papers. DeBERTa is ranked first among all NLP methods on SWAG. In the absolute ranking, AIBERT also obtains the third position after a modified version of

RoBERTa, called ALUM-RoBERTa-large (Liu et al., 2020a). From the leaderboard, we notice the predominant use of TLMs for this task.

According to the original papers, the top three TLMs in RACE are AIBERT (89.4), DeBERTa (86.8) and XLNet (85.4). In the leaderboard, top positions are taken by ensemble of TLMs, while AIBERT score from Lan et al. (2020) is not reported in the leaderboard, and the same holds for the XLNet score.

From the leaderboards of Yelp-5 and Yelp-2, we retrieve the XLNet accuracy scores, whereas the XLNet paper reports the error rates. On XLNet paper it is also reported the error rate of BERT on Yelp-2 and Yelp-5, from which the accuracy scores can be calculated. The leaderboard of Yelp-5 shows also the accuracy score of BigBird (which matches the F1 micro-averaged score reported in the original paper). Performance on Yelp-2 is not reported, since the only available accuracy score refers to XLNet and it is shown in the relative leaderboard.

On IMDB, two leaderboards are available (cf. Table 2.2), which report different accuracy scores for XLNet, with the one matching the performances of XLNet and BigBird shown in their respective papers (for the latter, the original paper reports the micro-averaged F1 for both BigBird and RoBERTa), and the other one including the accuracy scores of BERT and DistilBERT. In TACRED, the top-3 TLMs are LUKE, SpanBERT and ERNIE. According to the available leaderboard, the best absolute method uses SpanBERT. The BERT score is from the SpanBERT work (Joshi et al., 2020).

As regards HotpotQA, results are known from the original papers of Longformer (Beltagy et al., 2020) and BigBird (Zaheer et al., 2020), both using the distractor setting of the benchmark and F1 metric. BigBird achieves better performance on test set, with a score of 73.6 against 73.2 by Longformer. In the current leaderboard, we observe that the method in the top position (submitted on 28 Jan 2022) is based on AIBERT and achieves a score of 76.54, while BigBird and Longformer are ranked 20th and 21th, respectively. Analogously, BigBird and Longformer are tested on WikiHop benchmark. From the original papers and the available leaderboard, BigBird outperforms Longformer (82.3 against 81.9). Bigbird is also the best absolute method according to the leaderboard. In Natural Questions, BigBird achieves in the current leaderboard the 6th position for long answer setting (77.78 on test F1) and the 24th position in short answer setting (57.85 on test F1).

Concerning semantic textual similarity, S-BERT and S-RoBERTa were tested on STS-2016 without using any specific STS training data (Reimers and Gurevych, 2019). From the reported performance, S-RoBERTa achieves the best score against S-BERT and the competitors. In the current leaderboard, S-RoBERTa is ranked 9th.

In abstractive summarization tasks, we notice the predominant presence of Pegasus in the top-3 TLMs. From the available leaderboard of CNN/DM, the current top model w.r.t. Rouge-1 metric (more details on Rouge metrics are provided further down in the manuscript in Section 4.4.2) is *GLM* (Du et al., 2022). On Arxiv, XSum and BigPatent, best performance of BigBird is obtained by leveraging Pegasus; this version is known as BigBird-Pegasus (Zaheer et al., 2020). In Beltagy et al. (2020) BigBird and LED are ranked the same for Arxiv on Rouge-1 metric, but LED outperforms BigBird on Rouge-2 and Rouge-L; however, in the Arxiv leaderboard, the

LED score is missing. On XSum, the “Mixed and Stochastic” version of Pegasus outperforms BigBird and BART with a score of 47.6. In the leaderboard of XSum, the score of BigBird-Pegasus is missing, while the reported Pegasus score refers to the version of the model trained on HugeNews. As regards BigPatent, BigBird-Pegasus and Pegasus achieve Rouge-1 score of 60.64 and 53.63, respectively. In the BigPatent leaderboard, it is just reported BigBird-Pegasus, with no competitors.

It should be emphasized that it is not always possible to compare all the models applied to the same benchmark, due to different choices adopted by researchers in the experimental settings. In NewsQA, SearchQA, HotpotQA and Natural Questions, the authors of SpanBERT use a simplified version of the benchmarks taken from the MRQA shared task (Fisch et al., 2019); for example, in Natural Questions, only the short answers (SA) are considered, and long answers (LA) are used as context, while in NewsQA unanswerable questions or without annotator agreement are discarded. The authors experimented also with BERT on those MRQA tasks highlighting the superior performance of SpanBERT. In Pearce et al. (2021), we can observe a fair comparison of BERT, XLNet, RoBERTa, ALBERT and BART on NewsQA in terms of F1 score and using the base version of all of them. RoBERTa achieved the best results, followed by BART, XLNet, BERT and ALBERT.

Main findings, limitations and further perspectives. We have discussed how most existing TLMs perform on well-known NLP benchmarks. Our general objective was to provide a short guide on the ability of TLMs to achieve significant performance on fundamental NLP tasks. We believe this can be useful not only to practitioners who may want to know which TLM should be preferred on given tasks, but also to NLP specialists looking for baselines and competitors for their own methods, or for methods that can inspire new solutions for the NLP problems.

In the attempt of answering about the extent to which TLMs are able to perform well, or even close to human level, we have ascertained that TLMs can achieve very high or near-optimal performance in some tasks (e.g., IMDB and SQuAD v1.1), whereas others (e.g., Yelp-5 and TACRED) indicate that the margins of improvement for TLMs are still large.

Nonetheless, our major findings in this work go beyond the TLM performance cases. In fact, we have noticed that the picture on NLP benchmark scores cannot always be complete or totally certain, which depends on various factors. For instance, the benchmark leaderboards can report methods’ scores that are not present in the original works (sometimes due to the unavailability of test sets during the benchmark competition, or simply because others than the method’s authors applied the method to the benchmark data). It also happens that the benchmark leaderboards are not up-to-date w.r.t. all relevant TLM scores (e.g., T5 in SQuAD v1.1 on the dev set), or they can miss new emerging TLMs. Overall, for benchmarks like GLUE and SQuAD, the involved methods turn out to be comparable and we have indeed shown the top performers, but for other benchmarks, such as NewsQA and Natural Questions, it is difficult to get a fair comparison due to differences between the benchmark and the method in terms of experimental setup (e.g., open-domain vs. closed-domain), dataset modifications or simplifications, and variations in the assessment criteria.

Our study has also a number of limitations. First, the length constraints of this paper forced us to make a selection from all existing TLMs and benchmarks; in this respect, our preference went to most popular and/or relevant models and tasks, although we had to leave out of consideration challenging problems such as coreference resolution, or machine translation and related multi-/cross-lingual tasks, as well as emerging TLMs such as REALM, ProphetNet-X and MASS. Second, we restricted our analysis to the original versions of TLMs, although existing variants (in terms of architecture components, training procedures, and parameter settings) could have also been used on NLP benchmarks. Third, our study discarded tasks for specific domains, such as the legal and biomedical ones, which have their own competitions and benchmarks. Therefore, we recognize the need for further investigating the performance of TLMs on NLP benchmarks to fill the lack of knowledge on the above points.

2.4.2 Data and benchmarks for the legal-domain

We provide a summary of the main benchmarks and datasets that have been recognized as relevant in the TLM-based legal learning context (Greco and Tagarelli, 2024a).

Our main focus is on those corpora that were used by the approaches covered in this work, which will be described next (Section 3). Note that we shall leave out of consideration the datasets used in the COLIEE Competitions, since they have already been described in Section 2.3.2.

Our presentation is organized into three parts, which describe corpora concerning caselaw documents, codes, and a combination of both, respectively; moreover, each subsection is further organized by possibly grouping corpora that are cohesive in terms of data type and task. Table 2.4 summarizes the datasets that we shall describe through this section, according to the legal document category, the data type, the source, the size, and the tasks for which the benchmarks were designed.

Caselaw data. Strickson and Iglesia (2020) propose a corpus of about 5K labeled UK court judgments, gathered from the web, for the task of JLP. Each law case is divided into separate judgments issued by individual judges, and each sentence in a judgment is labeled as “allow” or “dismiss” through a pattern matching approach. The dataset is used for the JLP task as a classification problem, whereby classic machine learning classifiers (e.g., support vector machine, random forest, logistic regression) are evaluated.

ECHR (Chalkidis et al., 2019a) contains allegations of violated provisions regarding the European Convention of Human Rights.⁴² Each case includes a list of facts and a score, provided by the Convention, representing the importance of the case in the case law’s development. Also, each case is mapped to the violated articles of the Convention. Moreover, (Quemy and Wrembel, 2022) present *ECH-OD*, a new

⁴² <https://archive.org/details/ECHR-ACL2019>

database for storing and managing ECHR cases. This is designed to be automatically maintained and used as a unified benchmark to compare machine learning methods for the legal domain. The authors have provided the whole pipeline for the benchmark data extraction, transformation, integration, and loading as open-source software.

Swiss-Judgment-Prediction (SJP) (Niklaus et al., 2021) comprises 85K cases, in diachronic order, from the Federal Supreme Court of Switzerland (FSCS).⁴³ The evaluation task is a binary classification of the judgment outcome (i.e., approval or dismissal). The dataset includes cases written in German, French and Italian, and is annotated with publication years, legal areas and cantons of origin. Niklaus et al. (2021) evaluate XLNet, RoBERTa, ALBERT (cf. Section 2.2), GermanBERT, UmBERTo, CamemBERT (cf. Section 3.2.5), and two variants namely Hierarchical BERT and Long BERT, both in monolingual or multi-lingual versions (cf. Section 3.2.6).

GerDaLIR (Wrzalik and Krechel, 2021) is a dataset for *legal precedent retrieval* on German language.⁴⁴ It is based on case laws gathered from the *Open Legal Data* (Ostendorff et al., 2020). Passages containing references are considered queries, while the referenced law cases are labeled as relevant. The authors evaluate a set of retrieval methods on this dataset with Transformer-based re-ranking. In particular, they fine-tune GBERT and GELECTRA base versions using top-100 BM25 passage rankings and test the final models on top-1000 BM25 passage ranking; the use of ELECTRA for re-ranking has shown to lead to higher performances in most cases. Urchs et al. (2021) introduce two further legal corpora for the German law. The first corpus⁴⁵ contains about 32K decisions, enriched with metadata, from hundreds of Bavarian courts. There are 22 different types of decisions in the corpus, such as resolutions, judgments and end-judgments. This corpus is not intended for a specific task (for example, it can be used to detect the type of the decision). The second corpus⁴⁶ is a subset of the former and contains 200 judgments, whose sentences (about 25K) were annotated by a domain expert w.r.t. four components of the text (written in the *Urteilsstil* style): “conclusion” (i.e., the overall result of the case), “definition” (i.e., abstract legal facts and consequences), “subsumption” (i.e., the ensemble of concrete facts and determination sentence), and “other” (i.e., sentences not labeled with any of the three previous labels). This corpus is intended for the automatic detection of conclusion, definition and subsumption components.

Zhong et al. (2019b) provide 92 expert-annotated extractive summaries of *Board of Veterans’ Appeals* (BVA) cases focused on post-traumatic stress disorder (PTSD), along with 20 test cases quadruple-annotated for agreement evaluation and two summaries for each test case and written by law experts.⁴⁷ Each sentence is annotated considering six labels, namely issue, procedural history, service history, outcome, reasoning, evidential support. Also, Walker et al. (2019) introduce a dataset to test

⁴³ https://huggingface.co/datasets/rcds/swiss_judgment_prediction

⁴⁴ <https://github.com/lavis-nlp/GerDaLIR>

⁴⁵ https://zenodo.org/record/3936726#.ZAdMIXbMJJD_

⁴⁶ https://zenodo.org/record/3936490#.ZAdN7HbMJJD_

⁴⁷ <https://github.com/luimagroup/bva-summarization>

the performance of rule-based script classifiers, comprising 50 fact-finding decisions of BVA cases focused on veterans' appeals to a rejected disability claim for service-related PTSD. Each sentence of the dataset is assigned a rhetorical role by domain experts as follows: finding sentence, if it states a finding of the fact, evidence sentence, if it states the content of a testimony, reasoning sentence, if it reports the reasoning of the judge underlying the findings of facts, legal-rule sentence, if it states legal rules in the abstract, and citation sentence, if it refers to legal authorities and other materials. The dataset is used to test two hypotheses: whether distinctive phrasing allows automatic classifiers to be developed on a small set of labeled decisions, and whether semantic attribution theory can provide a general approach to develop such classifiers. Results demonstrate that some use cases can be addressed using a very small set of labeled data.

*Multi-LexSum*⁴⁸ is a collection of almost 9K expert-edited abstractive summaries for 40K writings of the Civil Rights Litigation Clearinghouse (CRLC),⁴⁹ which provides information on federal US civil rights cases for various target audiences (lawyers, scholars, and the general public) (Shen et al., 2022). It is designed for multi-document and single-document summarization tasks. The source documents are extremely long, with cases often having more than two hundred pages. Multi-LexSum provides multiple summaries with different granularity (from an extreme one-sentence summaries to summaries with more than five hundred words). Although the provided summaries are abstractive, they present a high fraction of terms included also in the source document.

RulingBR (Feijó and Moreira, 2018)⁵⁰ comprises 10K Brazilian rulings for summarization on legal tasks, which were retrieved from the decision documents of the highest court in Brazil, *Supremo Tribunal Federal* (STF).⁵¹ Each decision document is composed of the following four parts: “Ementa” (i.e., summary), “Acórdão” (i.e., judgment), “Relatório” (i.e., report), and “Voto” (i.e., vote). The Ementa part is used as gold summary for the dataset. Lage-Freitas et al. (2022) also propose a dataset consisting of about 4K legal cases from a Brazilian State higher court (*Tribunal de Justiça de Alagoas*), with a focus on the Brazilian appeals system, assigning the appeals with labels regarding court decisions. Following Aletras et al. (2016), the authors assume that there is enough similarity between the case description of legal judgments and appeals lodged by attorneys. Brazilian courts data are scraped from the Web and segmented into sections, identifying the description, decision and unanimity parts; then, description sentences are labeled according to the decision outcome (yes, no, or partial) and unanimity information (unanimity vs. non-unanimity).

CAIL2019-SCM (Xiao et al., 2019) is a dataset of about 8K triplets of cases of the *Suprem People's Court of China*, concerning private lending.⁵² It was collected from

⁴⁸ <https://multilexsum.github.io/>

⁴⁹ <https://clearinghouse.net/>

⁵⁰ <https://github.com/diego-feijo/rulingbr>

⁵¹ <https://portal.stf.jus.br/>

⁵² <https://github.com/china-ai-law-challenge/CAIL2019/tree/master/scm>

the *China Judgments Online*⁵³ for the *CAIL* competition, where participants were required to perform a similar case matching task, i.e., to detect which pair of cases in the triplet contains the most similar cases. Every document in the triplet refers to the fact description of a case. The most similar pair within each triple is detected by legal experts. The authors provide some baselines to compare the participants' performance, one of which uses BERT to obtain embeddings of the two cases, for which the similarity score is computed. The *CAIL* competition was first held in 2018 (Xiao et al., 2018). In *CAIL2018*, participants were required to perform a legal judgment prediction task divided in three sub-tasks: law article prediction, charge prediction, and term-of-penalty prediction. The input is the fact description of a criminal case, and the associated dataset is divided into two sub-datasets: *CAIL-big* (with more than 1.6M cases) and *CAIL-small* (about 130K cases). Yu et al. (2022b) extended the fact prediction task data of *CAIL* 2021 for the explainable legal case matching task. The sentences of a legal case in *CAIL* 2021 are associated with several tags regarding the issue of private lending. In the proposed dataset, called *eCAIL*,⁵⁴ the tagged sentences are considered as rationales. Given two legal cases, the cross-case sentences with identical labels are *pro-rationales* for the matching task, while sentences with different labels are *con-rationales*. A matching label is assigned for the case pair according the tag-overlapping: if there is a overlapping of more than 10 tags the cases are considered as matching, otherwise if there is no overlapping the label corresponds to mismatching, and an overlapping with less than (or equal to) 10 tags is considered as partially matching. The dataset provides 6K legal case pairs, with rationales and explanations (the concatenation of all the overlapped tags) for the matching labels. Yu et al. (2022b) also provide *ELAM*, a dataset for explainable legal case matching task, containing 5K legal case pairs with the associated matching label, rationales, their alignments and the explanations for the matching decision. The authors collected the legal cases online,⁵⁵ which refer to the crime of obstruction of the social management order. Each case is associated with several legal-related tags. To pair the legal cases, the authors randomly selected 1250 query cases and constructed a pool of candidates for each query. From the candidate pool, a case is retrieved based on the number of overlapping tags between the case and the query. Each sentence of a legal case pair is associated with a rationale label, with the support of legal experts. The possible rationale labels are the following: not a rationale, a key circumstance, a constitutive element of a crime, or a focus of disputes. The alignment of the rationales (i.e., pro and con rationales) and the matching label (matching, partially matching or not matching) are then marked. Legal experts are also asked to provide explanations for their matching decision.

ILDC (Indian Legal Documents Corpus) (Malik et al., 2021) comprises about 35K cases from the Indian Supreme Court, annotated with the court decisions.⁵⁶ This is a corpus intended for court judgment prediction and explanation, which

⁵³ <https://wenshu.court.gov.cn/>

⁵⁴ <https://github.com/ruc-wjyu/IOT-Match>

⁵⁵ <https://www.faxin.cn/>

⁵⁶ <https://github.com/Exploration-Lab/CJPE>

requires a model to predict the final outcome (accept or reject, w.r.t. the appellant) and to provide explanations for the given prediction. In this regard, a portion of the corpus is annotated with explanations given by legal experts, and ranked in order of importance in such a way that a higher rank corresponds to an explanation that is more important for the final judgment. The dataset is divided in $ILCD_{single}$ and $ILCD_{multi}$, depending on whether there is a single decision for documents having one or more petitions, or different decisions for documents with multiple appeals.

Kalamkar et al. (2022) propose a corpus of 354 Indian legal judgment documents, annotated via crowd-sourcing activity with 12 different rhetorical roles, from different courts (Supreme Court of India, High Courts and district-level courts).⁵⁷ The annotation process is designed with the support of legal experts. The corpus is intended for the automatic structuring of legal documents. A Transformer-based model is proposed as baseline for the benchmark. Moreover, the authors propose extractive/abstractive summarization and court judgment prediction tasks as two applications of rhetorical roles, as they test how rhetorical roles could be useful for those tasks. For extractive and abstractive summarization, they experiment with the *LawBriefs* corpus, which comprises 285 expert-authored extractive summaries of Indian court judgments. For the court judgment prediction task, experiments were conducted using the ILDC corpus (Malik et al., 2021).

Two further legal datasets for rhetorical role identification are introduced in Bhattacharya et al. (2021). One dataset contains 50 cases from the Supreme Court of India belonging to five law domains: criminal, land and property, constitutional, labour/industrial and intellectual property rights. Such documents are gathered from Thomson Reuters Westlaw India website.⁵⁸ The other dataset contains 50 cases from the UK Supreme Court, gathered from the official website of the court.⁵⁹ Both the datasets are labeled with the following seven rhetorical roles: “Facts”, “Ruling by Lower Court”, “Argument”, “Statute”, “Precedent”, “Ratio of the decision” and “Ruling by Present Court”.

Paul et al. (2022b) introduce a pre-training corpus consisting of about 5.4M Indian court cases. The documents are gathered from several web platforms and come from the Supreme Court and many High Courts of India. The corpus covers various court case domains as well as more than 1K central government acts. The authors further pre-train Legal-BERT@aub and Legal-BERT@stanford on the proposed corpus and assess its pre-training effectiveness considering several downstream benchmarks, for the Indian as well as English languages. The performance of the pre-trained models have been compared to the BERT, the original Legal-BERT@aub and Legal-BERT@stanford.

Bhattacharya et al. (2019a) gather about 17K legal cases of the Supreme Court of India through the website of Westlaw India, which provides documents and related summaries written by domain experts. The authors perform a systematic comparison of several summarization algorithms, such as traditional unsupervised extractive

⁵⁷ <https://legal-nlp-ekstep.github.io/Competitions/Rhetorical-Role/>

⁵⁸ <http://www.westlawindia.com>

⁵⁹ <https://www.supremecourt.uk/decided-cases/>

methods (e.g., latent semantic analysis), neural unsupervised extractive methods (e.g., *Restricted Boltzmann Machines* (Verma and Nidhi, 2018)), and summarization methods specifically conceived for legal documents, both unsupervised (*CaseSummarizer* (Polsley et al., 2016)) and supervised (*LetSum* (Farzindar and Lapalme, 2004)).

Shukla et al. (2022) provide three legal summarization datasets⁶⁰ gathering documents from the Indian and UK laws. The first dataset is *Indian-Abstractive dataset (IN-Abs)*, with about 7K cases of Indian Supreme Court judgments, obtained from the website of *Legal Information Institute of India*,⁶¹ and corresponding abstractive summaries. The second dataset is *Indian-Extractive dataset (IN-Ext)*, with 50 case documents of the Indian Supreme Court labeled with six rhetorical roles (i.e., facts, argument, statute, precedent, ratio of the decision, and ruling by present court) and extractively summarized by domain experts, providing a summary for each rhetorical segment separately (with the exception of ratio and precedent segments that are summarized together). The third dataset is *UK-Abstractive dataset (UK-Abs)*, with 793 case judgments gathered from the website of the UK Supreme Court, which provides also the press (abstractive) summaries of the cases, divided in three segments: “Background to the Appeal”, “Judgment”, and “Reasons for Judgment”. The authors specify three criteria for the evaluation of methods: document-level summaries, segment-wise evaluations (i.e., how the summary represents the logical rhetorical segments in the legal case), and how the summaries are evaluated by domain-experts.

Niklaus et al. (2022) augment the Swiss-Judgment-Prediction (SJP) dataset introduced in Niklaus et al. (2021) via machine translation, i.e., translating a document written in one of the three languages (German, Italian, French) into the remaining two languages. A second version of the dataset is also provided by further augmenting SJP with Indian cases of the ILDC corpus, provided by Malik et al. (2021). In this regard, they translate all the Indian cases reported in the corpus to German, French and Italian. The authors evaluate several TLMs in relation to cross-domain (i.e., different legal areas), cross-regional (i.e., different regions) and cross-jurisdiction (from Indian to Swiss) transfer learning, whose discussion is demanded to Section 3.2.6.

LEX Rosetta (Savelka et al., 2021) propose a multi-lingual dataset of about 89K annotated sentences for the task of *automatic functional segmentation*, i.e., segmenting adjudicatory decisions of cases according to the functional role of the parts.⁶² The sentences are from 807 documents of several courts, gathered from different sources that include seven countries (Canada, Czech Republic, France, Germany, Italy, Poland, USA), and annotated according to the following types: out of scopes (i.e., sentences that are outside the main document, such as editorial content and appendices), heading (i.e., markers of a section), background (i.e., sentences explaining facts, claims and procedural background), analysis (i.e., sentences containing the court reasoning and applications of law to the facts), introductory summary (i.e., a

⁶⁰ available at <https://github.com/Law-AI/summarization>

⁶¹ <http://www.liiindia.org/in/cases/cen/INSC/>

⁶² https://github.com/lexrosetta/caselaw_functional_segmentation_multi-lingual

summary of the discussed case), and outcome (i.e., sentences describing the final decision). The dataset is used to test whatever GRU-based models generalize on different contexts (countries), in the segmentation of cases in three functional types (Background, Analysis and Outcome). To this end, the authors analyze the use of multi-lingual sentence embeddings of predictive models in three versions: training the model on a single context and evaluating transfer learning on other unseen contexts; training the model on a set of contexts and evaluating transfer learning on other unseen contexts; and pooling the data of the target context with data from the other contexts. Results have shown that the second and third versions of the model are more effective.

Law code data. SARA (Holzenberger et al., 2020) is a dataset for *statutory reasoning* on US tax law.⁶³ This dataset is comprised of a set of rules extracted from the statutes of the US Internal Revenue Code (IRC),⁶⁴ along with a set of questions which would require to refer to the rules for being answered correctly. In fact, IRC contains rules and definitions for the imposition and calculation of taxes, and it is subdivided into sections defining one or more terms (e.g., employment, employer and wages). Each section is normally organized around a general rule, followed by a number of exceptions, and each of its subsections refers to a certain number of slots, which may be filled by existing entities. IRC can hence be framed as a set of predicates formulated in human language so as to require a system to determine whether a subsection applies, and to identify and fill the slots mentioned. Statutory reasoning is addressed as a task of entailment and a task of question answering. In the first task, two paragraphs are manually created for each subsection as test cases: the one describes a case to which the statutes apply, the other one describes a case to which the statutes do not apply. In the second task, test cases are created to predict how much tax a person owes, considering all the statutes and applying arithmetic calculations. In general, this dataset offers some features that allow for reasoning on several aspects, such as reasoning on time, numbers, cross-reference and common-sense. To test the abilities of NLP models on the statutory reasoning problem, the authors have pre-trained the models (e.g., Legal-BERT@jhu) on a large legal corpus obtained extracting tax law documents from the *Caselaw Access Project*,⁶⁵ private letter ruling from the *International Revenue Service* (IRS)⁶⁶ and unpublished US tax Court cases.

BSARD (Louis and Spanakis, 2022) is a French dataset composed of more than 1.1K legal questions labeled by domain experts with relevant articles selected from the 22K law articles gathered from 32 publicly available Belgian codes.⁶⁷ The set of questions and associated relevant articles are obtained in collaboration with *Droits Quotidiens* (DQ), an organization composed of a team of experienced jurists, which

⁶³ <https://nlp.jhu.edu/law/>

⁶⁴ <https://uscode.house.gov/browse/prelim@title26&edition=prelim>

⁶⁵ <https://case.law/>

⁶⁶ <https://www.irs.gov/tax-exempt-bonds/teb-private-letter-ruling-some-basic-concepts>

⁶⁷ <https://github.com/maastrichtlawtech/bsard>

every year receives many questions from citizens seeking advice on legal issues, retrieves the articles relevant to the questions asked, answers the questions in a manner comprehensible to the applicant and categorizes the set of questions, legal references and answers with tags. The resulting corpus contains a large number of legal topics (related to social security, work, family, justice and so on). BSARD is intended for statutory article retrieval.

GCL (Papaloukas et al., 2021) is a dataset of about 47K documents regarding Greek legislation and designed for the multi-granular topic classification task, which requires to detect the thematic topic that is representative of a legal document.⁶⁸ The thematic topics are available in multi-level hierarchy. The main data source for this dataset is the *Permanent Greek Legislation Code - Raptarchis*, a catalogue of Greek legislation available through the portal e-Themis.⁶⁹ The portal provides a thematic index for the catalogue, reflecting the thematic hierarchical categories (topics). The hierarchy is dictated by the structural division in volumes, chapters and subjects, which reflect the levels of thematic topics. The classification task in GLC is divided into three sub-tasks, each of them deals with a level of the hierarchy.

Besides statutes, several benchmarks have been developed about *contracts* of different types. *CUAD* (Hendrycks et al., 2021b) is a dataset specialized for *legal contract review*.⁷⁰ It includes more than 500 contracts, varying in type and length, with 13K annotations across 41 category labels provided by legal experts. Such category labels regard general information, such as party names, dates, renewal terms and so on, as well as restrictive covenants and revenue risks. Language models are required to detect the portions of a contract (the clauses) related to each label. Evaluations of such models as BERT, ALBERT, RoBERTa and DeBERTa have highlighted that better performance are influenced by model design and training set size.

Leivaditi et al. (2020) provide a dataset containing 179 annotated documents regarding lease contracts. The annotations consist of entities (related to parties, property, terms and rent conditions, dates/periods) and red flags, i.e., terms or sentences indicating a potential risk for one or more parties (e.g., break option, guarantee transferable, right of first refusal to lease, bank guarantee), so that the dataset is mainly intended for supporting red flag detection and entity extraction tasks. The documents in the dataset are gathered through the EDGAR database, which is accessible through the US Securities and Exchange Commission (SEC).⁷¹ The process of selecting the contracts to be annotated is performed using the BM25 ranking function, which evaluates the relevance of documents w.r.t. keywords/queries that may suggest the presence of red flags. The identification of such keywords/queries and the process of annotation are supervised by domain experts.

⁶⁸ https://huggingface.co/datasets/greek_legal_code

⁶⁹ <https://www.secdigital.gov.gr/e-themis/>

⁷⁰ <https://github.com/TheAtticusProject/cuad/> The CUAD dataset is also available at atticusproject.ai.org/cuad

⁷¹ <https://www.sec.gov/edgar.shtml>

ContractNLI (Koreeda and Manning, 2021) contains 607 annotated contracts regarding non-disclosure agreements.⁷² Such documents are gathered from Internet search engines and EDGAR. By comparing different non-disclosure agreements, a set of 17 hypotheses is obtained. Each document is annotated with respect to its relation with one of the hypotheses (i.e., entailment, contradiction, or not mentioned). If a document is annotated as entailing or contradicting, the spans (i.e., sentence or list item within a sentence) composing the documents are annotated as evidence or not (binary label) of the associated entailment relationship.

ToS (Aumiller et al., 2021) is a dataset consisting of Term-of-Service documents, specifically collected for *topic similarity* task. The documents include heterogeneous topics due to the different web sources. Some of the most frequent topics regard limitation of liability, law and jurisdiction, warranty, and privacy. Topics are obtained in a hierarchical way splitting the documents into smaller chunks. The authors define and test a system built on TLMs, which revealed to largely outperform segmentation baselines based on TF-IDF and bag-of-words.

MAUD (Wang et al., 2023a) is a dataset for the legal multi-choice reading comprehension task and consists of legal texts extracted from 152 public merger agreements gathered from EDGAR. Merger agreements are legal documents regarding public target company acquisitions. In these documents there are special clauses, called deal points, that establish the conditions under which the parties are obliged to complete the acquisition. The deal points are extracted from the merger agreements by lawyers working on the American Bar Association’s 2021 Public Target Deal Points Study (“ABA Study”). Moreover, a set of multiple-choice questions are answered by the lawyers for each deal point. One or more questions can be asked for a deal point, and each question can be answered by one or more answers. MAUD contains 92 questions, 8K unique deal points annotations, 39K question-answer annotations (the examples) and 7 deal point categories (e.g., Conditions to Closing, Deal Protection and Related Provisions, Material Adverse Effect).

Manor and Li (2019) provide a dataset containing legal contracts and summaries gathered from two websites, *TL;DRLegal*⁷³ and *TOS;DR*,⁷⁴ the purpose of which is to clarify the content of contracts through summaries. More precisely, the former, which deals mainly with software licences of companies, is used as a source for collecting 84 sets of contract agreement sections and corresponding summaries, whereas 412 sets are obtained from TOS;DR website, which focuses on user data and privacy topics of companies. The quality of the proposed summaries is verified by authors through an analysis of levels of abstraction, compression and readability.

Another important target of interest for the development of benchmarks is represented by *privacy policies*. *PolicyIE* (Ahmad et al., 2021) is a corpus for automating fine-grained information extraction of privacy policies, especially through intent classification and slot filling tasks.⁷⁵ PolicyIE consists of about 5K intent and 11K

⁷² <https://stanfordnlp.github.io/contract-nli/>

⁷³ <https://tldrlegal.com/>

⁷⁴ <https://tosdr.org/>

⁷⁵ <https://github.com/wasiahmad/PolicyIE>

slot annotations of several privacy policies related to website and mobile applications. The retrieved policy documents cover four privacy practices that are included in the *General Data Protection Regulation* (GDPR). Thus, sentences of such policy documents are categorized into the following GDPR-like intent classes: data collection/usage (i.e., what user information is collected, as well as the reason and the modality in which it collected), data sharing/disclosure (i.e., what user information is shared with third parties, as well as the reason and the modality in which it is shared), data storage/retention (i.e., location and time period in which user information will be saved), data security/protection (i.e., what protection measures for user information are taken), other (i.e., privacy practices not included in the other categories). Each sentence is annotated with 18 slot labels, which can be categorized into two overlapping types: type-We, which comprises data and participants to the policy practices (e.g., data provider, data collected, data collector) and type-II, i.e., purposes, conditions, polarity and protection methods. The annotation procedure was performed and monitored by domain experts.

PrivacyQA (Ravichander et al., 2019) contains 1750 questions with over 3.5K annotations of relevant answers regarding to privacy policies of mobile applications.⁷⁶ Questions for a particular privacy policy are provided by crowd-workers, while the identification of the related answers are committed to legal experts, which also provide meta-annotations on the relevance of the question, OPP-115 category, subjectivity, and the likelihood that the answer to the input question is contained into a privacy policy. The authors test the ability of different baselines on two tasks: deciding if a question is answerable, and identifying evidences in the policies for a given question.

PolicyQA (Ahmad et al., 2020) comprises about 25K triplets of question, passage and answer text, derived from segments of website privacy policy documents.⁷⁷ The corpus is designed so that the answer consists of small portions of the text that better identify the target information in relation to the question. It is curated from the existing *OPP-115* corpus (Wilson et al., 2016), which consists of 115 website policies (about 3.7K segments) annotated following domain-experts annotation schemes. The annotation schemes categorize the policy segments in ten data practice categories (e.g., first party collection/use), which are further categorized in several practice attributes (e.g., user type), and each practice attribute is assigned a set of values; for instance, user without account, user with account, other and unspecified. The annotated segments with the associated practice, attribute and value categories are used to form the *PolicyQA* corpus. Segments and categories are provided to skilled annotators to manually label the questions, for a total of 714 individual questions. The associated Question-Answering task is answer span prediction given a policy segment. In this regard, two neural baselines are evaluated, one of this is based on BERT.

⁷⁶ https://github.com/AbhilashaRavichander/PrivacyQA_EMNLP

⁷⁷ <https://github.com/wasiahmad/PolicyQA>

Bui et al. (2021) introduce a corpus⁷⁸ for the extraction and visualization in privacy policies of personal data objects, i.e., spans in the text expressing types of user data, and related privacy actions. The proposed corpus contains about 4.1K sentences and 2.6K annotated fine-grained data objects concerning several real-world privacy policies. It is obtained exploiting the OPP-115 dataset as a starting point, opting for the top US websites that cover several domains like banking, e-commerce, social network. The data objects in the privacy policies are detected by annotators with experiences in privacy and security research. The data objects are then labeled by the annotators, choosing among “collect”, “not_collect”, “share” and “not_share” labels. Such labels indicate the privacy action performed on the user data (collection or sharing). The resulting annotation has also been revised with a semi-automated process to improve the annotation quality, which involves the use of tools for correction and pre-annotation. The final corpus is used to train and evaluate a neural NER model, called *PI-Extract*, on the extraction of personal data objects and privacy actions. The task is formulated as a sequence labeling problem, which is to assign a label for each token of a given sentence.

Relevant benchmarks have been built by considering multi-lingual and/or multi-task evaluation scenarios. For instance, *COVID-19 Exceptional Measures* (Tziafas et al., 2021) is a collection of legal, manually-annotated documents regarding COVID-19 exceptional measures across 21 European countries for multi-lingual classification task. To this end, feature-based methods and XLM-RoBERTa pre-trained on the collection have been evaluated, showing best results in the use of the domain-adapted TLMs.

MultiEURLEX (Chalkidis et al., 2021a) consists of European Union laws, annotated with multiple labels and translated in 23 languages, with legal topic classification as supported task.⁷⁹ The authors experiment with monolingual BERT models, pre-trained in Romance, Slavic, Germanic and Uralic languages, and multi-lingual models (mT5 and XLM-RoBERTa) which are evaluated for cross-lingual legal text classification on this benchmark. The experimentation focuses mainly on the zero-shot cross-lingual transfer, namely one-to-many setting, in which a multi-lingual model is fine-tuned on one language and evaluated in the other 22 languages. However, models are also evaluated on one-to-one (training and testing on the same language) and many-to-many (training and testing on all languages) settings. Adaptation strategies on the multi-lingual models are applied to avoid the catastrophic forgetting of multi-lingual knowledge when fine-tuning on one source language only, significantly improving the zero-shot cross-lingual transfer. In the one-to-one setting, multi-lingual models prove to be competitive against monolingual models.

EUR-Lex-Sum (Aumiller et al., 2022) is a multi- and cross-lingual dataset containing about 32K pairs of documents and summaries in 24 languages.⁸⁰ Each language comprises up to 1500 pairs. Documents consist of legal acts retrieved from the

⁷⁸ https://github.com/um-rtcl/piextract_dataset

⁷⁹ https://huggingface.co/datasets/multi_eurlex

⁸⁰ <https://github.com/achouhan93/eur-lex-sum>

European Union law website,⁸¹ 375 of which are legal acts written in each of the languages. Summaries are structured following particular guidelines, for example there are sections dedicated to key points, background, key terms and so on. The authors evaluate several zero-shot extractive baselines, one of which is a version of LexRank that receives chunks (based on existing separators in the text) and uses embeddings generated by S-BERT, and cross-lingual baselines, including one based on LED with capability of greedily chunking the text if document sizes exceed the model’s maximum input length.

LegalNERo (Pais et al., 2021) contains 370 legal documents, designed for NER tasks and manually annotated according to five coarse-grained classes: person, location, organization, time expressions, and legal document references.⁸² The documents are extracted from a larger corpus, called *MARCELL-RO*,⁸³ containing several documents from national legislation (e.g., decrees, regulation, laws) of seven countries, Romania included. The authors evaluate a baseline based on BiLSTM and CRF, which takes as input a text representation obtained through FastText.⁸⁴

PLUE (Chi et al., 2023) is a multi-task benchmark that collects several privacy policy datasets (including the aforementioned datasets) to evaluate NLP methods over various privacy policy tasks, namely classification, question-answering, intent classification, slot filling, name entity recognition.⁸⁵ In particular, PLUE contains the following datasets: OPP-115, APP-350 (Zimmeck et al., 2019), PrivacyQA, PolicyQA, PolicyIE, and PI-Extract (the dataset used in Bui et al. (2021)). To enable model pre-training on the domain, Chi et al. (2023) also provide a large corpus using *MAPS* (Zimmeck et al., 2019), a corpus of 441K mobile application privacy policies, and the *Princeton-Leuven Longitudinal Corpus* (PLLC) (Amos et al., 2021), containing about 130K privacy policies of websites. From the combination of the two corpora, a pre-training corpus with 332M words is created. The authors evaluated several TLMs as baseline, previously pre-trained on MAPS and PLLC and then fine-tuned on the PLUE datasets.

Drawzeski et al. (2021) introduce a multi-lingual corpus for the analysis of *fairness of online terms of service*. The dataset contains 100 contracts, derived from 25 ToS documents annotated in four languages (English, German, Italian and Polish) and extracted from an existing corpus (Lippi et al., 2019). In each contract, potentially unfair clauses are labeled with one of the nine possible unfairness categories, namely arbitration, unilateral change, content removal, jurisdiction (i.e., which courts will have jurisdiction over disputes under the contract), choice of law (i.e., which law will regulate the contract), limitation of liability, unilateral termination, contract by using (i.e., the use of a service binds the consumer to the terms of use of the service without being required to indicate that she/he has read and accepted them), and privacy included (i.e., the use of the service implies the acceptance of the

⁸¹ <https://eur-lex.europa.eu/>

⁸² <https://lod-cloud.net/dataset/racai-legalnero>

⁸³ <https://marcell-project.eu/>

⁸⁴ <https://fasttext.cc/>

⁸⁵ <https://github.com/JFChi/PLUE>

related privacy policy). Moreover, for each category the degree of the unfairness is indicated with three numerical values: 1 for clear fairness, 2 for potential unfairness, and 3 for clear unfairness. Four types of discrepancies are observed across the language versions of the same contract, relating to the sentence structures, errors or inaccuracies in translation into the target languages, absence of some clauses in the different language versions and the choice of legal terminology.

Hybrid data. *LexGLUE* (Chalkidis et al., 2022b) is a collection of seven existing legal NLP datasets⁸⁶ for evaluating models across several legal tasks, which include multi-label classification, multi-class classification and multiple choice question-answering:

- *ECtHR Tasks A & B* for multi-label classification of allegations regarding violations of the European Convention of Human Rights (ECHR) provisions. The dataset is used to test models on article violation prediction (Task A, (Chalkidis et al., 2019a)) and alleged violation prediction (Task B, (Chalkidis et al., 2021c)). In both Task A and Task B, the total number of ECHR articles is reduced to 10, discarding those that are rarely discussed, cannot be violated or are not depending on the facts of a case.
- The English part of *Multi-EURLEX* (Chalkidis et al., 2021a) for multi-label classification on European Union (EU) legislation. It includes different labeling granularity levels (from 21 to 7K EuroVoc concepts). In LexGLUE, the 100 most frequent labels from the second level of granularity (567 total labels) are considered.
- *SCOTUS*⁸⁷ for multi-class classification on US Supreme Court opinions. In LexGLUE, SCOTUS opinions are associated with 14 issue areas (e.g., Economic Activity, Criminal Procedure, Civil Rights) obtained through the Supreme Court DataBase.⁸⁸
- *LEDGAR* (Tugener et al., 2020) for multi-class classification on contract provisions gathered from the US Securities and Exchange Commission (SEC) documents. In LexGLUE, a subset of the original dataset is considered, derived from the 100 most frequent labels.
- *UNFAIR-ToS* (Lippi et al., 2019) for multi-label classification on contracts between providers and users of services (i.e., terms of service) with 8 classification labels.
- *CaseHOLD* (Zheng et al., 2021) for multiple choice question-answering about holdings of US court cases gathered from the Harvard Law Library case law corpus.

Chalkidis et al. (2022b) evaluate BERT, RoBERTa, DeBERTa, Longformer, BigBird, Legal-BERT@aub and Legal-BERT@stanford on LexGLUE. For ECtHR and SCOTUS, the authors employ a hierarchical variant of the models, following Chalkidis et al. (2021c). The domain-adapted models, i.e., Legal-BERT@aub

⁸⁶ https://huggingface.co/datasets/lex_glue

⁸⁷ <https://www.supremecourt.gov/>

⁸⁸ <http://supremecourtdatabase.org/>

and Legal-BERT@stanford, performed overall better than competitors, with large improvement in US case law data.

FairLex (Chalkidis et al., 2022c) comprises four datasets (ECtHR, SCOTUS, Swiss-Judgment-Prediction and CAIL (Wang et al., 2021b)) for the evaluation of *model fairness*.⁸⁹ To this end, it includes three groups of fairness attributes: demographics, regional and legal area. The first group regard biases relating to factors such as gender, age and race. The second and third groups aim to alleviate disparity, respectively, in regions of a given jurisdiction and in different areas of law. Moreover, it contains five languages (Chinese, English, French, German and Italian) and four jurisdictions (China, USA, Switzerland and European Council). The authors also provide four hierarchical BERT-based models, one for each dataset, as baselines for the benchmark. Such models are similar to Chalkidis et al. (2021c) and further pre-trained on the specific dataset. The models are warm-started from MiniLMv2 checkpoints, using a distilled version of RoBERTa for the English version and a distilled version of XLM-R for the other languages. Experimental results show that the models have some disparity in performance. In particular, in the ECtHR task there is a disparity related to defendant state and applicant’s gender, while for the FSCS task there is disparity related to language (Italian versus French and German), legal areas (penal law versus the others) and court regions (Switzerland courts versus federation courts). Court regions disparity is noted also in the CAIL task (Beijing courts versus Sichuan courts). However, disparities in performance can be influenced by general factors based on the distribution of data.

Bhattacharya et al. (2020) collect documents of the Supreme Court of India and statutes in the Indian judiciary through the Thomson Reuters Westlaw India website,⁹⁰ for a task of document similarity. In particular, they propose and evaluate an approach based on *precedent citation network* augmented with the hierarchy of legal statutes, in order to encompass also the knowledge of legal text’s hierarchy.

Pile of Law (Henderson et al., 2022) is a legal corpus designed to address *ethical issues* in the pre-training phase.⁹¹ It is collected from 35 EU and US data sources and covers several legal sub-domains, such as court opinions, administrative rules, contract and legislative records, for a total of about 10M documents. Such data has already implicit filters which reflect legal standards of the specific jurisdiction, but the authors note that not all of them have been detected and such norms can vary respect to the jurisdiction. By performing filtering rules on the data, the proposed dataset respects the legal norms of governments and courts regarding the presence of toxic (offensive or obscene terms) or private content and prevents a model to learn such information. The authors demonstrate that the dataset can be used to learn contextual privacy/toxicity rules, as it respects the variation in the different privacy/toxicity norms. For example, they demonstrate models pre-trained on Pile of Law can learn contextual privacy rules with regard to the use of pseudonyms in immigration court and in civil litigation. In particular, a BERT-based model is trained

⁸⁹ <https://huggingface.co/datasets/coastalcph/fairlex>

⁹⁰ <https://www.westlawasia.com/>

⁹¹ <https://huggingface.co/datasets/pile-of-law/pile-of-law>

on the data to predict whether a pseudonym should be used. Moreover, Henderson et al. (2022) provide a baseline by pre-training BERT on the corpus from scratch. The resulting model, called *PoL-BERT-Large*, is fine-tuned and evaluated for a legal reasoning task on CaseHOLD, reaching about the same performance reported in LexGLUE, and outperforms BERT. However, it does not outperform a BERT model trained exclusively on case law data. This is probably due to the extreme data diversity in the corpus that limits the pre-training efficacy respect to competitors trained exclusively on task-oriented data such as CaseHOLD.

Table 2.4: Summary of benchmarks and datasets discussed in Section 2.4.2.

Reference	Category	Data type	Data source	Data size	Tasks
Strickson and Iglesia (2020)	cases	UK court judgments	web sources	5K documents	LJP
Chalkidis et al. (2019a)	cases (allegations)	ECHR cases	ECHR's public database (HUDOC)	11.5K documents	LJP (binary violation), ALVP, case importance prediction
Chalkidis et al. (2021c)	cases (allegations)	ECHR cases	ECHR's public database (HUDOC)	11K documents	ALVP
Niklaus et al. (2021)	cases	Federal Supreme Court of Switzerland cases	entscheidsuche platform	85K documents	LJP
Wrzalik and Krechel (2021)	cases	case laws from Open Legal Data	Open Legal Data platform	123K query passages, 131K documents	LPR
Urchs et al. (2021)	cases	cases from Bavarian courts	gesetze-bayern platform	32K docs, 200 docs with 25K sentences	automatic detection of sections
Zhong et al. (2019b)	cases (appeals)	US Board of Veterans' appeals cases	US Department of Veterans Affairs platform	92 documents	ES
Walker et al. (2019)	cases (appeals)	US Board of Veterans' appeals cases	US Department of Veterans Affairs platform	50 documents, 6K sentences	SC
Shen et al. (2022)	cases	writings providing information on US federal civil rights cases	Civil Rights Litigation Clearinghouse platform	9K documents	AS
Zheng et al. (2021)	cases (holdings)	US court cases	existing corpus (Harvard Law Library corpus)	53K questions	MCQA
Feijó and Moreira (2018)	cases	Brazil federal court decisions	Supremo Tribunal Federal platform	10K documents	ES
Lage-Freitas et al. (2022)	cases	Brazilian State higher court cases	web sources	4K documents	LJP, unanimity decision

Continued on the next page

Reference	Category	Data type	Data source	Data size	Tasks
Xiao et al. (2019)	cases	Supreme People's Court of China cases	China Judgments Online platform	8K document triplets	CM
Yu et al. (2022b)	cases	Supreme People's Court of China cases	existing dataset (CAIL)	6K documents pairs	explainable CM
			Faxin platform	5K documents pairs	
Malik et al. (2021)	cases	Supreme Court of India cases	IndianKanoon platform	35K documents	CJPE
Kalamkar et al. (2022)	cases	Indian courts' legal judgments	Indian court platforms (Supreme Court of India platform, eCourt India services platform)	354 documents, 40K sentences	automatic legal document structuring
Bhattacharya et al. (2021)	cases	Supreme Court of India cases	Thomson Reuters Westlaw India platform	50 documents	RRL
		UK Supreme Court	UK Supreme Court platform	50 documents	
Paul et al. (2022b)	cases	Supreme Court and High Courts of India cases	web platforms (Supreme Court of India platform, IndianKanoon platform)	5.4M documents	pre-training
Bhattacharya et al. (2019a)	cases	Supreme Court of India cases	Thomson Reuters Westlaw India platform	17K documents	AS
Bhattacharya et al. (2020)	cases	Supreme Court of India cases, statutes in the Indian judiciary	Thomson Reuters Westlaw India platform	1.8K documents	document similarity
Shukla et al. (2022)	cases	Supreme Court of India case judgments, UK Supreme Court case judgments	Legal Information Inst. of India platform, UK Supreme Court platform	~7K docs (India), 793 docs (UK)	AS, ES
Niklaus et al. (2022)	cases	Federal Supreme Court of Switzerland	existing datasets	~180K Swiss docs, 96K Indian docs	LJP, cross-domain/regional/

Continued on the next page

Reference	Category	Data type	Data source	Data size	Tasks
		and Indian cases	(SJP and ILDC)		jurisdiction transfer learning
Savelka et al. (2021)	cases	documents from several courts	web platforms, random courts (CA); Constitutional/Supreme/Supreme admin. courts (CZ); Cour de cassation (FR); federal courts (DE); criminal courts (IT); Supreme Court, Constitutional tribunal (PL); Federal district court, Dept. of Labor (US)	807 documents with 89K sentences	functional segmentation
Holzenberger et al. (2020)	statutes (rules)	US tax law	Office of the Law Revision Counsel platform - US Internal Revenue Code	276 train, 100 test docs	entailment and tax amount prediction
Louis and Spanakis (2022)	statutes	Belgian codes (federal/regional authorities)	Droits Quotidiens organization	1.1K questions, 22K articles	SAR
Papaloukas et al. (2021)	statutes, regulations	Greek legislation (laws, Royal/Presidential decrees, regulations and decisions, from the Official Government Gazette)	Permanent Greek Legislation Code (Raptarchis) through e-Themis platform	47K documents	multi-level TpC
Hendrycks et al. (2021b)	contracts	25 types of legal contracts (e.g., license, consulting, service, manufacturing)	SEC's EDGAR database	510 documents	legal contract review, relevant text prediction
Leivaditi et al. (2020)	contracts	lease contracts	SEC's EDGAR database	179 documents	sentence detection, entity extraction
Tuggener et al. (2020)	contracts	Exhibit-10 material contract agreements	SEC's EDGAR database	60K documents	text classification

Continued on the next page

Reference	Category	Data type	Data source	Data size	Tasks
Koreeda and Manning (2021)	contracts	non-disclosure agreements	web sources and SEC's EDGAR database	607 documents	NLI
Aumiller et al. (2021)	contracts	terms of service	web sources	40K documents	topic similarity
Wang et al. (2023a)	contracts	merger agreements	SEC's EDGAR database	39K question-answer annotations	MCRC
Manor and Li (2019)	contracts	contracts relating to software licenses, privacy policies, terms of service/use	web platforms (TL;DRLegal, TOS;DR)	84 agreement-summary pairs, 421 agreement-(up to 3)summary pairs	abstractive similarity
Ahmad et al. (2021)	contracts	privacy policies	mobile applications and web sources	5.2K sentences	intent classification, SF
Ravichander et al. (2019)	contracts	privacy policies	mobile applications	1.7K questions, 3.5K sentences	question answerability, evidence extraction
Ahmad et al. (2020)	contracts	privacy policies	web sources	25K triples	answer span prediction, QA
Bui et al. (2021)	contracts	privacy policies	existing dataset (OPP-115) based on web sources	4.1K sentences	personal data object extraction, NER
Tziafas et al. (2021)	regulations	COVID-19 exceptional measures	web platforms	4K sentences	SC
Chalkidis et al. (2021a)	statutes, regulations	European Union laws	EUR-Lex platform	65K documents	TpC
Aumiller et al. (2022)	statutes, regulations	legal acts of 20 domains (e.g., taxation, energy, transportation, industrial policy, health protection)	EUR-Lex platform	32K doc-summary pairs, ~375 docs as val/test sets for all languages, remaining docs as lang.-specific train set	AS, ES
Pais et al. (2021)	statutes, regulations	decrees, regulation, laws of several countries	existing dataset (MARCELL-RO)	370 documents	NER

Continued on the next page

Reference	Category	Data type	Data source	Data size	Tasks
			based on the Romanian legislative platform	370 documents	NER
Chi et al. (2023)	contracts	privacy policies	existing datasets (OPP-115, APP-350, PrivacyQA, PolicyQA, PolicyIE, PI-Extract) related to mobile applications and web sources	18.7K segments, 1.7K questions, 25K triples, 9.2K sentences	intent classification, SF, QA, NER
Lippi et al. (2019)	contracts	terms of service	web platforms	9.4K sentences	clause detection and type classification
Drawzeski et al. (2021)	contracts	terms of service	existing dataset (UNFAIR-ToS) related to web platforms	100 documents	sentence-level annotations projection
Chalkidis et al. (2022b)	cases (allegations, holdings), contracts	ECHR cases, US Supreme Court opinions (Criminal Procedure, Civil Rights, Economic Activity), US court cases, terms of service	existing datasets (ECtHR Tasks A & B, Multi-EURLEX, SCOTUS, LEDGAR, UNFAIR-ToS, CaseHOLD) based on web platforms,	~164 docs, 9.4K sentences, 53K questions	AVP, ALVP TpC, multi-class/label classification, MCQA
Henderson et al. (2022)	cases, regulations, contracts,	EU and US documents (Court Listener Opinions, US Board of Veterans' Appeals,	web sources, existing datasets	10M documents	pre-training

Continued on the next page

Reference	Category	Data type	Data source	Data size	Tasks
		EUR-Lex, EDGAR contracts, ECHR, US Code)			
Chalkidis et al. (2022c)	cases	ECHR cases, US Supreme Court opinions, Federal Supreme Court of Switzerland cases, Supreme People's Court of China cases	existing datasets (ECtHR, SCOTUS, SJP and CAIL), based on web platforms	209K documents	fairness on ALVP, TpC, case approval prediction, crime severity prediction

2.4.3 Data and benchmarks for psychosocial-domain

This section provides an overview of data and benchmarks relevant to the psychosocial domain, organized by text type: i) clinical notes, suicidal notes, last statements, social posts and text messaging for mental-health issues, ii) surveys, lexicons and annotated texts for morality. As mentioned in Section 2.3.3, the discussion on mental-health data and benchmarks is intentionally limited due to the relatively brief period dedicated to this area during the Ph.D. research (Greco et al., 2023).

Mental-health-related data and benchmarks. Existing studies on mental-health issues typically refer to four main types of data: clinical notes, suicidal notes, social media and text messaging.

Clinical notes. A major source of data refers to clinical notes and electronic health records (EHRs). EHRs contain information regarding patients in hospital-stay, including diagnoses and procedure codes (Meng et al., 2021), medications of specific treatments, and annotations about suicidal attempts or ideation (Rawat et al., 2022).

Suicidal notes and last statements. A related source comes from notes written by people who have committed or attempted suicide (Zhang et al., 2021a), and notes by people sentenced to death (Schoene et al., 2019).

Social media. Most existing approaches focus on the analysis of users’ opinions and statements posted on social media, with emphasis on Reddit and Twitter text contents; for instance, Twitter posts containing hashtags like #mentalhealth and #healthanxiety (Vajre et al., 2021) or suicidal words and phrases (Ji et al., 2018). Also, Matero et al. (2022) exploit the Facebook app *MyPersonality*, in which Facebook users are involved in a psychological test and are asked to fill a personality questionnaire.

Text messaging. Instant messaging also plays an important role, such as SMS-like texts hinting at factors of daily stressors (Mauriello et al., 2021) and conversations extracted from online services supporting people with emotional distresses (Liu et al., 2021d).

MFT-related data and benchmarks. The MFT is a cornerstone of datasets focusing on the morality field (Guo et al., 2023b). Insights from current studies on morality have primarily been obtained via analyzing moral vignettes, questionnaire data, and social media datasets (Hofmann et al., 2014). Therefore, we categorize the data sources into 3 categories: Moral surveys, Moral lexicons, MFT-annotated text (Zangari et al., 2024).

Table 2.5 provides a summary of the data sources utilized in our overview research (Section 7). These datasets represent some of the most suitable resources existing at the time of our study for integrating TLMs and MFT.

Table 2.5: Summary of the data sources used in the revised works (Dataset), specifying whether they are Moral survey, Moral lexicon or MFT-annotated dataset (Category), along with their size in terms of number of rows (Size). Symbol “-” indicates information unavailable or not applicable.

Dataset	Size	Labels	# annotators	Source
Moral survey				
MFQ (Graham et al., 2011)	30	5 (Care/Harm, Fair/Cheat., Loyal./Betr., Auth./Subv., Sanct./Degrad.)	-	Link
MFQ-2 (Atari et al., 2023)	36	6 (Care/Harm, Equal./Ineq., Propor./Disprop., Loyal./Betr., Auth./Subv., Sanct./Degrad.)	-	Link
Moral lexicon				
MFDv1 (Graham et al., 2009)	324	10 (Care, Harm, Fair., Cheat., Loyal., Betr., Auth., Subv., Sanct., Degrad.)	-	Link
MFDv2 (Frimer, 2019)	2014	10 (Care, Harm, Fair., Cheat., Loyal., Betr., Auth., Subv., Sanct., Degrad.)	-	Link
eMFD (Hopp et al., 2021)	3,270	5 (Care/Harm, Fair/Cheat., Loyal./Betr., Auth./Subv., Sanct./Degrad.)	-	Link
MoralStrength (Araque et al., 2020a)	1000	5 (Care/Harm, Fair./Cheat., Loyal./Betr., Auth./Subv., Sanct./Degrad.)	at least 5 per item	Link
MFT-annotated dataset				
MFTC (Hoover et al., 2020a)	35,108	10 (Care, Harm, Fair., Cheat., Loyal., Betr., Auth., Subv., Sanct., Degrad.)	at least 3 per item	Link
MFRC (Trager et al., 2022)	16,123	8 (Care/Harm, Equal./Ineq., Propor./Disprop., Loyal./Betr., Auth./Subv., Sanct./Degrad., Thin, Impl./Expl.)	at least 3 per item	Link
MFNC (Weber et al., 2021)	35,985	10 (Care, Harm, Fair., Cheat., Loyal., Betr., Auth., Subv., Sanct., Degrad.)	557	Link
Covid (Rojecki et al., 2021)	2,648	10 (Care, Harm, Fair., Cheat., Loyal., Betr., Auth., Subv., Sanct., Degrad.)	-	-
Congress (Johnson and Goldwasser, 2018)	2,050	10 (Care, Harm, Fair., Cheat., Loyal., Betr., Auth., Subv., Sanct., Degrad.)	2	-
Extended Congress (Roy and Goldwasser, 2021)	161,295	10 (Care, Harm, Fair., Cheat., Loyal., Betr., Auth., Subv., Sanct., Degrad.)	-	-
Facebook (Beiró et al., 2023)	4,498	12 (Care, Harm, Fair., Cheat., Loyal., Betr., Auth., Subv., Sanct., Degrad., Liber., Oppr.)	9	-

Continued on the next page

Dataset	Size	Labels	# annotators	Source
Covid-19 (Pacheco et al., 2022)	750	5 (Care/Harm, Fair./Cheat., Loyal./Betr., Auth./Subv., Sanct./Degrad.)	3	Link
Exteded ArgQuality (Kobbe et al., 2020)	320	5 (Care/Harm, Fair./Cheat., Loyal./Betr., Auth./Subv., Sanct./Degrad.)	2	Link
MoralArg (Alshomary et al., 2022)	230,000	5 (Care/Harm, Fair./Cheat., Loyal./Betr., Auth./Subv., Sanct./Degrad.)	-	Link
MIC (Ziems et al., 2022)	38,000	5 (Care/Harm, Fair./Cheat., Loyal./Betr., Auth./Subv., Sanct./Degrad.)	186 total, 3 per item	Link
MAAFS (McCurrie et al., 2018)	69	6 (Care/Harm, Fair./Cheat., Loyal./Betr., Auth./Subv., Sanct./Degrad., Liber./Oppr.)	575	Link
MFVs (Clifford et al., 2015)	132	6 (Care/Harm, Fair./Cheat., Loyal./Betr., Auth./Subv., Sanct./Degrad., Liber./Oppr.)	510	-

Table 2.6: Example of items for each moral foundation in the Moral Foundations Questionnaire (MFQ) (Graham et al., 2008).

Foundation	Example item
Care	“Compassion for those who are suffering is the most crucial virtue”
Fairness	“I think it’s morally wrong that rich children inherit a lot of money while poor children inherit nothing”
Authority	“Respect for authority is something all children need to learn”
Loyalty	“People should be loyal to their family members, even when they have done something wrong”
Purity	“I would call some acts wrong on the grounds that they are unnatural”
Liberty	“The government interferes far too much in our everyday lives”

Moral surveys. Moral surveys correspond to set of close-ended questions designed to gauge individuals’ moral beliefs. In this regard, The Moral Foundations Questionnaire (MFQ) (Graham et al., 2008) is a psychometric tool designed to evaluate the extent to which individuals endorse the five different moral foundations of the MFT. It is widely used in social psychology to understand moral diversity across cultures and to examine individual differences in moral scenarios. The MFQ evaluates each foundation by utilizing 30 questions and six items, which are graded on a 6-point Likert scale (Likert, 1932), anchored from 1 (strongly disagree) to 6 (strongly agree). Table 2.6 shows moral foundations with corresponding example items statements used in the MFQ to assess an individual’s moral values related to each foundation (Graham et al., 2008).

The Moral Foundations Questionnaire–2 (MFQ-2) (Atari et al., 2023) contains 36 questions built on top of the MFQ by refining Fairness/Cheating into two distinct dimensions: Equality/Inequality and Proportionality/Disproportionality. Equality/Inequality is a concept that promotes social justice and equality, while Proportionality/Disproportionality focuses on remuneration based on merit, contributing to virtues like productivity and deservingness, but also fostering corruption and nepotism. Like the MFQ, the MFQ-2 uses Likert scale assessments to gauge moral priorities, from 1 (“slightly describes me”) to 5 (“describes me extremely well”). The studies conducted to develop MFQ-2 has shown how the structure and influence of moral foundations vary across cultural contexts.

Moral Lexicons. Moral lexicons refer to lists of word-score pairs related to moral values. The Moral Foundations Dictionary (MFD) (Graham et al., 2009) is the first vocabulary developed to assess moral values from textual data. The MFD provides a list of words and phrases associated with the five moral foundation dimensions. The average number of words included in MFD for each moral category is 32, hence having a size of 324 words. MFDv2 (Frimer, 2019) is an extension of MFD, which consists of a lexicon of 2014 words, with each word associated to a single moral category. Another extension of the MFD is the extended Moral Foundation Dictionary (eMFD) (Hopp et al., 2021), a dictionary of text samples extracted from the *Moral Foundations News Corpus* (MNFC) (Weber et al., 2021), containing English news articles on a variety of topics. Lemmas and stems are annotated with a moral value and words in the eMFD are weighted according to a probabilistic scoring procedure.

MoralStrength (Araque et al., 2020a) is a semi-automated lexicon of approximately 1,000 lemmas, obtained as an extension of the MFD, which quantifies the relevance and strength of words related to the five moral foundations of the MFT. The lexicon is constructed by initially selecting seed words related to each moral foundation and then expanding this set through semantic similarity measures derived from word embeddings.

LibertyMFD (Araque et al., 2022) is in turn an extension of MoralStrength for detecting the Liberty moral dimension. Roy and Goldwasser (2021) extend the dataset proposed by Johnson and Goldwasser (2018) to build a topic indicator lexicon, for topics like Gun Control and Immigration.

MFT-annotated texts. MFT-annotated texts refer to data manually or (semi-) automatically labeled with the dimensions of the MFT. The Moral Foundations Twitter Corpus by Hoover et al. (2020a) comprises 35,108 tweets annotated with 11 moral values based on the MFT. Each tweet is categorized into 10 moral values, reflecting the dyadic dimensions such as care/harm, which are treated separately, i.e., care and harm are separate labels. Additionally, the dataset includes a *non-moral* label for tweets that do not exhibit any moral sentiment, resulting in 11 possible labels per tweet. It also includes seven different domains related to morally relevant issues: (i) All lives matter, referring to American social movement that emerged as a reaction to the Black Lives Matter movement of the African-American community; (ii) The Baltimore Protests, concerning tweets related to the protest over the murder of Freddie Gray in Baltimore; (iii) Black Lives Matter, including all tweets with the hashtag #BlackLivesMatter; (iv) 2016 Presidential Election, consisting of tweets collected from followers of US politicians and reputable newspaper agencies throughout the 2016 presidential US election; (v) Davidson, a collection of hate speech and inflammatory language (Davidson et al., 2017); (vi) Hurricane Sandy, encompassing all tweets that were shared prior to, during, and shortly following the occurrence of Hurricane Sandy; (vii) METOO, which contains information from 200 persons who have participated in the social movement against sexual assault and harassment.

Further exploring the domain of social media, the Moral Foundations Reddit Corpus (MFRC) (Trager et al., 2022) introduces a dataset of 16,123 English Reddit posts drawn from 12 different subreddits, chosen according to the buckets of US politics, French politics and everyday moral life. Unlike Hoover et al. (2020a), the MFRC does not deal with the Fairness/Cheating dimension of the MFT, but addresses the dimensions of Equality/Inequality and Proportionality/Disproportionality, following the framework of Atari et al. (2023). Equality/Inequality is a concept that promotes social justice and equality, while Proportionality/Disproportionality focuses on remuneration based on merit, contributing to virtues like productivity and deservingness, but also fostering corruption and nepotism. It also includes Thin morality and Non-moral categories, i.e., judgments which do not clearly refer to the six moral dimension, and judgments without moral content, respectively.

The Moral Foundations News Corpus (MNFC) (Weber et al., 2021) comprises news articles collected from major outlets (e.g., The New York Times) between 2013 and 2015 and analyzed with the MFD to detect words related to specific moral foundations. Several annotators used an online platform (Moral Narrative Analyzer),

to highlight portions of the text that reflected moral content. They also identified the most prominent moral foundation. Additionally, they annotated specific entities mentioned in the text (e.g., people or organizations) with moral foundations.

Beiró et al. (2023) gathered English Facebook comments focused on the topic of COVID-19 vaccination. They manually annotated 4,498 comments utilizing the same dimensions used by Hoover et al. (2020b). In a similar fashion, Kennedy et al. (2021) also collected data from Facebook volunteers for analyzing individual moral concerns. Specifically, status updates of English-speaking participants were collected, with a total of 107,798 status updates collected from 2,691 individuals, along with their responses on the 30-item MFQ, consisting of two 15-item sections: the first measures the significance that individuals attribute to each of the five moral foundations, while the second contains contextualized elements designed to assess real moral judgments.

Johnson and Goldwasser (2018) collected tweets of US politicians related to 6 topics: (i) abortion, (ii) the Affordable Care Act, (iii) guns, (iv) immigration, (v) LGBTQ rights, and (vi) terrorism. They annotated 2,050 tweets based on topics, policy frames, and the five moral foundational dimensions of the MFT, plus a non-moral category, enabling to frame the problem as an 11-label classification task. Roy and Goldwasser (2021) further extended this dataset by collecting more tweets from US congress members, using a lexicon-based approach.

Pacheco et al. (2022) built a corpus of 750 tweets about the COVID-19 vaccine manually annotated by morality foundation, and vaccination stance (pro, anti, neutral, disagree). Similarly, Rojecki et al. (2021) annotated 2,648 tweets related to the COVID-19 pandemic.

Kobbe et al. (2020) augmented the ArgQuality Corpus (Wachsmuth et al., 2017), which includes debate arguments for two stances on 16 topics like evolution vs. creation and banning plastic bottles. The dataset contains 320 arguments balanced by stance. Kobbe et al. (2020) annotated it with moral dimensions, enabling analysis of the relationship between moral values, argument quality, stance, and audience reactions. Similarly, Alshomary et al. (2022) used Hulpus et al. (2020)’s lexicon linking moral foundations to Wikipedia concepts, creating a dataset of 230k argumentative texts balanced across five moral dimensions and annotated with moral values. To assess the quality of the proposed dataset, a sample of 100 example have been manually validated.

The Moral Integrity Corpus (MIC) (Ziems et al., 2022) was built up from Social Chemistry 101 Forbes et al. (2020), which is a corpus of norms describing binary judgments of everyday situations. MIC extends Social Chemistry 101 by providing moral annotations related to the MFT on 38,000 prompt-reply pairs.

The Moral and Affective Film Set (MAAFS) (McCurrie et al., 2018) comprises 69 crowd-sourced YouTube video clips representing MFT dimensions. Independent participants (575 total) rated these videos using standardized questionnaires assessing MFT moral foundations, arousal (emotional intensity), wrongness (moral judgment of actions), weirdness (perceived strangeness), and clarity (understanding of the moral content and context), among other factors. Two of the authors reviewed and selected videos based on alignment with moral vignettes or foundations.

The Moral Foundations Vignettes (MFVs) (Clifford et al., 2015) introduces a collection of moral vignettes, each violating one of the five moral foundations. It contains 132 vignettes, crafted to encompass diverse content and validated through factor analysis and human participants. The dataset was annotated by 510 respondents and evaluated by approximately 30 individuals to ensure consistency and reliability of the responses.

PART I: AI-BASED NLP METHODS AND APPROACHES FOR THE LEGAL DOMAIN

Artificial intelligence is increasingly finding its application to the legal domain. This is mainly motivated by an evidence that the volume of information produced in the legal domain is overwhelming, and also variegated due to the involvement of many different actors, such as legal professionals (lawyers, attorneys), law courts, legislators, law firms, and even citizens Surden (2019). AI tools, especially those for NLP, are indeed conceived to help legal actors handling huge amounts of legal documents, lighten the workload in searching for and understanding text passages of interest, thus ultimately easing streamline processes and saving time for tackling more complex tasks.

AI is powering legal tech companies with the advanced technical backbone for better serving their clients more cheaply, efficiently and accurately than ever. Prominent examples of required applications of AI in the legal tech industry include contract review and analysis,⁹² and prediction of case outcomes or court rulings.⁹³

Moreover, AI development is highly fostered by political institutions and governments. For example, the European Union's approach to AI focuses on excellence and trust, aiming to boost research and industrial capacity while ensuring safety and fundamental rights.⁹⁴ This has been translated into concrete actions such as a review of the Coordinated Plan on AI (with EU member states) and a proposal for a regulation laying down harmonized rules on AI, known as *AI Act*.

Whilst the hype for successfully using TLMs is expected to hold true also for such a challenging domain as law, however, like for other specialized fields, legal language has distinct characteristics compared to general natural languages. Given the variety of legal sources and their heterogeneous functionality, legal language has indeed specific terminology and linguistic patterns, formal syntax, and semantics relying on a particular knowledge domain, to the extent that it is often regarded as a *sublanguage*.

⁹² <https://www.klarity.ai/>

⁹³ <https://lexmachina.com/>, <https://www.bluej.com/> <https://www.deepjudge.ai/>

⁹⁴ <https://digital-strategy.ec.europa.eu/en/policies/european-approach-artificial-intelligence>

Chapter 3

Bringing order into the realm of Transformer-based Language Models for Artificial Intelligence and Law

We examine the landscape of works that aim to solve legal problems by means of a subset of artificial intelligence and NLP methodologies, which corresponds to the state-of-the-art in legal AI and is represented by TLMs. We believe this is not trivial since, despite being a relatively recent technology, its evolution in the last few years has been dramatic and progressed at an unprecedented rate, also in the legal domain. To the best of our knowledge, this is the first systematic study on problems, tasks and benchmarks in the legal AI area, for which AI approaches and methods based on TLMs have been developed to address several tasks, such as retrieval, classification, prediction, entailment, information extraction and many others.

Despite our effort, this work cannot be intended to discuss definitive solutions to legal problems, tasks and applications. Moreover, please note that other deep learning technologies, from convolutional and recurrent networks to word embeddings, as well as topic modeling and classic machine-learning techniques are regarded as out of scope of this work, unless they are used in combination with TLMs.

This work is also necessarily a snapshot in time, until mid 2022, with an update to early 2023. It is hence likely that recently appeared relevant works were missing at the time of writing of this work. As such, the presented classification will need to be updated as new challenges come into focus and additional perspectives are brought to bear on legal problems.

3.1 Related Works

The use of AI within law has long been a topic of interest in the law area, mainly focusing on understanding the relations of AI to the practice and administration of law as well as on the ethical limits of the application of AI to legal data (e.g., (Callister, 2020; Surden, 2019)). While a purely law-based perspective is out of scope of this survey, it is nonetheless important to place our work into the computer-science literature that has recently surveyed the topic.

Zhong et al. (2020) overview existing tasks and applications in legal AI, organized into three categories, namely judgment prediction, similar case matching, and legal question answering. In this regard, a summary of main characteristics is provided about existing approaches, also including a mention to early TLMs used for legal AI. An overview on legal information retrieval approaches, divided into natural language based, ontology-based, and deep-learning-based systems is also presented in Sansone and Sperl  (2022). Francesconi (2022) describe the vision of the IAAIL president at ICAIL 2021 on the status of the AI and law discipline, including possible future perspectives that envisage the use of machine and deep learning to extract knowledge from legal data, combined with legal knowledge representation and models for legal reasoning. Also, Song et al. (2022) provide a comparative empirical evaluation of TLMs on various NLP tasks, which aims to gain insights into the performance difference between domain-specific models and general domain models.

While the above surveys are general in the domain of AI and law, a focus on the task of querying for ad-hoc case law retrieval is taken in Locke and Zuccon (2022), where approaches and methods are discussed into the following categories: Boolean and natural language, conceptual search and case-based retrieval, question answering, query expansion, query reduction, search result diversification, use of citation networks, and deeper understanding of texts.

None of the above works, however, provides a systematic analysis of approaches and methods based on TLMs for legal problems and tasks, which is instead the objective of our research.

3.2 Methods for Legal Understanding Tasks

In this section we discuss the main methods based on the TLMs and their application to the problems and tasks presented in Section 2.3.2. More specifically, we organize our presentation of the methods into six parts. Starting from early methods to more complex architectural settings, the first part (Section 3.2.1) includes methods that are mainly based on fine-tuning BERT and the other TLMs. Given the importance of the COLIEE competition over the past few years, we dedicate a separate part (Section 3.2.2) to the methods that were designed and competed for the COLIEE tasks. Section 3.2.3 discusses domain-adaptive approaches, which require pre-training of TLMs on legal corpora. Section 3.2.4 focuses on approaches coping with the text limitation length of BERT and other TLMs. Section 3.3 discusses GPT-based methods. Sections 3.2.5 and 3.2.6 are devoted to non-English and multilingual approaches, respectively. Finally, Section 3.2.7 discusses explainability and interpretability in TLM-based methods.

As a guide to our discussion, Tables 3.1-3.4 report on main characteristics of the approaches that we shall describe through this section, focusing on the adopted TLMs, the downstream tasks for which the approaches were designed, the types of data and languages, and whether a method is conceived to deal with long documents.

It should be noted that our objective is not providing a fully detailed description of each work, and most details about techniques that are not pertinent to TLMs might be discarded; also, we will not delve into the experimental results. Clearly, we refer the interested reader to the original works for further information.

3.2.1 Task-adaptive methods

Early applications of BERT and related models to the legal domain concern approaches that make TLMs *adaptive* to a legal specific task, i.e., they directly fine-tune an original, general-domain pre-trained model to the task at hand. Accordingly to the date of release of BERT, such applications have started being developed since 2019. For instance, Howe et al. (2019) provide a comparative evaluation of various text classification approaches, including BERT, topic models and word embeddings, for classifying Singapore Supreme Court judgments into legal areas. Gain et al. (2019) fine-tune BERT for a sentence-pair classification task to address the COLIEE-2019 Task 4 context. Rabelo et al. (2019a) fine-tune BERT on the training data available for the COLIEE-2019 Task 2, and use the model in the post-processing stage of a framework that exploits a combination of a multi-word token similarity score and a noun-phrase similarity score to identify candidate entailing paragraphs. This combined approach was ranked first for the COLIEE-2019 Task 2, with an F-score of 0.70 on the official COLIEE test dataset. In the same context of COLIEE-2019 Task 2, Yamada and Tokunaga (2019) fine-tune BERT for a sentence-pair classification task, where each entailed fragment and target paragraphs are regarded as sentences. Three variants are defined to handle different lengths of an input sentence-pair, and an ensemble model BERT-vote is implemented using a simple voting method based on a majority prediction for each instance according to the three BERT-based models.

In the context of the AILA-2020 task of rhetorical role labeling, Gao et al. (2020) combine TF-IDF features with the fine-tuned BERT embeddings (i.e., the feature vectors of the [CLS] token) to feed logistic regression, SVM, and AdaBoost classification models for final training and prediction. Results have shown that the joint semantic-statistical feature modeling improves upon the individual models.

Elwany et al. (2019) fine-tune BERT on a proprietary corpus of legal contracts for the task of extracting the fixed-term and the auto-renewing-term portions of an agreement. Two main strategies are developed to build the model depending on whether the BERT layers are frozen or left unfrozen during the fine-tuning process.

Table 3.1: Summary of main methods discussed in Sections 3.2.1 and 3.2.2.

Method	Ref.	Downstream Tasks	Lang.	Data	Long docs?
Fine-tuned BERT	Howe et al. (2019)	TC	EN	legal cases	No
Fine-tuned BERT	Rabelo et al. (2019a) Yamada and Tokunaga (2019)	COLIEE-2019 Task 2	EN	legal cases	No
Fine-tuned BERT	Gain et al. (2019)	COLIEE-2019 Task 4	EN	civil code	No
Fine-tuned BERT w/ TF-IDF	Gao et al. (2020)	RRL	EN	legal cases	No
Fine-tuned BERT w/ and w/o frozen layers	Elwany et al. (2019)	NER	EN	contracts	No
Fine-tuned BERT in regression form	Sanchez et al. (2020)	PR	EN	legal news articles	No
Fine-tuned BERT	Ravichander et al. (2019)	QA	EN	contracts	No
Fine-tuned BERT	Ahmad et al. (2020)	QA	EN	contracts	No
BiLSTM-CRF + BERT (PI-Extract)	Bui et al. (2021)	NER	EN	contracts	No
Fine-tuned BERT	Mistica et al. (2021)	TC	EN	help requests	No
Fine-tuned RoBERTa	Westermann et al. (2020)	COLIEE-2020 Task 4	EN	civil code	No
Fine-tuned RoBERTa + BiLSTM	Majumder and Das (2020)	RRL	EN	legal cases	No
Fine-tuned RoBERTa	Vold and Conrad (2021)	QA	EN	PRIVACYQA	No
Fine-tuned BERT and RoBERTa	Chalkidis et al. (2020a)	TC	EN	EURLEX57K; MIMICIII; AMAZON13K	No
Fine-tuned RoBERTa	Savelka et al. (2020) Savelka and Ashley (2021)	TC	EN	legal cases	No
Fine-tuned T5	Hudzina et al. (2020)	COLIEE-2020 Task 4	EN	civil code	No
Fine-tuned BERT and SpanBERT	Kruiper et al. (2021)	MWE discovery	EN	SPAR.txt	No
Fine-tuned RoBERTa + BigBird	Hong et al. (2021)	IE	EN	parole hearings	No

Continued on the next page

Method	Ref.	Downstream Tasks	Lang.	Data	Long docs?
Fine-tuned XLM-RoBERTa w/ GLoVe embeddings	Trias et al. (2021)	NER	EN	CAP	No
Fine-tuned BERT, RoBERTa, DistilBERT, XLNet and m-BERT	Shaheen et al. (2020)	LMTC	EN	JRC-Aquis and EURLEX57K	No
Fine-tuned BERT	Chalkidis et al. (2019b)	LMTC	EN	EURLEX57K	No
Triplet-loss-network w/ fine-tuned S-BERT encoder	Sarkar et al. (2021)	TC	EN	promissory and non- promissory sentences	No
Fine-tuned BERT, RoBERTa, Transformer + BiLSTM + CRF; Fine-tuned MiniLM, UniLM, UniLMv2, MASS and BART	Ahmad et al. (2021)	intent classification, SF	EN	contracts	No

BERT; fine-tuned BERT; BERT in regression form; ensemble of BERT and fine-tuned BERT; BERTLaw	Nguyen et al. (2020)	COLIEE-2020 Tasks 1-4	EN	legal cases; civil code	No
Ensemble of fine-tuned BERT and BM25	Nguyen et al. (2021a)	COLIEE-2021 Tasks 1, 2	JP	legal cases	No
Fine-tuned BERT, BART, and RoBERTa + Random Forest	Rabelo et al. (2020)	COLIEE-2020 Task 2	EN	legal cases	No
Fine-tuned BERT	Alberts et al. (2020)	COLIEE-2020 Task 2	EN	legal cases; NLI	No
monoT5-zeroshot; fine-tuned monoT5 and DeBERTa; DeBERTa + monoT5 Ensemble	Rosa et al. (2021)	COLIEE-2021 Task 2	EN	case paragraphs, decision fragments	No
Fine-tuned BERT w/ truncated texts	Shao et al. (2020b)	COLIEE-2020 Task 2	EN	legal cases	No
Fine-tuned BERT and AIBERT	Shao et al. (2020a)	COLIEE-2020 Tasks 3, 4	JP	civil code	No
Fine-tuned RoBERTa w/ NLI type detection	Rabelo et al. (2020)	COLIEE-2020 Task 4	EN	civil code; SNLI	No
Fine-tuned BERT	Schilder et al. (2021)	COLIEE-2021 Task 3	JP	civil code	No
Fine-tuned S-BERT + TF-IDF vectorization + data enrichment; Legal-BERT-FP@aueb + TF-IDF vectorization + data augmentation; BERTScore	Wehnert et al. (2021)	COLIEE-2021 Task 3	EN	civil code	No
Fine-tuned BERT + XGBoost	Alberts et al. (2020)	COLIEE-2020 Task 4	EN	civil code	No
Ensembles of fine-tuned ELECTRA, T5 and Multee	Schilder et al. (2021)	COLIEE-2021 Task 4	JP	civil code	No
DistilRoBERTa and GPT-3	Schilder et al. (2021)	COLIEE-2021 Task 5	JP	civil code	No

Continued on the next page

Method	Ref.	Downstream Tasks	Lang.	Data	Long docs?
Fine-tuned BERT ensemble w/ augmentation	Yoshioka et al. (2021) Yoshioka et al. (2021)	COLIEE-2021 Tasks 3-5	JP	civil code	No
Fine-tuned BERT and Legal-BERT@aueb; Fine-tuned BERT w/ thesaurus concept number	Kim et al. (2021)	COLIEE-2021 Tasks 2, 4, 5	EN	legal cases; civil code	Yes

Marginal yet practically valuable improvements in both accuracy and training speed were observed for the fine-tuned BERT model.

Sanchez et al. (2020) employ BERT in its regression form for passage retrieval, i.e., to predict a relevance score for a query-document pair in input, where each document is represented by a combined title and summary field along with the content of a news article. On a similar fashion, Mistica et al. (2021) fine-tune BERT on a dataset of online help requests for a multi-label classification task pertinent to different areas of law.

Ravichander et al. (2019) fine-tune BERT on the PrivacyQA benchmark for two tasks: to identify if a question is answerable and to identify the evidence in a privacy policy that answers the question. to predict only the answerability of the question, then at inference time, if the question is considered answerable, the model produces the evidences. Analogously, Ahmad et al. (2020) fine-tune BERT on the PolicyQA benchmark, whose task is to predict the answer span of a related question, given the policy segment. This time, however, the model is first pre-trained on a dataset, SQuAD, which is regarded as related to the PolicyQA corpus. The positive impact of pre-training with SQuAD and fine-tuning on the benchmark is demonstrated by the fact that the same model without pre-training or fine-tuning (or both) performed lower. A different use of BERT in the policy context is adopted by Bui et al. (2021), which propose *PI-Extract* for the extraction of personal data and related actions in privacy policies. *PI-Extract* consists of four BiLSTM-CRF-based NER models trained on a privacy policy corpus introduced by the authors (cf. Section 2.4.2) which contains four possible labels to be assigned to the personal data objects contained in an input sentence, where each label expresses the type of privacy action performed on the personal data. Each NER model of *Pi-Extract* is specifically trained to predict only one of four possible labels. In *PI-Extract*, BERT is used as a more effective alternative to word embedding based on GloVe Pennington et al. (2014) to encode the text input.

A few studies have also focused on RoBERTa. Westermann et al. (2020) fine-tune RoBERTa for sentence-pair classification on the COLIEE-2020 Task 4. Majumder and Das (2020) provides RoBERTa embeddings to a BiLSTM for the AILA-2020 task of rhetorical role labeling, reaching the highest rank at the competition. Vold and Conrad (2021) fine-tune RoBERTa on the PRIVACYQA dataset, a corpus of questions about privacy policies associated with mobile applications, for binary or graded QA pair classification. Chalkidis et al. (2020a) evaluate BERT and RoBERTa, along with RNN-based Label-Wise Attention Networks (LWANs) and Probabilistic Label Trees (PLTs) on three English datasets for large-scale multi-label text classification: EURLEX57K (Chalkidis et al., 2019b), MIMICIII, and AMAZON13K. Experimental results have shown that BERT and RoBERTa outperform both PLT-based methods and LWANs, and that a combination between BERT and LWAN can further improve performance.

On the task of determining whether a sentence describes facts or not, Savelka et al. (2020) develop a meta type system to train models on data from different domains and jurisdictions of annotated legal cases and assess performance of the models across the various domains. Datasets have been re-labeled in such a way as

to obtain from the original rhetorical roles only “Fact” and “Non-Fact” labels. To this purpose, RoBERTa base is fine-tuned on the different datasets and combined datasets.

Savelka and Ashley (2021) focus on the identification of explanatory sentences, i.e., to detect sentences that are useful for explaining selected legal concepts. In order to classify sentences according to four levels of usefulness, three RoBERTa-base models are fine-tuned on different tasks: classifying retrieved sentences in terms of their value for explaining the legal concepts, sentence pair classification, and another type of sentence pair classification where one sentence is a provision of written law and the second is a retrieved sentence. Results point out important interactions among a legal concept, the provision of law in which it is embedded, and retrieved sentences.

Besides the RoBERTa alternative to BERT, Hudzina et al. (2020) resort to the pre-trained T5-base model for the COLIEE-2020 Task 4 and for denoising tasks on Japanese Civil Code article texts and titles and on Wikibook articles about the Japanese Civil Code. Results have shown that T5 appears to overfit the training data despite the multiple domain-specific tasks. Kruiper et al. (2021) utilize pre-trained BERT-base and SpanBERT-base¹ to learn a sequence tagger for a particular task of multi-word expression (MWE) for Automated Compliance Checking (ACC), which is to identify low-level constituent parts of a sentence, i.e., spans. The objective is to learn a semantic lexicon for ACC, which is essential for semantic parsing as it enables the grounding of information units identified in a text (e.g., objects, interactions, and constraints). SpanBERT shows to be less effective than BERT, which is likely due to a mismatch between the span types and sizes used for the training of the original SpanBERT and the ones used to train the sequence tagger.

Hong et al. (2021) investigate on information extraction of case factors from a corpus of parole hearing transcripts, which is characterized by longer texts with fewer labels than in other NLP datasets. A two-step open-domain question answering approach is followed by using a Reducer to extract relevant text segments from a hearing and a Producer model to generate answers from the text segments selected by the Reducer. A combination of a rule-based Reducer and a neural Producer show to yield improved performance. In particular, as default Producer, a combination of RoBERTa and BigBird base is chosen due to its balance of long input length, low computation requirements, and performance on different choices of prediction heads, namely classification head, MLM head and QA head.

Trias et al. (2021) propose an ensemble language model consisting of a combination of a Transformer architecture with a finite state machine to extract names from American English documents.² This exploits one pre-trained general-purpose English language NER model based on Flair³ trained on CoNLL03 data, and another one trained on Harvard Caselaw Access Project, using GLoVe embeddings and XLM-RoBERTa.

¹ <https://github.com/rubenkruiper/SPaR.txt>

² <https://harvard-almit.github.io/legal-nlp>

³ <https://github.com/flairNLP/flair>

Shaheen et al. (2020) focus on the problem of large multi-label text classification (LMTC) in the legal domain, using JRC-Aquis and EURLEX57K annotated with the EuroVoc labels. In LMTC, the label space is comprised of thousands of labels, typically following a power-law label distribution and hierarchically organized. Pre-trained BERT, RoBERTa, DistilBERT, XLNet and m-BERT are applied on this task, using various training strategies such as gradual unfreezing⁴ and slanted triangular learning rates,⁵ in addition to fine-tuning. Results indicate that DistilBERT is better in retrieving the most probable label compared with RoBERTa and BERT. Also in the LMTC context, Chalkidis et al. (2019b) apply a number of methods, including BERT, to investigate which portions of the documents in EURLEX57K are more informative. Results show that the title and recitals of each document contains enough information (and also allows to overcome the BERT’s maximum text length limit); however, the approach fails in zero-shot learning.

Sarkar et al. (2021) define a triplet network for legal sentence classification. A triplet network consists of three instances of the same neural network with shared parameters, it takes as input three objects, i.e., the positive example and the anchor, which belong to the same class, and the negative examples, which belongs to a different class, and it outputs the distance between the anchor and the positive example and the distance between the anchor and the negative example. The representation of the anchor, positive and negative sentences are then used to compute the triplet loss function to minimize. The network encodes each input sentence using S-BERT as encoder, and the contextual sentence embedding is then fed to a two-layer perceptron. For the downstream classification task, a SVM with Radial Basis Function (RBF) kernel is used to compute the probability that an input sentence is promissory; the sentence is eventually classified as either promissory or non-promissory depending on a user-specified threshold on the SVM output probability.

Ahmad et al. (2021) deal with fine-grained information extraction of privacy policies, specifically to identify sentences expressing privacy practices (intent classification task) and to detect specific details as text spans into the sentences (slot filling task). Two alternative approaches are proposed to tackle the two tasks: modeling intent classification and slot tagging either jointly as a sequence tagging task, or separately by generating respectively labels and spans. For the first approach, BERT, RoBERTa, a Transformer model along with a BiLSTM encoder are trained to get contextual representations of the input; in particular, the embedding of the special token [CLS] is used by a softmax classifier to predict the target intent, while the embeddings of the input tokens are used by another softmax classifier to predict the slot labels combined with a conditional random field (CRF). Results show that BERT-based and RoBERTa-based sequence tagging models outperform the other baselines by a wide margin, and the use of CRF is beneficial for the slot tagging task but slightly degrades the performance for intent classification. For the second

⁴ The training process is divided into multiple cycles, each consisting of several training epochs. Training starts after freezing all layers except for the last few layers in the first cycle one, then more layers are unfrozen gradually (from last to first layers) during later cycles.

⁵ The learning rate is increased at the beginning of a training epoch up to the maximal learning rate, then slowly reduced to refine the parameters.

approach, MiniLM, UniLM, UniLMv2, MASS and BART are fine-tuned on the benchmark, and results show that seq2seq models outperform the sequence tagging models in the slot filling task, with BART and UniLM-based models reaching best performances, while they have similar results on the intent classification task.

3.2.2 Enhanced methods for COLIEE tasks

Case Law Retrieval (Task 1). In the COLIEE-2021 competition, Nguyen et al. (2021a) consider BERT as semantic model and Rank-BM25⁶ as lexical model. In particular, the lexical model restricts the searching space of candidate cases to the top-100 cases for a given query, according to a lexical similarity score. After that, query and candidates are split into paragraphs and a lexical score matrix for each query-candidate pair is derived. In a similar way, a semantic score matrix for each query-candidate pair is derived using BERT, which is fine-tuned on a silver dataset based on the Task 1 raw data. Finally, the most relevant cases for the given query is obtained by combining lexical and BERT-based scores. The overall performance on the test set turns out to be quite poor, probably due to prior passage of searching space restriction performed by the lexical model.

In the previous edition of COLIEE competition, the same team uses a similar approach, where BERT is fine-tuned for a text-pair classification task with the support of specific heuristics to extract text-pairs, and BM25 is used as lexical model Nguyen et al. (2020). In that case, for each query the searching space consists of only 200 possible candidates, thus leading to significantly better results.

Rossi and Kanoulas (2019), Li et al. (2021b), Althammer et al. (2021), Ma et al. (2021) and Shao et al. (2020b) propose methods dealing with the document length limitation; therefore a discussion of such works is postponed to Section 3.2.4.

Case Law Entailment (Task 2). One of the first approaches to case law entailment is provided by Rabelo et al. (2019a), which has been described in Section 3.2.1. In the 2020 edition, Rabelo et al. (2020) propose a Random Forest classifier to score each pair query-paragraph using features from the entailment score of BERT fine-tuned on the Task 2 data, BART and RoBERTa fine-tuned on a generic entailment dataset, BERT fine-tuned on paraphrase detection (zero-shot setting), cosine similarity between sentence embeddings of the input pair and cosine similarity on bag-of-words of the noun phrases contained in the input pair. Data augmentation based on back translation is also carried out, however without significant evidence of improvement.

Alberts et al. (2020) exploit BERT in two ways to address the task in the COLIEE-2020 competition: the first one is to compute cosine similarity scores between BERT embeddings after fine-tuning on the COLIEE training data, and the second way is to derive features for a natural language inference (NLI) task by applying BERT on the NLI dataset. The two types of features, combined with BM25 based ones, are fed

⁶ <https://pypi.org/project/rank-bm25/>

as input to a XGBoost-based classifier (Chen et al., 2015), which achieved the third place in the competition.

Nguyen et al. (2020) experiment with the same solution proposed for Task 1 (described in Section 3.2.2) together with BM25 scores. This is also improved by fine-tuning BERT on the training data provided for this task. A further variant is proposed in which BM25 is replaced with BERT fine-tuned on SigmaLaw,⁷ a corpus of legal cases including cases from the US Supreme Court. Among the three versions, the second one achieved the best F1 score reaching the leading position in the competition.

In the COLIEE-2021 competition, Nguyen et al. (2021a) starts from their own approach implemented for Task 1 to define a binary classification task for BERT fine-tuned using both the silver dataset, obtained from the Task 1 data, and a further gold dataset obtained from the Task 2 data. One of their submissions is a combination of lexical score and a model trained on the *Next Foreign Sentence Prediction (NFSP)* task (Nguyen et al., 2021b), that is a binary classification task in which the text-pairs are composed of sentences with different languages and the model has to understand the correct meaning of each sentence to determine if their semantics belong to two consecutive sentences (cf. Section 3.2.6). However, best results are obtained using their earlier approach.

Rosa et al. (2021) apply pre-trained Transformer models directly to the Task 2 of COLIEE 2021, without any preliminary fine-tuning on the legal task.⁸ More specifically, they test monoT5 in two settings: zero-shot and fine-tuned on the Task 2 data of COLIEE 2020. In addition to monoT5, they also fine-tune DeBERTa on the legal task using the Task 2 data of COLIEE 2021. To balance the amount of negative and positive examples, they expand the positive examples generating a set of artificial fragments from the base case paragraphs through a sliding window strategy. As a further solution, an ensemble of monoT5 and DeBERTa, both fine-tuned on COLIEE 2020 data, is proposed. The models estimate a score for each fragment-candidate paragraph pair to select the best set of candidate paragraphs for each case fragment. A candidate paragraph is chosen if its score complies with specific rules. For the ensemble model, the selected paragraphs are the concatenation of the final set of DeBERTa and monoT5. The fine-tuned version of DeBERTa, monoT5 and the ensemble model achieve the first 3 positions in the competition, with the ensemble model on top position. However, once the ground-truth annotations of the test set have been released, the authors find out that monoT5 with zero-shot setting performs better than the single fine-tuned monoT5 and DeBERTa. This result leads the authors to the conclusion that, in case of limited labeled data, a model with no adaptations to the target domain can be more robust than fine-tuned models.

In the attempt of dealing with the text length limitation of BERT, Shao et al. (2020b) consider either to truncate the input queries and paragraphs symmetrically, or to limit the tokens of the query to 128 and truncate the paragraph if the total tokens of query and candidate exceeds the limit. This is motivated by the observation that

⁷ <https://osf.io/qvg8s/>

⁸ <https://github.com/neuralmind-ai/coliee>

most of the decision fragments in the training data is no longer than 128 tokens. In both cases, the final hidden vector corresponding to the [CLS] token is fed to a fully connected layer for the classification. The results on test set demonstrate that the asymmetric truncation significantly improves the performance of the model. A third solution for the task proposed by the authors is to combine the output vectors of BERT (both with symmetric and asymmetric truncation), BM25 score and some structural features to be fed to a RankSVM model.⁹

Note that there have been other approaches, such as those proposed in Li et al. (2021b); Kim et al. (2021), effectively dealing with the length of the documents; as such, they will be discussed in Section 3.2.4.

Statute Law Retrieval (Task 3). In COLIEE 2020, Nguyen et al. (2020) fine-tune `bert-base-uncased` on a text-pair classification task considering all query-article pairs. Due to the unbalancing between positive and negative examples, they filter all the possible candidates to top- k articles according to the cosine similarity calculated comparing the TF-IDF vectorization of article and query. They combine a `bert-base-uncased` model fine-tuned on text-pair classification and a variant of BERT, dubbed *BERT-CC*, previously fine-tuned on all data of the Civil Code with a masked language modeling task and further fine-tuned on text-pair classification. This is motivated by the assumption that BERT-CC can compensate for the lack of domain-specific knowledge of BERT, so that together they can learn different linguistic features. This hypothesis is confirmed by the results on the test set in the competition, where the ensemble model performed higher than the two separate models.

Shao et al. (2020a) address the task as a binary classification problem using the Japanese version of the data provided by organizers. Each question is paired with every possible article and, due to the imbalance between positive and negative samples, the positive samples are oversampled up to 100 times. A Japanese version of BERT with whole word masking mechanism, called `BERT-base-mecab-ipadic-char-4K-whole-word-mask` (for short, *BERTjpcwwm*), and a Japanese version of ALBERT¹⁰ are trained. In the first case, *BERTjpcwwm* is equipped with two settings, considering a maximum text length of 384 and 512, respectively. The relevance of an article for a query is decided through a threshold. The ensemble of these two *BERTjpcwwm* models and ALBERT model ranked first at COLIEE-2020.

For the 2021 edition, Schilder et al. (2021) fine-tune a pre-trained Japanese BERT model on pairwise sequence classification task, where a text-pair is composed of queries and articles. The task is addressed by varying the BERT score along with the maximum cosine similarity on validation set and using cosine similarity to sort entailing articles. However, the proposed system produced unsatisfactory results, probably due to the lack of a sufficient number of training examples to adequately fine-tune BERT on COLIEE language domain.

⁹ https://www.cs.cornell.edu/people/tj/svm_light/svm_rank.html

¹⁰ <https://github.com/alinear-corp/albert-japanese>

Wehnert et al. (2021) propose three approaches based on BERT model for the English version of the task. In the first approach, S-BERT embeddings are combined with TF-IDF vectorization. To this end, the training data is increased with document enrichment strategies, in order to obtain article vectors that are as much unique as possible. In particular, each article can be enriched with hierarchical relations between articles as metadata, crawling the Japanese open source commentary on the article¹¹ and using the training data of Task 4. For each query-article pair, the cosine similarity scores are calculated over TF-IDF and S-BERT embeddings (`paraphrase-distilroberta-base-v1`). In the second approach, `LEGAL-BERT@aub` is fine-tuned on sentence classification task. Data manipulation is performed in order to have an augmented dataset. The softmax score derived from the model is then combined to the cosine similarity calculated on TF-IDF vectorization (cf. Section 3.2.3). The third approach involves the use of BERTScore, which works as follows: the pairwise cosine similarities of all token-wise contextual embeddings from two sentences are first computed, then the token pairs between the two sentences which have the highest cosine similarity are selected, and those similarities are summed up and discounted by the words in the sentence to obtain precision, recall and the according F1-score. The first approach proved to be the best for the competition, which was ranked first at the 2021 competition.

Note that Aydemir et al. (2020) address the COLIEE-2020 task based on a multi-language approach, which is described in Section 3.2.6. Also, in Section 3.2.4, we discuss the COLIEE-2021 methods by Nguyen et al. (2021a) and Yoshioka et al. (2021) dealing with the length limitation of BERT.

Statute Law Entailment (Task 4). For the statute law entailment task, early Transformer-based solutions have been proposed in Gain et al. (2019), Westermann et al. (2020), and Hudzina et al. (2020) (cf. Section 3.2.1).

In COLIEE 2020, Alberts et al. (2020) use the last hidden layers `bert-base-cased` as the input for the XGBoost classifier, inspired by Gain et al. (2019). However, such a combination reaches poor performance, probably due to the fact that BERT was not trained on legal tasks. Shao et al. (2020a) apply the same ensemble system used for Task 3, without however replicating the enhancement of the training samples, as for Task 4 the goal is to answer the queries based on the provided relevant articles. This led to lower performance of the model, which ranked sixth in the competition.

Rabelo et al. (2020) address the statute law entailment problem using RoBERTa for natural language inference. They identify multiple NLI types in the statute law dataset, from which the model needs to recognize condition, conclusion and exceptional cases. To add more training data, they include the SNLI dataset (Bowman et al., 2015). A pre-processing step is performed to help the detection of the different NLI types. This step consists in splitting the data in condition, conclusion and exceptional case, replacing references to prior cases or paragraphs with the referenced text, re-generating a sentence by negating the conclusion for the exceptional case and extracting one relevant sentence that matches most terms with the query. The model reaches the second position in the competition. The authors also adopt

¹¹ <https://ja.wikibooks.org/>

a BERT model for natural language inference to address Task 4 of COLIEE 2021 (Kim et al., 2021). They select the most relevant candidate sentences for the query following their previous approach (Rabelo et al., 2020) and training a BERT classifier over triplets of query, article’ sentence and binary label (yes/no). To help the pragmatic reasoning of BERT, the authors incorporate additional features consisting of the semantic category numbers of the Kadokawa thesaurus to the content words. The inclusion of such a semantic information boosts the performance of the model, reaching the fourth position in the competition for Task 4, while the same model without the additional information reached the third-to-last position.

Nguyen et al. (2020) exploit the questions and the results of their system for the Task 3 (cf. Section 3.2.2) as input pairs for the classification problem. On the development set, top-2 articles extracted by TF-IDF from the Civil Code are added to the gold data for each question. The answer for the question is positive if at least one pair is classified as positive by `bert-base-uncased`. As a second setting, only the content of the questions is used as input, following the approach they used in COLIEE 2019 (Rabelo et al., 2019b). Essentially, the entailment problem is translated to a *lawfulness* classification task. BERT and BERTLaw (cf. Section 3.2.3) are trained with law sentences of the Civil Code and previous years’ bar questions available from COLIEE. As shown in the final scores on test set, the second setting with BERTLaw reaches a score considerably higher than the other solutions proposed by the same team as well as by other teams that contributed to the competition.

Schilder et al. (2021) define ensemble models based on ELECTRA and T5 to address the Japanese version of the statute entailment task. In the first case, a Japanese ELECTRA-*small* variant is chosen,¹² which turns out to be the best pre-trained architecture for Japanese according to the Jensen-Shannon divergence (Lin, 1991), and measures the embedding distribution distances between positive and negative examples. The model is trained layer by layer on the few training examples available. Although the model obtains strong performance in the validation step, it achieves a low accuracy on test set, probably due to a potential over-fitting on the available data. As regards T5, the best performing approach consists in fine-tuning on COLIEE dataset and evaluating T5 embeddings (the other two approaches consist in adding span corruption in the fine-tuning step and pre-training T5 embeddings from scratch, respectively). Besides ELECTRA and T5, the authors also investigate a multi-sentence natural language inference model, *Multee* (Trivedi et al., 2019). The final submissions comprise ELECTRA, Multee and an ensemble of the first T5-based approach, ELECTRA and Multee. However, none of the three submissions obtained satisfactory results.

Yoshioka et al. (2021) address the Japanese statute law entailment task using BERT-based ensemble methods with data augmentation. In particular, they augment training data with a focus on the logical mismatches between articles and questions. The positive examples are composed of the same sentence used both as the query and as the article’s sentence, while negative examples are pairs of original sentences and their versions with judicial decision flipped. If one article contains multiple

¹² <https://huggingface.co/Cinnamon/electra-small-japanese-discriminator>

decisions, then it is divided into smaller sentences containing one judicial decision, and if the split sentence expresses an exceptional case, then the judicial decision is flipped. Then, ten BERT-japanese models¹³ are fine-tuned on this dataset to evaluate if the article entails a question or not. The models differ in the non-deterministic characteristics of fine-tuning process and the randomly selection of training set. Three ensemble models are then obtained by combining subsets of the aforementioned ten models. In the competition, they reach the best three accuracy scores, thus taking the top three positions.

Further studies on Task 4 are Nguyen et al. (2021a) and Kim et al. (2021), which address the task by accounting for the document length limitation (cf. Section 3.2.4), Aydemir et al. (2020), which propose a multi-language approach (cf. Section 3.2.6), and Goebel et al. (2021), which address the task based on a domain-adaptive solution (cf. Section 3.2.3).

Statute Law Question Answering (Task 5). Schilder et al. (2021) propose two solutions for the Japanese version of Task 5: using DistilRoBERTa without any domain-specific training, and using GPT-3 with a few-shot learning. DistilRoBERTa is a distilled version of RoBERTa for paraphrase detection. The authors use the implementation¹⁴ of S-BERT containing a pre-trained DistilRoBERTa model trained on paraphrase text. The relevant articles to a given query are determined according to a similarity score. GPT-3 is a massive auto-regressive language model based on Transformer that obtained impressive results in generation and question-answering tasks with a few-shot learning scenario. The model is available via OpenAI interface¹⁵ in different size versions. The authors use the largest model, dubbed *davinci*, and the fastest model, dubbed *ada*. From the experimental results, it is observed that GPT-3 models could not reach a sufficient level of comprehension on the legal domain, while the system based on DistilRoBERTa shows promising performance on the task.

Kim et al. (2021) combine the output of a traditional TF-IDF technique with the NLI system described for the English version of Task 4 (cf. Section 3.2.2). As for Task 3, the additional semantic information enhances the model performance, allowing the team to achieve the second position at the competition.

Yoshioka et al. (2021) also consider the Japanese version of Task 5, although they could not meet the deadline for the submissions. Three configurations of ensemble models are defined by combining the submitted runs for the Task 3 (Yoshioka et al., 2021) (cf. Section 3.2.2) and for Task 4 (Yoshioka et al., 2021) (cf. Section 3.2.4), under the same conditions of the other participants. One of the three configurations, based on the best run of Task 4, obtained similar performance compared to the best model of the competition (Nguyen et al., 2021a). Also, Nguyen et al. (2021a) propose a multilingual approach for the task, therefore its discussion is referred to Section 3.2.6).

¹³ <https://github.com/cl-tohoku/bert-japanese>

¹⁴ <https://github.com/UKPLab/sentence-transformers>

¹⁵ <https://openai.com/api/>

Table 3.2: Summary of main methods discussed in Sections 3.2.3 and 3.2.4.

Method	Ref.	Downstream Tasks	Lang.	Data	Long docs?
Pre-trained BERT w/ legal or hybrid tokens; Pre-trained DistilBERT w/ legal or hybrid tokens	Shaghighian et al. (2020)	PR, NER, TM, SA	EN	US SEC legal agreements	No
From-scratch pre-trained BERT (BERTLaw)	Nguyen and Nguyen (2021)	COLIEE-2020 Task 4	EN	civil code	No
Further pre-trained BERT (Legal-BERT-FP@aub); From-scratch pre-trained BERT (Legal-BERT-SC@aub)	Chalkidis et al. (2020b)	TC, NER	EN	EURLEX57K; ECHR cases; CONTRACTS-NER	No
Further pre-trained BERT (Legal-BERT@jhu)	Holzenberger et al. (2020)	CTR	EN	CAP cases	No
Further pre-trained BERT (Legal-BERT-FP@stanford); From-scratch pre-trained BERT (Legal-BERT-SC@stanford)	Zheng et al. (2021)	QA, OR, ToS DS	EN	CaseHOLD; Overruling; Terms-of-Service	No
Fine-tuned Legal-BERT@aub	Mahari (2021)	LPP	EN	CAP judicial opinions	No
Further pre-trained AIBERT (ALeaseBERT)	Leivaditi et al. (2020)	red flag detection, entity extraction	EN	contracts	No
Further pre-trained BERT, SpanBERT and ELECTRA	Chi et al. (2023)	intent classification, SF, QA, NER	EN	contracts	No
Further pre-trained Legal-BERT@aub (InLegalBERT); Further pre-trained Legal-BERT@stanford (InCaseLawBERT)	Paul et al. (2022b)	SAR, RRL, CJP	EN	legal cases	Yes
Fine-tuned BERT, Legal-BERT@aub, Legal-BERT@jhu, RoBERTa, S-BERT, S-RoBERTa, Longformer	Ostendorff et al. (2021)	DR	EN	Open Case Book; Wikisource US Courts	Yes
Fine-tuned BERT and LegalBERT@aub + BiLSTM and CRF	Bhattacharya et al. (2021)	RRL	EN	legal cases	No
Fine-tuned BERT, S-BERT and LegalBERT@aub	Chalkidis et al. (2021b)	RIR	EN	EU2UK; UK2EU	No
Fine-tuned BERT (C-BERT)	Chalkidis et al. (2021b)	EUROVOC concept classification	EN	EU2UK; UK2EU	No
Fine-tuned BERT + KERMIT	Wehnert et al. (2022)	COLIEE-2021 Task 4	EN	civil code	No
Fine-tuned BERT, RoBERTa, LegalBERT@aub, DeBERTaV3, BigBird	Wang et al. (2023a)	MCRC	EN	contracts	Yes
Fine-tuned Legal-PEGASUS, Legal-LED, BART, BERT-BART; BERTSUM	Shukla et al. (2022)	ES, AS	EN	legal cases	Yes

Continued on the next page

Method	Ref.	Downstream Tasks	Lang.	Data	Long docs?
Fine-tuned BERT w/ summarized texts	Rossi and Kanoulas (2019)	COLIEE-2019 Task 1	EN	legal cases	Yes
Two-stage fine-tuned BERT w/ attentive aggregation (HIER-BERT)	Chalkidis et al. (2019a)	AV, CIR	EN	ECHR cases	Yes
Two-stage fine-tuned BERT w/ attentive aggregation (BERT-PLI)	Shao et al. (2020c)	COLIEE-2019 Tasks 1, 2	EN	legal cases	Yes
LMIR + fine-tuned BERT-PLI	Shao et al. (2020b), Ma et al. (2021)	COLIEE-2020 Task 1	EN	legal cases	Yes
Fine-tuned BERT, ALBERT, RoBERTa, DeBERTa	Hendrycks et al. (2021b)	QA	EN	CUAD legal contracts	Yes
Fine-tuned RoBERTa and S-RoBERTa	Aumiller et al. (2021)	STP	EN	ToS documents	Yes
Transformer summarizer + fine-tuned BERT	Kim et al. (2021)	COLIEE-2021 Task 2	EN	legal cases	Yes
Fine-tuned BERT and RoBERTa	Nguyen et al. (2021a)	COLIEE-2021 Task 3	JP	civil code	Yes
Hier-BERT variation based on LegalBERT@aueb	Chalkidis et al. (2021c)	rationale extraction	EN	ECtHR	Yes
Pre-trained DPR (LawDPR); Fine-tuned Longformer-Encoder-Decoder	Althammer et al. (2021)	COLIEE-2021 Tasks 1, 2	EN	Canada case laws	Yes
Fine-tuned BERT (BERTSum)	Deroy et al. (2021)	ES	EN	legal cases	Yes
Fine-tuned BERT, RoBERTa, DistilBERT	Klaus et al. (2022)	ES	EN	EUR-LexSum	Yes
Further pre-trained BERT and RoBERTa	Lam et al. (2020)	employment notice prediction	EN	notice cases	Yes
Fine-tuned BERT, Legal-BERT@aueb and Legal-BERT@stanford; Fine-tuned BigBird and LongFormer	Limsopatham (2021)	AV, OR	EN	ECHR Violation dataset; Overruling Task dataset	Yes
From-scratch pre-trained BERT + Siamese Legal BERT + Fine-tuned Legal BERT	Khazaeli et al. (2021)	PR, QA	EN	caselaw headnotes, RFCs	No
Further pre-trained BERT (BERT-Legal) + LSTM	Li et al. (2021b)	COLIEE-2021 Task 1	EN	legal cases	Yes
Fine-tuned BERT-Legal w/ augmentation; Fine-tuned BERT-Legal w/ fast gradient;	Li et al. (2021b)	COLIEE-2021 Task 2	EN	legal cases	Yes
Fine-tuned Legal-BERT@aueb + TF-IDF vectors	Furniturewala et al. (2021)	AILA-2021 Tasks 2a-b	EN	legal cases	Yes
Fine-tuned Legal-BERT@aueb	Jain et al. (2021)	AILA-2021 Tasks 2a-b	EN	legal cases	Yes

Continued on the next page

Method	Ref.	Downstream Tasks	Lang.	Data	Long docs?
Ensemble of BM25 w/ KLI summarizer and fine-tuned BERT w/ LED summarizer	Askari and Verberne (2021)	COLIEE-2020 Task 1	EN	legal cases	Yes
Fine-tuned BERT (Span NLI BERT)	Koreeda and Manning (2021)	NLI, span detection	EN	contracts	Yes
Fine-tuned BART, PEGASUS, LED and PRIMERA	Shen et al. (2022)	AS	EN	legal cases	Yes
BERT + Bi-LSTM + CRF (SciBERT-HSLN); Fine-tuned BERTSUM + rhetorical roles (BERTSUM RR) Legal-PEGASUS + rhetorical roles (Legal-PEGASUS RR) XLNet + rhetorical roles (BERTSUM RR)	Kalamkar et al. (2022)	RRL, ES, AS, LJP	EN	legal cases	Yes
BERT, RoBERTa, DeBERTa, Longformer, BigBird, Legal-BERT@aub and Legal-BERT@stanford	Chalkidis et al. (2022b)	AVP, ALVP, TpC, TC, QA	EN	legal cases, contracts	Yes
Fine-tuned Longformer (Longformer-8192 & variants), Fine-tuned LegalBERT@aub + Longformer (LegalLongformer & variants) Fine-tuned LegalBERT@aub+TFIDF (TFIDF-SRT-LegalBERT)	Mamakas et al. (2022)	AVP, ALVP, TpC, TC, QA	EN	legal cases, contracts	Yes
Fine-tuned Hierarchical Attention Transformer (HAT)	Chalkidis et al. (2022a)	ALVP, TC, NLI	EN	legal cases, contracts	Yes
Fine-tuned BERT, DistilBERT, RoBERTa and XLNet	Malik et al. (2021)	CJPE	EN	legal cases	Yes

3.2.3 Domain-adaptive pre-training methods

In contrast to task-adaptive fine-tuning methods so far discussed, *domain-adaptive pre-training* allows for deeply tailoring an out-of-the-box model to a specialized language domain (Wang et al., 2020a; Gururangan et al., 2020; Song et al., 2022). In the legal domain, there are two main strategies that stand as alternative to the direct application of a TLM for the downstream task, namely (i) to continue pre-training the model on a legal corpus, or (ii) to pre-train the model from scratch on a legal corpus. In this section, we discuss methods that developed TLMs based on one or both the pre-training strategies and compared their performance effects on various tasks.

On different tasks of document review, Shaghaghian et al. (2020) propose to further pre-train and pre-train from scratch BERT models according to different strategies of tokenization and configuration of the initial weights of the model. More specifically, SentencePiece tokenization is performed on a corpus of US SEC legal agreements to generate the same number of cased tokens as in *bert-base-cased*; only 36% of tokens were found to be shared between tokens from the input legal corpus and the tokens in the original BERT. A hybrid tokenization is also devised whereby the 500 most frequent words in the input legal corpus are added to the token set if they do not occur as unbroken tokens in the set of general-domain tokens. Moreover, each of the BERT variants is also used as the teacher to customize as many DistilBERT models.

BERTLaw (Nguyen and Nguyen, 2021) is a model pre-trained on a collection of American legal cases with 8.2 million sentences (182 million words). BERTLaw has the same architecture and training objectives as BERT-base, however with an overlap between its vocabulary and the one in the original BERT-base model that is less than 50% of the size of the union of the two vocabularies. On the COLIEE-2019 Task 4, BERTLaw has shown to outperform BERT-base by almost 4% on validation data and over 16% on test data.

Chalkidis et al. (2020b) are the first to propose both further pre-training (FP) and pre-training from scratch (SC) BERT on legal corpora. Their models, hereinafter referred to as *Legal-BERT-FP@aueb* and *Legal-BERT-SC@aueb*,¹⁶ were trained on a collection of legal documents, including EU and UK legislations, ECJ cases, ECHR cases, US court cases, and US contracts. Legal-BERT-FP@aueb models are the result of prolonged in-domain pre-training of BERT-base up to 500K steps, in all or a selection of the training legal corpora, whereas Legal-BERT-SC@aueb has still the same architecture but with a new vocabulary equivalent in size to BERT-base. Also, a smaller Legal-BERT-SC@aueb model is derived with 6 layers, 512 hidden units, and 8 attention heads (35M parameters, 32% the size of BERT-base). All models were trained over the legal corpora in batches of 256 samples, including up to 512 sentence-piece tokens, with Adam optimizer and learning rate of 1e-4, as in the original BERT; nonetheless, for the fine-tuning stage, early-stopping was used instead of a fixed number of epochs, and more batch sizes, learning rates and

¹⁶ <https://huggingface.co/nlp-aueb>

drop-out regularization rate than the recommended ones in Devlin et al. (2019) were experimented. Legal-BERT-SC@aueb has shown to achieve a significantly lower training loss than the smaller model, while the latter turns out to be comparable to the original BERT-base. As for the Legal-BERT-FP@aueb models, the training loss can actually be improved when further pre-training on some selected corpora (but not on the whole collection of corpora). Considering the downstream tasks (EurLex57K, ECHR-Cases, Contracts-NER), Legal-BERT-FP@aueb and Legal-BERT-SC@aueb models outperform BERT-base and are as good as or better than the fine-tuned BERT-base, with more substantial gain in the ECHR-Cases multi-label task. Legal-BERT-FP@aueb and Legal-BERT-SC@aueb models also tend to be comparable to each other on all but the ECHR-Cases tasks, where Legal-BERT-SC@aueb is clearly better. The small Legal-BERT-SC@aueb is highly competitive with at least one of the larger variants in most tasks, suggesting that architecturally complex model may not be necessary when dealing with domain-specific sub-languages.

A legal BERT model is also developed by Holzenberger et al. (2020), in a context of evaluation of statutory reasoning (i.e., how to reason about an example case, based on rules extracted from statutes), particularly for US tax regulations. This model, hereinafter referred to as *Legal-BERT@jhu*,¹⁷ is obtained by further pre-training bert-base-cased on a portion of the case.law corpus, including both state and federal cases, for a total of 900M tokens. On the legal questions and answers of the corpus, Legal-BERT@jhu shows better perplexity than bert-base-cased (i.e., 2.7 against 14.4), which indicates improved adaptation to the legal queries. Legal-BERT@jhu was fine-tuned on the task of recognizing legal terms in the test set split of case.law, i.e., tokens or multi-token collocations that are defined in Black's Law Dictionary, a well-known legal dictionary.

Zheng et al. (2021) further investigate the use of BERT-based methods to pre-train TLMs on legal documents, starting from the premise that a substantial gain from law-specific pre-training might be prevented by fine-tuning tasks that are too easy and/or based on data that are inconsistent with the pre-training corpus. The proposed models, here dubbed *Legal-BERT@stanford*¹⁸ are pre-trained on the Harvard Law case corpus (i.e., case.law), where a random sample of 10% of decisions was extracted from the corpus and used as holdout set. More specifically, the variant Legal-BERT-FP@stanford is derived from the BERT-base model pre-trained for an additional 1M steps using the case.law corpus and the same vocabulary as BERT (uncased), while the variant Legal-BERT-SC@stanford is pre-trained from scratch for 2M steps using the case.law corpus, with a custom legal domain-specific vocabulary created using SentencePiece tokenization. Both variants use the same training objectives as BERT, i.e., MLM and next sentence prediction; for the latter, regular expressions are used to ensure that legal citations are included as part of a segmented sentence. Through a number of legal tasks varying in difficulty and domain-specificity, BERT has shown to already achieve high performance for easy tasks, while legal-specific pre-training

¹⁷ <https://nlp.jhu.edu/law>

¹⁸ <https://github.com/reglab/casehold>

takes small, resp. significant, advantage for mid-difficulty tasks that are not highly domain-specific, resp. high-difficulty and domain-specific tasks.

Leivaditi et al. (2020) further pre-train ALBERT on lease contracts. The resulting model, named *ALeaseBERT*, is fine-tuned and evaluated on the tasks of red flag detection and entity extraction. In particular, for the red flag detection, the additional pre-training proves to enhance the performance not only compared to lexical competitors, but also against ALBERT pre-trained from scratch as well as a fine-tuned ALBERT on the task.

Chi et al. (2023) further pre-train BERT, SpanBERT and ELECTRA on a pre-training corpus for privacy policy understanding provided in PLUE (cf. Section 2.4.2); the resulting pre-trained models are then fine-tuned on the benchmarks included in PLUE. Results reveal the advantage of these further pre-trained models compared to their only task-adapted counterparts (i.e., fine-tuned on the PLUE benchmarks). Moreover, the pre-trained models show comparable results w.r.t. a task-adapted Legal-BERT@aub.

Paul et al. (2022b) further pre-train Legal-BERT@aub and Legal-BERT@stanford on a corpus of about 5.4M Indian court cases. The resulting models, called InLegalBERT (derived from Legal-BERT@aub) and InCaseLawBERT (derived from Legal-BERT@stanford), are evaluated on three downstream tasks: legal statute retrieval, rhetorical role labeling, and court judgment prediction. In this regard, the two proposed models have been fine-tuned and assessed on various benchmarks against BERT, the original Legal-BERT@aub and Legal-BERT@stanford and the best performers according to the benchmark’s paper. In particular, for the legal statute retrieval task, Paul et al. (2022b) consider the ECtHR Task B of LexGLUE and a dataset of criminal case documents provided in Paul et al. (2022a). A hierarchical model — which consists of a lower-level encoder to provide the [CLS] representation of each sentence in the input text, followed by an higher-level encoder that aggregates the sentence embeddings to get the document representation — is used to get the relevant statutes, in a multi-label classification manner. Five versions of this model are defined, choosing the lower-level encoder among BERT, Legal-BERT@aub, Legal-BERT@stanford, InLegalBERT and InCaseLawBERT, whereas the high-level encoder is based on a LSTM with the attention mechanism. For the rhetorical role labeling task, the models are evaluated on a dataset of 50 documents from the Supreme Court of India (Bhattacharya et al., 2019b) and a dataset of as many documents from the Supreme Court of UK (Bhattacharya et al., 2021); a hierarchical model is again used, with the same variations of the lower-level encoder as in the previous task. For court judgment prediction, the authors consider the ILDC_{multi} dataset and the hierarchical BERT model provided by Malik et al. (2021), again with variations as above. Results show that both InLegalBERT and InCaseLawBERT improve the performances against LegalBERT@aub and Legal-BERT@stanford, respectively. Moreover, both InLegalBERT and InCaseLawBERT outperform the best performer of Paul et al. (2022a) and Malik et al. (2021), and InLegalBERT is better than the best performer of Bhattacharya et al. (2021), ECtHR Task B (Chalkidis et al., 2022b). Overall, results suggest that further pre-training

on Indian legal data can improve the performance of a model in several legal tasks across domains and countries.

Further applications of legal pre-trained models. The above discussed legal pre-trained models have recently attracted attention, especially Legal-BERT@aub models which have started to be widely used.

In Mahari (2021), Legal-BERT-SC@aub is fine-tuned for the legal precedent prediction (LPP) task, i.e., to predict relevant passages from precedential court decisions given the local context surrounding the passage in the opinion citing it (cf. Section 2.3.2). Using the US federal judicial opinions from the Case Law Access Project (CAP), the neighborhood of a passage is here comprised of 300 characters before the passage, 300 characters after the passage, 300 characters from the Introduction and 300 characters from the Conclusions. In 96% of unseen test examples, the correct passage is found to be within the top-10 predicted passages.

Ostendorff et al. (2021) compare Legal-BERT@aub and Legal-BERT@jhu — along with BERT (base and large), RoBERTa, S-BERT and S-RoBERTa, Longformer — against word-vector- and citation-based approaches on legal document recommendation as a text similarity task.¹⁹ A legal document is represented as a numerical vector, and the top-5 documents with the highest cosine similarity are selected through nearest neighbor search. Results have shown no absolute winner across all datasets and criteria, although a legal-domain adaptation of the FastText word embedding model Bojanowski et al. (2017); Joulin et al. (2017) was found the best performing method on average.

Bhattacharya et al. (2021) address the problem of rhetorical role labeling and propose different deep learning models to identify seven rhetorical classes over documents from five different law domains and from two different jurisdictions. BERT and LegalBERT@aub are fine-tuned on this task and used as bottom layers of a BiLSTM-CRF-based architecture. More specifically, the sentence embeddings produced by BERT and LegalBERT are used with a Hierarchical BiLSTM classifier and a Hierarchical BiLSTM CRF classifier with or without attention mechanism. Results show different outcomes in the model comparison, and in particular BERT and LegalBERT embeddings reveal to be beneficial for a Supreme Court of the UK dataset, but not for a Supreme Court of India dataset, compared to *Sent2Vec* (Pagliardini et al., 2018) embeddings. (*Sent2Vec* is a simple unsupervised model that considers sentence embeddings as the average of source word embeddings, considering both unigrams and n-grams).

Chalkidis et al. (2021b) address the task of regulatory information retrieval, where the query is a document (e.g. an EU directive) and the goal is to retrieve a set of relevant documents from the available law corpora. Upon two new datasets, EU2UK and UK2EU, a two-stage approach is defined, where document re-ranking is carried out after pre-fetching, i.e., retrieval of the top- k prominent documents for a query. As retrieval models, BERT, S-BERT and LegalBERT@aub are applied but showed to perform worse or comparably to BM25. As an improvement, following Chalkidis et al. (2019b), BERT is fine-tuned to predict EUROVOC concepts that describe the

¹⁹ <https://github.com/malteos/legal-document-similarity>

core subjects of each text: the resulting model, called C-BERT, turns out to be the best pre-fetcher by a large margin in EU2UK, while being comparable to BM25 in UK2EU; also, an ensemble combining C-BERT with BM25 further improves the results. For document re-ranking, several neural re-ranking methods are selected to yield a relevance score for each of the top- k documents returned by the best pre-fetcher: DRMM (Guo et al., 2016), PACRR (Hui et al., 2017), BERT-based re-rankers. DRMM is a deep matching model focused on relevance matching in ad-hoc retrieval. PACRR is a IR model that focuses on preserving positional information by incorporating domain insights. Chalkidis et al. (2021b) use DRMM and PACRR on top of BERT embeddings, and the resulting methods are called C-BERT-DRMM and C-BERT-PACRR. A major finding is that the re-rankers fail to improve the performance w.r.t. the ensemble pre-fetcher, as their term matching mechanisms tend to degenerate and over-utilize the pre-fetcher score.

For the statute law retrieval task of COLIEE 2021, Wehnert et al. (2021) fine-tune an ensemble model of Legal-BERT@aueb (legal-bert-base-uncased) and TF-IDF on sentence classification, to predict the relevance of an article to a query. Training and validation data are handled in three settings: in the original form (i.e., pairs of queries and relevant articles), decomposing instances with multiple relevant articles, and performing a data augmentation strategy to reduce data imbalance. In particular, the decomposed pairs are obtained separating the multiple relevant articles in different instances, producing a query-article pair for each relevant article. Data augmentation is performed by retrieving the top-50 non-relevant articles according to the highest cosine similarity between TF-IDF vectors of relevant articles and all the non-relevant articles, analogously to the approach in Nguyen et al. (2020) to reduce the unbalance between positive and negative samples. The resulting dataset is fed to Legal-BERT@aueb and TF-IDF, to obtain a softmax score of relevance and cosine similarity score, respectively. The overall score is computed as the average of the two scores after normalization. The top- n relevant articles are filtered using a threshold based on the precision-recall trade-off in the validation set.

The same team also applied for the Task 4 of COLIEE 2021 (Goebel et al., 2021), proposing two approaches: S-BERT combined with a *Graph Neural Network* (GNN) model and S-BERT combined with LegalBERT@aueb. In the first approach, GNN is expected to model the entailment relationships between nodes, where nodes represent articles/queries. The result is a bi-partite graph, in which queries can have connections to multiple relevant articles, and articles are enriched with metadata from the section titles. From the content of articles and queries, contextual sentence embeddings are generated using S-BERT and used as node features encoding external knowledge. Such node features are used to get query node embeddings following Morris et al. (2019), which encodes information taking into account also relevant articles as direct neighbor nodes. The query node embeddings are then used for query node classification to address the entailment task. In the second approach, the team adapt the LegalBERT@aueb encoder they proposed for Task 3 (Wehnert et al., 2021), re-initializing the classification head with new trainable parameters and splitting training data instances having multiple articles for a query in order to obtain additional instances. From the competition results, LegalBERT@aueb

submissions outperform the GNN-based approach; according to the authors, it is likely that the GNN was unable to generalize as well as LegalBERT@aub due to the limited amounts of data and the lack of document enrichment to encode external knowledge. Wehnert et al. (2022) further propose a new method for the COLIEE-2021 Task 4, which combines BERT and KERMIT (Zanzotto et al., 2020), an encoder that provides embeddings of syntactic parse trees. The idea is to inject further linguistic knowledge from the syntactic parse trees of query and articles. The output is the concatenation of the two embedding vectors. The trees for query and articles are obtained using the parser from the Stanford CoreNLP, as in Zanzotto et al. (2020). The concatenation of query and article is given to `bert-base-uncased` to get the sequence representation through the [CLS] resulting embedding. Both the embeddings of KERMIT and BERT are then passed to a fully connected decoder layer to get the entailment label. This third approach to Task 4 reaches higher performance than the GNN-based run and comparable results with the best LegalBERT@aub-based run.

Wang et al. (2023a) address the multiple-choice reading comprehension task of MAUD by fine-tuning several TLMs, in particular BERT, RoBERTa, LegalBERT@aub, DeBERTaV3 and BigBird for a single-task setting, and RoBERTa, LegalBERT@aub and DeBERTaV3 for a multi-task setting, which requires to fine-tune a model for all questions of MAUD. Newer and larger models provided better results, with BigBird and LegalBERT@aub being the best performers in the single-task and multi-task setting, respectively.

Shukla et al. (2022) evaluate different extractive and abstractive summarization methods on their own collected datasets (cf. Section 2.4.2): domain-agnostic methods, such as SummaRunner (Nallapati et al., 2017) and BERTSUM for extractive summarization, BART for abstractive summarization; domain-adapted methods, such as CaseSummarizer (Polsley et al., 2016) and Gist (Liu and Chen, 2019) for extractive summarization, Legal-PEGASUS²⁰ for short abstractive documents, Legal-LED²¹ for long abstractive documents; task-adapted methods, in which Transformers are fine-tuned on data generated with sentence similarity techniques. In addition, a newly hybrid extractive-abstractive method, called *BERT.BART*, is introduced and evaluated in both domain-adapted and task-adapted versions. As mentioned in Section 2.4.2, methods are to be evaluated w.r.t. document-level summaries, how well the summary represents the logical rhetorical segments in the legal case document, i.e., segment-wise evaluation, and how the summaries are evaluated by domain-experts. As regards the document-level evaluation, on the IN-Ext dataset, BERTSUM turns out to be the best extractive method, while the task-adapted Legal-PEGASUS and BERT-BART obtain the highest Rouge and BERTScore values, respectively; on the IN-Abs dataset, the task-adapted Legal-LED, BART and Legal-PEGASUS obtain the best BERTScore, R-1 Rouge and R-L Rouge scores, respectively; on the UK-Abs dataset, the task-adapted BART is the best performer in terms of R-L Rouge and BERTScore. As regards the segment-wise evaluation for the IN-Ext and UK-Abs

²⁰ <https://huggingface.co/ksi319/legal-pegasus>

²¹ <https://huggingface.co/ksi319/legal-led-base-16384>

datasets, results show a mixed picture, with Transformers obtaining the best score in most metrics; however, a discrepancy is noticed between the metrics used for summarization and the expert judgments: in fact, according to the involved law experts, extractive summaries are more satisfactory than the abstractive counterparts, since abstractive summaries are considered to be less organized and often incomplete.

3.2.4 Dealing with long legal documents

As discussed in Section 2.2, BERT and its early derivations can process texts having 512 tokens as maximum length. Input truncation to the maximum length has often been performed, which also holds for the methods discussed in the previous sections. However, this choice is clearly not optimal, especially when important information are spread through the whole document in different locations, which is also the case for legal documents. In this regard, text segmentation and text summarization can certainly be helpful.

In this section, we discuss methods that adopt different strategies for dealing with long documents. We organize this presentation by distinguishing between task-adaptive fine-tuning methods and domain-adaptive pre-training methods.

Task-adaptive methods. Rossi and Kanoulas (2019) address the COLIEE-2019 Task 1 as a ranking problem formulated as a pairwise relevance classification task, given the binary labels “Noticed” or “Not noticed”. This is treated as a downstream task for the pre-trained bert-base-uncased, where the hidden state of the [CLS] token is used as the pairwise embedding for a pair of summarized query case and candidate case, and the TextRank method (Barrios et al., 2016) is used for the summarization steps. The model has shown good generalization ability, although the widespread distribution of scores for the positive class and the small proportion of pairs with high score for the negative class might affect negatively the performance of the system.

Chalkidis et al. (2019a) face the length issue in legal cases proposing *Hier-BERT*. BERT is used to produce fact embeddings, then a self-attention mechanism applies to the fact embeddings and generates a single case embedding that is fed to a task-specific output layer. Three tasks are evaluated: binary violation classification, multi-label violation classification, and case importance prediction. The model is tested on ECHR (cf. Section 2.4.2). The list of facts in the fact description of each case are extracted using regular expressions. The violated articles are mapped to the violating case. An importance score is assigned for each case by the European Convention of Human Rights. The authors demonstrate that Hier-BERT significantly improves BERT in the aforementioned tasks.

BERT-PLI (Shao et al., 2020c) utilizes BERT to model paragraph-level interactions for legal case retrieval tasks, using a cascade framework.²² Initially, BM25 is used to filter the set of candidates and BERT is fine-tuned for sentence-pair binary

²² <https://github.com/ThuYShao/BERT-PLI-IJCAI2020>

classification on the COLIEE-2019 Task 2 data. The fine-tuned parameters are kept for the subsequent stage whose goal is to train the BERT-PLI model for relevance prediction. BERT-PLI receives a query and a selected document each broken into paragraphs. Each pair of query paragraph and document paragraph is an input to the fine-tuned BERT. The final hidden state vector of the [CLS] token is considered as the aggregate representation of the input pair. For each query and document, an interaction map of paragraph is finally produced. Then, for each query paragraph, a max-pooling strategy is used to capture the strongest signal with a candidate document. The ordered sequence of vectors resulting from the max-pooling on all paragraphs of the query is encoded by a RNN. An attention mechanism is applied to infer the importance of each position, and a document-level representation of the query-document pair is obtained via attentive aggregation. The resulting representation is provided to a fully connected layer with *softmax* to make a prediction of binary relevance. BERT-PLI has shown to achieve better recall and F1-score than the top-2 teams at COLIEE-2019.

To address the binary classification COLIEE-2020 Task 1, Shao et al. (2020b) apply a cascade framework in which a bi-gram LMIR (Song and Croft, 1999) is used to select top- N candidates to be further classified by BERT-PLI. The proposed system focuses mainly on the semantic understanding of documents, so that an additional solution is proposed to combine the semantic aspect of BERT-PLI with the exact matching among documents applied by a word-entity duet framework (Xiong et al., 2017). In particular, several features are extracted from BERT-PLI, BERT fine-tuned on sentence-pair classification using the first paragraph of query and candidate case, word-entity duet and similarity scores based on words and entities. A RankSVM model is then used to re-rank the top- N candidates, based on the aforementioned features. The resulting system performs better than the single BERT-PLI based solution, suggesting that the union of semantic understanding and exact matching can be beneficial. Analogously, in COLIEE 2021, Ma et al. (2021) apply BERT-PLI in Task 1 with some revisions. First, a LMIR model (Ponte and Croft, 1998) is used to filter candidates, then BERT is fine-tuned for a next sentence prediction task on the COLIEE-2019 Task 2. While Shao et al. (2020c) compute interactions with all query paragraphs and all top- N candidate paragraphs, Ma et al. (2021) consider only query paragraphs with sentences with a citation token, i.e., sentences with placeholders (for example ‘*REFERENCE_SUPPRESSED*’) that indicate the presence of mentions to other noticed cases. They also limit the number of paragraphs for a document. The proposed model outperformed the BERT-PLI based solutions and reached the first position.

Hendrycks et al. (2021b) fine-tune and evaluate BERT, RoBERTa, ALBERT and DeBERTa on the Contract Understanding Atticus Dataset (CUAD, cf. Section 2.4.2) for legal contract review, which is composed of contracts annotated with 41 category labels by legal experts. To deal with long documents, a sliding window mechanism is used over each document. DeBERTa turns out to perform the best over the other models, however the overall performances show that this is a challenging task, where the size of a model and the amount of training data play a very important role.

Aumiller et al. (2021) propose an approach to document segmentation of legal text to predict topical coherence among paragraphs and to detect topical change (cf. Section 2.3.2).²³ The approach consists in fine-tuning RoBERTa and Sentence-RoBERTa on a topic similarity task, which is formulated as a binary classification problem, i.e., to detect if paragraphs belong to similar topics or not. To this purpose, a dataset of Terms-of-Service documents, dubbed ToS (cf. Section 2.4.2), is created which includes hierarchical topic annotations. The classification task is performed under the assumption of topical independence of the context, i.e., the topic probability of a paragraph is not affected by the belonging of context paragraphs to the same topic. The fine-tuned models are then evaluated for sequential inference on entire documents, where segments are delimited by topical change. The approach has shown to outperform segmentation baselines based on TF-IDF and bag-of-words models.

For the COLIEE-2021 Task 2, Kim et al. (2021) fine-tune BERT on a binary classification problem using pairs of query and candidate paragraphs from the relative COLIEE dataset. To address the BERT token limit, a Transformer-based model generates a summary of the input. A subsequent phase of post-processing limits the maximum number of outputs for a given case. The final system is submitted to the competition in three runs, which differed in the setting of post-processing parameters.

Nguyen et al. (2021a) split the articles in multiple chunks using sliding windows to address the text length limitation of BERT and RoBERTa for the COLIEE-2021 Task 3. This is motivated from the observation that only some parts of an article entail the corresponding query. Training data are built as pairs of query and entailing article (positive sample) or any other article (negative sample), limiting the maximum number of negative samples. To avoid noisy training examples, a self-labeling technique is exploited, consisting of multiple fine-tuning phases in which training data labels are modified according to specific rules (Triguero et al., 2015). Different models, such as `bert-base-japanese`, `bert-base-japanese-whole-word-masking`, and `xlm-roberta-base`,²⁴ are trained on a sentence-pair classification task using the COLIEE-2020 dataset. The ensemble of those models performed better than the individual ones, reaching the second position in the competition. The same approach is also adapted for Task 4, using `bert-base-japanese-whole-word-masking`.

For the Japanese version of COLIEE-2021 Task 3, Yoshioka et al. (2021) improve their previous proposal, introduced in COLIEE 2020, based on the ensemble of a BERT-based retrieval system (Sakata et al., 2019) and the keyword-based Indri retrieval system.²⁵ A new article database is built combining articles and referred articles. To deal with long article sentences, the articles composed of a list of conditions or multiple judicial decisions are divided into small sequences through the numbered paragraph structure. Oversampling and negative sampling strategies are also applied to the training examples. The BERT-based and Indri scores for

²³ <https://github.com/dennlinger/TopicalChange>

²⁴ <https://huggingface.co/xlm-roberta-base>

²⁵ <https://www.lemurproject.org/indri/>

each article are combined to get an overall score. Three versions of the system are defined using different ensemble settings: the first version consists of BERT with oversampling, BERT with negative sampling and Indri, the second version uses BERT with oversampling only, and the third version corresponds to BERT with oversampling and Indri.

Chalkidis et al. (2021c) propose a method to learn automatically extracting paragraphs (i.e., the rationales) of the input so to justify decisions in alleged violation prediction. To this end, the ECtHR dataset (cf. Section 2.4.2) is used as it includes annotations for paragraph-level rationales. Extraction of the rationales is driven by rationale constraints. The proposed method is a variation of Hier-BERT (Chalkidis et al., 2019a), dubbed HierBERT-HA, following the framework of Lei et al. (2016) for the construction of rationales and consisting in three sub-networks: the first reads the text, the second extracts rationales through hard masking mechanism, and the third classifies the hard-masked text. More specifically, the paragraphs are first represented by the [CLS] embeddings obtained by LegalBERT@aueb (Chalkidis et al., 2020b). Then, a two-layer Transformer (Vaswani et al., 2017) obtains a contextualized representation of the paragraph embeddings which are fed to two separated fully-connected layers, the one producing paragraph encodings for the classifier sub-network, and the other one producing paragraph encodings for the rationale extraction sub-network. The rationale sub-network obtains a binarized attention score for each paragraph. Each paragraph is then hard-masked using the corresponding attention score. The resulting paragraphs are then fed to the classifier sub-network to predict the alleged violated articles. Different versions of the model using rationale constraints as regularizers (such as sparsity, comprehensiveness, singularity) are compared. A HierBERT-HA version without hard masking, dubbed HierBERT-ALL, and rationale constraints is tested too. Results show that models with hard attention and rationale constraints perform well and comparably to HierBERT-ALL.

Althammer et al. (2021) propose a two-stage pipeline for case law retrieval followed by re-ranking.²⁶ For the first stage, each query case and the cases in the corpus are split into their paragraphs, so that relevant prior cases can be retrieved for each paragraph of the query case based on the relevance of their paragraphs. Both lexical and semantic retrieval models are exploited, where the first is based on BM25 and the second corresponds to DPR based on two BERT-base-uncased Siamese Encoders (i.e., the one for encoding the query passage and the other one for the candidate passage). The relevance score between a query and a candidate passage is computed as the dot-product between their encoded vectors. The resulting model, called *LawDPR*, is trained on the entailing query-paragraph pairs of COLIEE-2021 Task 2. Results have shown that the paragraph-level retrieval is beneficial in terms of recall, and that LawDPR outperforms BM25, although a combination of both models can lead to further improved performance. In the stage of re-ranking of the retrieved cases, the approach is to summarize the texts of cases and queries, and then to predict whether the summarized query case is relevant to a summarized case.

²⁶ https://github.com/sophiaalthammer/dossier_coliee

To this purpose, the LED model is fine-tuned on the corpus for the corresponding binary classification task.

Deroy et al. (2021) analyze expert-generated summaries and model-generated summaries of legal case documents, to understand which parts of the documents are selected from models and domain-experts, and which sentences marked as important by domain-experts are easier or harder to identify for models. To this purpose, 15 extractive summarization models are evaluated over 50 case documents from the Indian Supreme Court. One of these models is BERTSum (Liu and Lapata, 2019), which is a BERT model fine-tuned for extractive summarization. Results show that BERTSum tends to select the initial portions of the document, even after fine-tuning. This is probably due to the pre-training of BERTSum, which is performed on news article corpus in which the first sentences are usually indicative of the content of the entire document.

Klaus et al. (2022) focus on extractive summarization of the European regulatory documents, in order to allow non-jurists (e.g., companies that want to ensure compliance with current regulations) a more facilitated comprehension of the documents and decide which regulations need more follow-up. They create the *EUR-LexSum* corpus,²⁷ consisting of 4K curated European regulatory documents with their summaries and structured into 32 policy fields (e.g., Taxation, Public Health and so on). Such documents are gathered from the Web,²⁸ which contains also the corresponding summaries in an abstractive form. In this regard, the authors perform a greedy strategy to obtain the extractive version of the summaries, before fine-tuning and feeding the data to a classifier. The obtained dataset is then split into 75% training, 15% validation, and 10% test documents. They fine-tune base uncased BERT, base uncased DistilBERT, and base RoBERTa (gathered from *TransformerSum*,²⁹ a library based on *BertSum* model (Liu and Lapata, 2019)), and address the summarization task as a binary classification of the sentence representations of the documents. Results prove that TLMs achieve superior performance compared to the *TextRank* (Mihalcea and Tarau, 2004) baseline, with BERT and DistilBERT performing better than RoBERTa. Furthermore, a combination of TextRank and TLMs is also evaluated, using the former as a pre-filter of candidate sentences, which achieves the highest F1 and precision scores in terms of *Rouge-1* metrics (Lin, 2004a).

Koreeda and Manning (2021) introduce *Span NLI BERT*, a BERT-based model that performs document-level NLI for non-disclosure agreements contracts. As explained in Section 2.3.2, the task is to find the implication relation (NLI) between a set of hypotheses and a contract (“entails”, “contradicts”, or “is neutral”), as well as to identify the spans in the contract which provide evidence for the associated relation. To deal with long documents, a dynamic context segmentation process is involved to split the text into overlapping contexts, assigning a pre-defined number of tokens to each context and marking the span boundaries within the context. The span boundaries are identified by special tokens [SPAN], whereas the input consists of a

²⁷ <https://github.com/svea-klaus/Legal-Document-Summarization>

²⁸ <https://eur-lex.europa.eu/browse/summaries.html>

²⁹ <https://transformersum.readthedocs.io/en/latest/>

[CLS] token followed by the hypothesis and the context separated by [SEP]. The task is thus reduced to a multi-label binary classification over the [SPAN] tokens for the evidence identification, and a classification task over the [CLS] token for NLI. Span-NLI-BERT was fine-tuned and tested on the ContractNLI dataset (cf. Section 2.4.2) against several baselines, based on SVM, TF-IDF and BERT. Span-NLI-BERT was also tested with different backbone models, including BERT without fine-tuning, BERT pre-trained from scratch with a case law corpus (Zheng et al., 2021), BERT fine-tuned on case law and contract corpora (Chalkidis et al., 2020b), DeBERTaV2 without fine-tuning and DeBERTa V2 fine-tuned on span identification (Hendrycks et al., 2021b); the latter turned out to be the best performing model. Overall, it was found that the correct identification of evidence greatly improves the NLI task, the performance of the model degrades in presence of rare labels, the negation by exception in the text damages the model’s accuracy for NLI, while the presence of references to definitions does not hurt performance.

Using the ILDC benchmark, Malik et al. (2021) compare BERT, DistilBERT, RoBERTa and XLNet models, including their hierarchical versions, against non-TLMs. An input text is divided into overlapping chunks of 512 tokens, and the [CLS] representation of each chunk produced by a TLM is passed to a sequential or feed-forward model to get the final case-decision prediction. The best performing model turns out to be the hierarchical model based on XLNet and a BiGRU on top. Malik et al. (2021) also consider a number of explainability approaches as a post-prediction step (cf. Section 3.2.7).

Inspired by SciBERT-HSLN (Brack et al., 2021), Kalamkar et al. (2022) propose a method based on BERT word embeddings along with Bi-LSTM and CRF, for the automatic prediction of rhetorical roles. Results show that SciBERT-HSLN outperforms a CRF model, which uses BERT sentence embeddings, and the model proposed in Cohan et al. (2019), which only uses BERT. Kalamkar et al. (2022) also experimented how rhetorical role prediction can be helpful in two applications: extractive and abstractive summarization of court judgments and court judgment prediction. For extractive summarization, BERTSUM is fine-tuned on the LAW-Briefs corpus, chunking the input in case of exceeding 512 tokens and pairing each input sentence with its rhetorical role. The resulting model, called *BERTSUM RR*, is compared against BERTSUM fine-tuned on data without information on the rhetorical roles. Similarly, for the abstractive summarization, Legal-PEGASUS is used by splitting the input in chunks (each of 1024 tokens) in one setting, and segmenting the document in terms of rhetorical roles in another setting, which is called *Legal-PEGASUS RR*). Results demonstrate that the use of rhetorical roles improves performance on both extractive and abstractive summarization. As regards the court judgment prediction task, XLNet is used on the model proposed in Malik et al. (2021), whereby a version of the model is trained on the last 512 tokens of the judgment text of the ILDC corpus, and another version is trained on the same data but filtered on the last 512 tokens sentences associated to the “analysis” role, which turns out to improve the predictions.

In Chalkidis et al. (2022b) hierarchical versions of several TLMs are evaluated on the LexGLUE data. These methods are particularly motivated by the inclusion of

three datasets in LexGLUE, i.e., ECtHR (A and B), SCOTUS and Multi-EURLEX, whose texts substantially exceed the limit of 512 tokens. Even models that handle longer text sequences, such as Longformer, are not able to avoid the truncation of texts in these datasets. Following Chalkidis et al. (2021c), each paragraph is encoded using the corresponding TLM-based encoder, then a shallow version of the TLM is fed with paragraph encodings and obtains a new representation of a paragraph that is aware of the surrounding paragraphs. The document representation is then obtained from the context-aware paragraph representations through max-pooling. Results show the benefit of this hierarchical approach.

Using the same three LexGLUE datasets, Mamakas et al. (2022) propose to modify Longformer in order to deal up to 8192 tokens and to combine Legal-BERT@aub with TF-IDF representations. In the former case, the proposed *Longformer-8192* is designed to increase the maximum input length up to 8192 tokens and to decrease the local attention window size from 512 to 128 to reduce computational burden. A variant called *Longformer-8192-PAR* is also introduced which is Longformer-8192 with the addition of the global token [SEP] between paragraphs to get paragraph-level representations. Also, the following models are introduced: *LegalLongformer*, which is a Longformer warm-started from LegalBERT@aub and able to handle up to 4096 tokens, *LegalLongformer-8192* and *LegalLongformer-8192-PAR*, which are Longformer-8192 and Longformer-8192-PAR, respectively, based on LegalLongformer. For the second approach, *TFIDF-SRT-LegalBERT* is proposed as a version of LegalBERT@aub in which duplicate sub-words are removed from the input text to decrease its length. The remaining sub-words are ordered by decreasing TF-IDF, so that the model is induced to attend more the earlier sub-words (with higher TF-IDF). This model is also extended by including a TF-IDF embedding layer. Both the approaches are fine-tuned and evaluated on LexGLUE tasks. The models of the first approach achieve better results on almost all tasks, and surpass the best model for Chalkidis et al. (2022b), i.e., the hierarchical version of LegalBERT. This demonstrates the effectiveness of warm-start from a legally pre-trained model and the adding of further positional embeddings and global tokens.

A hierarchical approach for handling long documents is also proposed in Chalkidis et al. (2022a) by using *Hierarchical Attention Transformer* (HAT) models. The HAT models are based on a multi-level hierarchical attention pattern, which comprises a segment-wise encoder (SWE) that applies to segments independently, followed by a cross-segment encoder (CSE), to get contextual representation across segments. Each segment is preceded by the [CLS] token, which represents the segment encoding. These two encoders can be arranged to form different architectural topologies. The text segmentation process is based on a dynamic segmentation strategy that aims to preserve sentence integrity while minimizing padding. The HAT models are pre-trained with the MLM objective task and the parameters of the HAT models are warm-started following BERT and RoBERTa. Thus, they are fine-tuned and evaluated on benchmarks of different domains, including ContractNLI and ECtHR. Results demonstrate that HATs perform better across all tasks than sparse-attention models like Longformer and BigBird, and also they are less demanding in terms of computation and memory resources.

Shukla et al. (2022) also evaluate three methods to deal with long legal documents: TLMs designed for long text (e.g., Longformer), chunking approaches along with TLMs designed for short summaries (e.g., BART and Legal-PEGASUS), and extractive summarization techniques to reduce the size of the input document combined with abstractive methods. For the latter approach, Shukla et al. (2022) propose BERT_BART, which first uses a BERT-based extractive summarization model to extract relevant sentences, then carries out a chunking-based BART model to get the final abstractive summary. Nonetheless, the chunking-based methods (i.e., Legal-PEGASUS and BART task-adapted through fine-tuning) and Legal-LED turn out to be the best performers for the document-level evaluation and for the segment-wise evaluation, respectively.

Shen et al. (2022) fine-tune and evaluate BART, PEGASUS, LED and PRIMERA for abstractive summarization on the Multi-LexSum benchmark. In particular, for the multi-document summarization task, LED and PRIMERA, which deal with long inputs, perform better than PEGASUS and BART on all the three tasks of Multi-LexSum, although no model has shown to be able to generate long summaries of length similar to those provided by humans. When considering short summaries, gold summaries provided in Multi-LexSum are used as input to summarize, instead of the source document. In this setting, on single-document summarization tasks, PRIMERA reaches the performance of PEGASUS and BART, while the performance degrades when BART-generated summaries are used as input instead of the gold ones. Long and tiny versions of generated summaries have benefited from multi-document summarization compared to the summaries of the single-document counterpart.

Domain-adaptive pre-training methods. There is also a body of works jointly focusing on domain-adaptive pre-training and handling long documents. The common approach is to apply in-house or existing (e.g., Legal-BERT@aueb) further/from-scratch pre-trained models on segmented or summarized versions of long legal documents.

In the context of employment notice prediction, Lam et al. (2020) adapt BERT-*base* and RoBERTa-*base* on a corpus of notice case texts in addition to the Harvard case law dataset. Using only the MLM criterion objective, the two models are further pre-trained for text classification with ten epochs and fine-tuned on 409 remaining cases. The goal is to predict reasonable notice based on a free-text summary of the case law, where the summaries are unstructured text data written in plain, non-legal English collected from WestLaw’s Quantum service. Surprisingly, however, results have shown a decrease in performance by the domain-adapted models against the out-of-the-box RoBERTa.

Limsopatham (2021) compare BERT, Legal-BERT@aueb, Legal-BERT@stanford, and RoBERTa on a text classification task. After removing tokens in the rear, resp. front, of the input texts to ensure the 512 token length, each model is applied on chunks of 200 tokens. An average pooling layer, or alternatively max pooling layer, is used before the classification layer. Results obtained on the ECHR Violation dataset and the Overruling Task dataset, which are multi-label and binary classification tasks, respectively, show that Legal-BERT@aueb and Legal-BERT@stanford achieve bet-

ter results than the other methods. However, fine-tuned BigBird and fine-tuned Longformer (even though not pre-trained on in-domain documents) outperform the other approaches that adapt BERT to deal with long documents by truncating them.

Khazaeli et al. (2021) propose a QA system for both factoid and non-factoid questions. It comprises a search repository, which contains the passages and their learned embeddings, a search engine, which retrieves passages by text and embedding similarity, and an answer finder, which re-ranks the retrieved passages. In the stage of passage retrieval, it makes use of a Siamese BERT architecture (a.k.a. S-BERT) (Reimers and Gurevych, 2019). A Siamese Legal BERT system is hence trained on a collection of headnotes to retrieve similar passages, with a regression objective function with cosine loss. A legal BERT model pre-trained from scratch, with a custom legal vocabulary on US case-law documents, receives in input a sentence embedding, and uses mean pooling of the tokens embedding. The answer finder is trained by fine-tuning the in-house Legal BERT model, whose classifier uses the [CLS] representation with two fully connected layers with a final softmax layer. For each input question-passage pair (q, p) , the answer finder concatenates their texts in the form '[CLS]q[SEP]p[SEP]', and computes the probability that the passage answers the question.

Li et al. (2021b) consider the COLIEE-2021 Task 1 as a ranking problem and address the BERT's input length limitation handling the text at paragraph and document level. Firstly, the input is divided into paragraphs. Each paragraph is encoded by a `bert-base-uncased` model that is further pre-trained on the Task 1 corpus, dubbed *BERT-Legal*. The pre-training task is similar to the MLM task but relies on the n-gram masking method to obtain the masked input. Each paragraph is represented by the hidden vector obtained for the [CLS] token. Then, a paragraph level coherence matrix of the query-candidate pairs is derived. The document level representation is obtained encoding the coherence matrix through a max-pooling strategy and a LSTM model. The output of the LSTM is adopted to get the contextual based semantic ranking through relevance prediction. The BERT-Legal ranking is part of a retrieval pipeline, in which the first stage is to recall top-50 relevant cases using BM25-based similarity score. To reduce the training consumption, the authors set a maximum number of paragraphs for a document. The proposed solution has not achieved satisfactory results, probably due to a bad candidates selection operated by the BM25-based recall method. For the Task 2, BERT-Legal is fine-tuned for binary classification on the case law entailment dataset. The authors propose two further versions of the model: the one leverages Task 1 dataset to augment the positive and negative samples in the training set, the other one adversarially train BERT-Legal with Fast Gradient Method (Miyato et al., 2017) to be robust to embedding perturbations. However, both the strategies have shown to decrease the performance of BERT-Legal on the test set, probably due to the different data distribution between Task 1 and Task 2.

Furniturewala et al. (2021) address both summarization tasks at AILA-2021. For Task 2a, two approaches are proposed: (i) to fine-tune Legal-BERT@aub, and (ii) to concatenate statistical feature vectors obtained by TF-IDF with the semantic feature vectors obtained by the Legal-BERT@aub in correspondence to the [CLS] token.

Table 3.3: Summary of main methods discussed in Section 3.3.

Method	Ref.	Downstream Tasks	Lang.	Data	Long docs?
Zero-shot GPT-3	Il and Katz (2022)	MCQA	EN	bar exams	Yes
Zero-shot/Few-shot GPT-3	Blair-Stanek et al. (2023)	NLI, QA	EN	CAP cases	Yes
Zero-shot/Few-shot/Fine-tuned GPT-3	Yu et al. (2022a)	COLIEE 2021 Task 4	EN	civil code	Yes
Fine-tuned GPT-3 (LawGPT 1.0)	Nguyen (2023)	QA	EN	legal cases	Yes
Zero-shot ChatGPT	Choi et al. (2023)	QA	EN	constitution	Yes

The final joint features are classified by a SVM model and logistic regression. For Task 2b, the sentences that were labeled as relevant in Task 2a were concatenated into one summary. The authors reached the first position for Task 2a and the best scores as regards some of the Rouge metrics for Task 2b.

Concerning the same tasks at AILA-2021, Jain et al. (2021) use Legal-BERT@aub to get contextual embeddings of the sentences and fed them to an MLP model. Five training datasets are obtained, so that, for Task 2a, the probability for a sentence to be relevant is the average of all the five probabilities computed by the models. a sentence is labeled as relevant if the average probability computed over the five models exceeds a certain threshold. A second run is proposed for Task 2a, in which both rhetorical and relevance labels are considered using a multi-task learning MLP model. This is motivated by the intuition that rhetorical labels can be helpful for predicting relevance labels. A third run is also proposed, combining the results of the first two runs by averaging all the individual probabilities of each individual trained model. For Task 2b, the average probabilities computed by each run of Task 2a are used to rank the sentences. The sentences in top positions are picked to get the summary, until the summary length limitation imposed by the organizers is reached. All the runs for Task 2a reached leading positions, surpassed only by Furniturewala et al. (2021). For Task 2b, the first run achieved the highest F1 score for all the Rouge metrics, the best recall for some of them, and the second-highest precision for all the metrics.

Askari and Verberne (2021) address the COLIEE-2020 and 2021 Task 1 by creating shorter query documents based on three approaches: term extraction with Kullback-Leibler divergence for informativeness (KLI), noun phrase or entity extraction, and abstractive summarization using LED fine-tuned on the COLIEE-2018 data containing human-written summaries of the case documents. Such summaries are then combined using lexical ranking methods based on BM25 and statistical language modeling as well as neural ranking methods based on DRMM, Legal-BERT@aub, and a version of BERT and Longformer designed for document ranking following MacAvaney et al. (2019). Ensemble models are also proposed using the scores of neural and lexical rankers as features for a classifier (the authors experimented with SVM, Naive Bayes and MLP). Such methods are primarily evaluated on COLIEE-2020 data. The results show that the best performance is obtained by the ensemble of BM25 (with KLI summarizer) and BERT (with LED summarizer), surpassing the first team at COLIEE-2020 (Westermann et al., 2020). On COLIEE 2021, BM25 with KLI summarizer surpassed the first team at the competition (Ma et al., 2021).

3.2.5 Methods for non-English legal languages

Prompted by the success in English text data, in the past few years there has been an increased interest in replicating the advances in NLP based on TLMs for many other languages. As a result, a plethora of non-English BERT and related models have been trained and developed, such as CamemBERT (Martin et al., 2020), BART_{hez} (Eddine et al., 2021) and FlauBERT (Le et al., 2020) for French, GermanBERT,³⁰ GBERT and GELECTRA (Chan et al., 2020) for German, ALBERT_o (Polignano et al., 2019), GILBERT_o³¹ and UmBERT_o³² for Italian, KoBERT³³ and KoBART³⁴ for Korean, RobBERT (Delobelle et al., 2020) and BERT_{je} (de Vries et al., 2019) for Dutch, RoBERT (Masala et al., 2020) and Romanian BERT (Dumitrescu et al., 2020) for Romanian, Greek-BERT (Koutsikakis et al., 2020) for Greek, RoBERT_{tuito} (Pérez et al., 2022) and BETO³⁵ for Spanish, AraBERT (Antoun et al., 2020) for Arabic, MacBERT (Cui et al., 2020), PERT (Cui et al., 2022) and MarkBERT (Li et al., 2022) for Chinese, AlephBERT (Seker et al., 2021) for Hebrew, BERT_{imbau} (Souza et al., 2020), RoBERT_a-PT-BR³⁶ and GPT_{uguese}-2 (Guillou, 2020) for Brazilian, and so on.

In specialized domains, such as the legal one, the challenge underlying the development of TLMs is even more evident, due to the lack of large domain-specific training data for specific languages. In this section, we briefly describe main methods for non-English TLM-based legal learning, organized by language.

Brazilian. Feijó and Moreira (2019) experiment with extractive and abstractive models to summarize documents of legal rulings. To this purpose, the *RulingBR* dataset is used, which contains 10K rulings from the Brazilian Supreme Court, and is divided in 60% training, 20% validation and 20% test samples. Each ruling is organized in sections, the first of which is the summary. The abstractive approaches involved are based on neural networks, and include a Transformer model and a *TransformerAAN* model (Zhang et al., 2018). The latter uses a cumulative average attention as an alternative to the self-attention in the decoder side to accelerate the decoding procedure. Results show that abstractive approaches outperform extractive methods, with Transformer getting the highest scores.

Lage-Freitas et al. (2022) evaluate classic machine learning techniques and deep learning models, including BERT_{imbau}, on court decision prediction and unanimity decision prediction. In general, due to the small size of the dataset they propose (cf. Section 2.4.2), deep learning techniques are outperformed by other methods. One exception is for BERT_{imbau}, which obtains comparable or slightly higher performance when predicting on three-label case outcomes (yes, no, partial) with

³⁰ <https://www.deepset.ai/german-bert>

³¹ <https://github.com/idb-ita/GILBERTo>

³² <https://github.com/musixmatchresearch/umberto>

³³ <https://github.com/SKTBrain/KoBERT>

³⁴ <https://github.com/SKT-AI/KoBART>

³⁵ <https://github.com/dccuchile/beto>

³⁶ <https://huggingface.co/josu/roberta-pt-br>

imbalanced setting of the dataset. Regarding unanimity prediction, deep learning models perform better than the classic ones in F1 score when the data set is balanced.

Polo et al. (2021) release a library³⁷ containing available pre-trained language models, including BERT, for the Brazilian legal language. The library also includes a package with useful functions and demo examples to facilitate the use of the models. The models are trained using datasets from several sources, in particular BERT is trained using three datasets, regarding clippings and motions from different Brazilian courts and longer documents from the Court of Justice of São Paulo, starting from a checkpoint of Souza et al. (2020). The resulting model is dubbed *Bertikal*.

Aguiar et al. (2021) evaluate different NLP methods along with several combinations of embeddings of Portuguese language models on the lawsuit classification task. Such methods are trained on a collection of petitions and indictments from the Brazilian Court of Justice of the State of Ceará, in which BERTimbau obtains best performance when lawsuit is embedded as the concatenation of all petitions containing one or more references to a legislation.

Serras and Finger (2022) focus on the categorization of case laws considering a Brazilian legal dataset.³⁸ In particular, they fine-tune m-BERT and the *base* and *large* version of BERTimbau. In the fine-tuning process, pairs of summary and header are given, and the aim is to generate the terms of the header from the summary. From the experimental results, it is observed that BERT-based methods outperform a statistical-based baseline. Larger BERT models perform slightly better than the base ones, highlighting that base models are robust enough to the deal with the task.

³⁷ <https://github.com/felipemaiapolo/legalnlp>

³⁸ Two reformulated versions of the Kollemata dataset: <https://www.kollemata.com.br/>. The dataset is not publicly available.

Table 3.4: Summary of main methods discussed in Sections 3.2.5 and 3.2.6.

Method	Ref.	Downstream Tasks	Lang.	Data	Long docs?
Fine-tuned Transformer and TransformerAAN	Feijó and Moreira (2019)	AS	BR	legal judgments	Yes
Fine-tuned BERTimbau	Lage-Freitas et al. (2022)	case law classification	BR	legal judgments	No
Fine-tuned BERTimbau	Aguiar et al. (2021)	lawsuit classification	BR	petitions, indictments	No
Fine-tuned m-BERT and BERTimbau	Serras and Finger (2022)	case law classification	BR	summaries, headers	Yes
Fine-tuned Transformer + GCN (R-former)	Dong and Niu (2021)	LJP	CN	legal cases	No
BERT + BiLSTM + CRF	Sun et al. (2021)	NER	CN	legal cases	No
CEEN (BERT as alternative to BiLSTM module)	Lyu et al. (2022)	LJP	CN	legal cases	No
Further pre-trained LegalBERT@OpenClap+attention (EPM)	Feng et al. (2022)	LJP	CN	legal cases	No
Inverse optimal transport-based rationale extraction approach + fine-tuned Chinese T5-PEGASUS (IOT-Match)	Yu et al. (2022b)	case matching prediction	CN	legal cases	Yes
Fine-tuned BERT + BiLSTM +CRF; UniLM (CLASS)	Li et al. (2021a)	RRL; AS	CN	legal cases	Yes
Fine-tuned Transformer and BERT + Graph Transformer	Huang et al. (2021)	AS	CN	legal opinion news	Yes
Pre-trained Longformer (Lawformer)	Xiao et al. (2021)	LJP; CR; RC; QA	CN	CAIL-Long; LeCaRD; CJRC; JEC-QA	Yes
Fine-tuned FlauBERT	Salaün et al. (2020)	lawsuit classification	FR	legal cases	No
From-scratch pre-trained BERT (JuriBERT); Further pre-trained CamemBERT (JuriBERT-FP)	Douka et al. (2021)	pleading classification	FR	legal cases	No
Further pre-trained BARThez (CriminelBART)	Garneau et al. (2021)	charge prediction (Cloze tests)	FR	legal cases	No
Zero-shot and fine-tuned CamemBERT	Louis and Spanakis (2022)	SAR	FR	federal/regional codes	Yes
Further pre-trained CamemBERT+Transformer+graph model (G-DSR)	Louis et al. (2023)	SAR	FR	federal/regional codes	Yes
Fine-tuned GBERT and GELECTRA	Wrzalik and Krechel (2021)	case law classification	DE	legal cases	No

Continued on the next page

Method	Ref.	Downstream Tasks	Lang.	Data	Long docs?
GermanBERT and DistilBERT	Tang and Clematide (2021)	document retrieval	DE	articles and legal cases	No
Fine-tuned m-BERT, XLM-RoBERTa, Greek-BERT, and Greek-Legal-BERT	Papaloukas et al. (2021)	article classification	GR	civil code	No
Fine-tuned m-BERT	Tarasconi et al. (2020)	NER; TC	IT	legal cases	No
Fine-tuned BERT (LamBERTa)	Tagarelli and Simeri (2022)	SAR	IT	civil code	No
Further pre-trained BERT (ITALIAN-LEGAL-BERT)	Licari and Comandè (2022)	NER, SC, SS	IT	legal cases	No
Fine-tuned KoBERT and KoBART	Yoon et al. (2022)	AS	KR	legal cases	No
From-scratch pre-trained BERT	Masala et al. (2021)	case law classification	RO	legal cases	No
From-scratch pre-trained RoBERTa	Gutiérrez-Fandiño et al. (2021a)	NER; TC	ES	various legal texts	No

BERT and m-BERT	Aydemir et al. (2020)	COLIEE-2020 Tasks 3-4	EN, JP	civil code	No
Fine-tuned hierarchical BERT + BiLSTM; BERT (Long BERT)	Niklaus et al. (2021)	LJP	IT, DE, FR	legal cases	Yes
Fine-tuned monolingual German-BERT, Camembert, UmBERTo	Niklaus et al. (2022)	LJP	IT, DE, FR	legal cases	Yes
Fine-tuned multi-lingual XLM- RoBERTa					
Fine-tuned m-BERT and m-DistilBERT	Shaheen et al. (2021)	TC	EN, DE, FR	JRC-Acquis; EURLEX57K	No
Fine-tuned Legal-BERT@aueb, m-BERT, WikiBERT, BERT	Avram et al. (2021)	TC	various EU lang.	JRC-Acquis; OPOCE	No
Fine-tuned m-BERT- and DistilBERT-based ParaLaw Nets	Nguyen et al. (2021b)	COLIEE-2021 Task 5	EN, JP	civil code	No
zero-shot mGPT, GPT-J-6B, GPT-NeoX-20B	Trautmann et al. (2022)	LJP	EN, IT, DE, FR	legal cases	Yes

Chinese. Dong and Niu (2021) address the legal judgment prediction task as a graph node classification problem, handling constraints among articles, charges and terms of penalty. Using the training set, a global consistency graph is derived, composed of all the possible relations of class labels (i.e., articles, charges and terms-of-penalty) treated as graph nodes. To prevent logically conflicting judgment results, relational learning is introduced into a Transformer model to achieve global and local consistency. The resulting model is dubbed *R-former*. *R-former*'s architecture includes a node encoder module and a node classification module. The first module is composed of two Transformers (Dai et al., 2019): the former obtains the article representations from raw text, the latter uses masking mechanisms to extract consistency information among nodes belonging to different tasks (article prediction, charge prediction and terms-of-penalty prediction tasks) and distinction information of the same task. The second module consists of a graph convolution network to obtain the relevance score of each node according to the neighbors in the consistency graph. Experiments are conducted using *CAIL-small* and *CAIL-big* datasets from Chinese AI and Law challenge (CAIL2018) (Xiao et al., 2018), with *R-former* performing the best among all the evaluated competitors.

Sun et al. (2021) combine BERT, BiLSTM and Conditional Random Fields (CRF) to identify legal case entities. To build a name entity recognition model, the authors leverage BERT's ability of text feature extraction, using it as input layer in order to get word embeddings. In addition, they employ a BiLSTM to get long-term memory information. Finally, state-transition matrix in CRF is used to output the globally optimal sequence, with more accurate labeling according to specific conditions. The resulting model is evaluated using the corpus of People's Daily newspaper along with Legal Case Texts from the CAIL Law Research Cups and other legal texts manually labeled. The authors find that using BERT's word embeddings instead of Word2Vec (Mikolov et al., 2013) improves greatly the recognition accuracy.

Lyu et al. (2022) present the so-called Criminal Element Extraction Network (CEEN) for reinforced criminal element extraction useful for predicting legal judgments. CEEN is designed to handle misleading law articles, having very similar TF-IDF representations, and indistinguishable descriptions of facts, having different targets and criminals. CEEN is composed of four parts: a fact description encoder, a reinforcement-learning-based element extractor, a criminal element discriminator and a multi-task judgment predictor. The fact description encoder is used to obtain contextual sentence representations of facts. The encoder is typically performed by four hierarchical BiLSTMs, although it can also be a BERT-based variant of CEEN, dubbed $CEEN_{BERT}$, which exploits the addition of the [CLS] token at the beginning of each sentence in the input, like Liu and Lapata (2019). After obtaining sentence representations, a reinforcement-learning-based element extractor is used to distinguish confusing fact descriptions, then a criminal element discriminator gets the discriminative criminal element representations. Finally, a multi-task judgment predictor outputs the judgment results. Experiments were conducted using *CAIL-small* and *CAIL-big* datasets from Chinese AI and Law challenge (CAIL2018) (Xiao et al., 2018) and a set of baselines, including Chinese BERT, RoBERTa and Lawformer. Experimental results show that the combination of CEEN and BERT significantly

overcomes BERT and all the baselines, indicating that language knowledge from pre-training and criminal elements' identification are complementary to each other.

Feng et al. (2022) propose to leverage event extraction from the fact description of criminal cases to constrain models for the LJP task on the CAIL-2018 benchmark. Indeed, most of the incorrect predictions are due to an inadequate identification of key events in the fact description that determine the final judgment, and the lack of consistency constraints among CAIL sub-tasks (law article, charge and term of penalty prediction). In this regard, the *EPM* model is introduced, which is based on a legal BERT (Zhong et al., 2019a) referred to as LegalBERT@OpenClap, and an attention mechanism, which is guided by event-based and cross-task consistency constraints. EPM is pre-trained on the training set of CAIL and fine-tuned on *LJP-E*, a portion of CAIL-small dataset, provided by the authors and manually-annotated with event triggers and roles. Results have shown that EPM outperforms several baselines for the benchmark and the event extraction process performed jointly with the LJP task is beneficial also for the performance of the competitors.

Yu et al. (2022b) propose an explainable method, called IOT-Match, for the task of case matching prediction. Given the sentences of a pair of legal cases, it extracts rationales based on semantics and legal characteristics, and generates explanations so that the matching prediction (carried out on eCAIL and ELAM, cf. Section 2.4.2) is based on the extracted rationales and explanations. More details on IOT-Match are discussed later in Section 3.2.7.

Li et al. (2021a) propose *CLASS* (Chinese Legal judgmentS Summarization), a method to generate abstractive summaries of Chinese legal judgments. Once relevant sentences are extracted from the input, the legal judgments are split into rhetorical roles, then a summary of each rhetorical role is generated. The extraction module uses BERT to get the embedding of each sentence, then a Bi-LSTM model encodes the sequence of sentences. A BERT-BiLSTM-CRF model is trained to split the legal judgments into the rhetorical roles. The UniLM model is chosen to get the abstractive summary of each rhetorical role. CLASS is evaluated using the legal judgment summarization dataset of CAIL2020, divided in 80% training, 10% validation and 10% test samples. Results show that CLASS can achieve higher performance compared to sequence-to-sequence competitors.

Huang et al. (2021) propose a graph-augmented abstractive summarization model for the automatic summarization of legal public opinion news. A separate graph encoder generates a structural representation of the source document. The components of a document are organized into two graphs, named element relational graph (ERG) and topic interaction graph (TIG). In ERG, the elements extracted from the document (i.e., entities, keywords, and event triples) are nodes that are connected with virtual nodes representing Person, Location, Keyword and so on, in order to get the document-level interconnections. In TIG, similarities of different nodes are represented, where the nodes are the elements of the document that can be viewed as topics and the graph can aid to detect the main topic of the document. The model is composed of a sequence encoder, to yield the sequential representation of the document, a graph encoder, to yield the structural representation of the document, and a sequence decoder, to generate the summary constrained by a dual attention mech-

anism on the sequential and structural representation of the encoders. The graph encoder is a Graph Transformer Network (Yun et al., 2019), which incorporates global structural information while learning from the neighbor nodes. The sequence encoder and the decoder are based on a vanilla Transformer. A BERT-based embedding mechanism is applied to improve the performance by enhancing the sequence encoder’s embedding and by initializing the nodes of the element graph. The model is evaluated using a legal public opinion summarization corpus, named *LPO-news*, consisting of article-summary pairs from the Sina Weibo website. Extensive experiments demonstrate that the proposed model outperforms competing baselines in terms of both Rouge and BERTScore metrics, also when compared on general news-oriented datasets, and that it can generate more coherent, faithful and informative summaries.

The first pre-trained model for legal long documents is *Lawformer* (Xiao et al., 2021).³⁹ Following Longformer, Lawformer is pre-trained with MLM objective, continuing from the checkpoint RoBERTa-wwm-ext on a collection of criminal and civil cases published by the Chinese government from China judgment Online. Lawformer is then fine-tuned on a number of tasks, namely legal judgment prediction, legal case retrieval, legal reading comprehension, and legal question answering, for which datasets CAIL-Long, LeCaRD, CJRC, and JEC-QA are used, respectively.

French. Salaün et al. (2020) fine-tune the FlauBERT base cased language model (Le et al., 2020), which is a pre-trained BERT on French corpora, to a collection of lawsuits submitted to a tribunal specialized in disputes between tenants and landlords in Quebec, i.e., the Régie du Logement du Québec (RDL). The RDL lawsuits are organized into three sections: fact description, legal reasoning and the verdict. The documents also contain metadata describing court location, judge, presence and type of each party: legal person (juridical entity) or physical person, single person or multiple people, male or female and so on. The task is to classify the outcomes of the judgments with three possible labels: penalty (judge convicts the defendant), agreement (judge imposes an agreement) and rejection (judge rejects the the plaintiff’s claims). FlauBERT is trained without the use of metadata and it achieves better performance when trained on all texts (discarding verdicts).

Douka et al. (2021) propose to pre-train small sizes of BERT (base, small, mini and tiny) from scratch using an MLM task and French legal datasets. The latter consist of the decisions of the Court and the Claimant’s pleadings from the Court of Cassation, together with raw legal texts crawled from *Légifrance* website⁴⁰. The resulting model, dubbed JuriBERT,⁴¹ is then tested on two downstream classification tasks: to assign the Court’s Claimant’s pleadings to Chambers and Sections of the Court of Cassation (8 labels), and to classify the Claimant’s pleadings to a set of subjects (151 labels). A variant of JuriBERT, pre-trained on Claimant’s pleadings, is also provided in *tiny* and *mini* versions, as well as another JuriBERT (JuriBERT-FP) obtained by

³⁹ <https://github.com/thunlp/LegalPLMs>

⁴⁰ <https://www.legifrance.gouv.fr/>

⁴¹ <http://master2-bigdata.polytechnique.fr/FrenchLinguisticResources/resources#juribert>

further pre-training CamemBERT (Martin et al., 2020). From the evaluation results, it is observed that JuriBERT-*small* outperforms larger models when training on specific sub-languages like the legal one. However, due to limited resources, larger models like JuriBERT-FP and JuriBERT-*base* have been pre-trained with the use of smaller batch sizes than the other models. By fixing a model size, it is shown that JuriBERT pre-trained from scratch on the same task-specific data used in the fine-tuning can lead to better performance compared with only domain-specific models. JuriBERT-FP outperforms JuriBERT-*base*, demonstrating that further pre-training a general-purpose model can be a preferable choice.

Garneau et al. (2021) propose a further pre-training of BARThez (Eddine et al., 2021) on a French legal corpus for criminal law, collected from the Criminal and Penal Chamber and mined from the Société Québécoise d'Information Juridique (SOQUIJ) website⁴². The legal comprehension of the resulting model, dubbed CriminalBART, was evaluated through Cloze tests regarding the prediction of criminal charges, legal provisions, and privacy.

Louis and Spanakis (2022) evaluate lexical and dense models on the BSARD benchmark (cf. Section 2.4.2). The lexical models are based on TF-IDF and BM25 and are used to retrieve the top- k articles for a given question, based on scores computed for each article. The dense models are bi-encoder models, in two architectures: *siamese*, which uses a unique word embedding model for both questions and articles, and *two-tower*, which uses two separated word embedding models. For each question, a similarity score is computed for all articles that are pre-encoded, and finally the top- k articles based on calculated scores are retrieved. The dense models are evaluated in a zero-shot setting, using word2vec and CamemBERT. To handle long articles for CamemBERT, each text is divided in overlapping chunks. The fine-tuned CamemBERT has shown to outperform all the other competitors but, considering only the zero-shot variants, the word2vec-based model gives significantly higher performance w.r.t. BERT-based models.

The BSARD benchmark is also employed in Louis et al. (2023), where a graph-augmented dense retrieval model, called *G-DSR*, is proposed to exploit the statute hierarchy to enrich the article information. *G-DSR* includes two components: a dense statute retriever and a legislative graph encoder. The first component is a bi-encoder for articles and questions, so that a question and an article relevant to the question get similar representations. Questions are encoded using the [CLS] token representation obtained by CamemBERT further pre-trained on BSARD articles, while the encoding process for statutory articles is left to a hierarchical encoder to handle the length of the documents. Given a pair of article and question, a similarity score is calculated to evaluate the relevance of the article for the question. The bi-encoder is trained through contrastive learning to get effective embedding functions. The second component of *G-DSR* represents the statute's hierarchy as a directed acyclic graph with two types of nodes: section nodes, corresponding to the headings of the code subdivisions, and the article nodes, representing the textual content of the articles. The edges of the graph model the hierarchy between sections and articles. The semantic information

⁴² <https://soquij.qc.ca/a/fr/>

of a node is initialized through the article encoder of the first block and used as the initial node features. A graph neural network is then employed to update the node features by aggregating the local neighborhood information based on the graph structure. The resulting model show higher performance against baselines such as BM25, DPR and a combination of BM25 and mT5. Also, the enriched information given by the legislative graph encoder to the dense statute retriever contributes significantly to improve performance.

German. Wrzalik and Krechel (2021) fine-tune GBERT and GELECTRA *base* models to re-rank retrieved documents of *GerDaLIR* (cf. Section 2.4.2). Passages that exceed the sequence length limitation of BERT and ELECTRA are divided along sentence boundaries, then the maximum score is assigned to the passage. ELECTRA demonstrates to obtain higher re-ranking quality compared to BERT.

Tang and Clematide (2021) address paragraph-level semantic similarity for legal document retrieval⁴³, using a corpus of legal cases and statutes in German language gathered from the Swiss Federal Court⁴⁴ and Swiss Government⁴⁵ websites, and extracting from the cases citations that point to statutes. They use GermanBERT, a BERT model trained on partially legal German text, and the German variant of DistilBERT. An extended attention mask mechanism is performed to combine the *idf* scores of non-neural methods with neural models. This mechanism suppresses low informative tokens in the input, thus impeding the self-attention calculation on those tokens. The authors developed a link-based similarity method to estimate paragraph-level semantic similarity, considering the relations between paragraph cases that share citations to the same statutes. From the experimental results, it appears that GermanBERT, along with the use of the extended attention mask mechanism, offers a clear added value compared to non-neural competitors at specific *idf* threshold.

Greek. In Papaloukas et al. (2021) multiple methods are evaluated on the GLC dataset (cf. Section 2.4.2) for classifying legal texts. In particular, the authors experimented with m-BERT, XLM-RoBERTa, Greek-BERT (Koutsikakis et al., 2020) and Greek-Legal-BERT (Athinaios, 2020). Greek-Legal-BERT is pre-trained on documents gathered from a Greek legislative Knowledge Base, called Nomothesia⁴⁶ (Chalkidis et al., 2017). Excluding a few exceptions, Greek-Legal-BERT proves to be the best performer for each evaluation level (i.e., volume, chapter and subject level), followed by Greek-BERT, m-BERT and XLM-RoBERTa, thus confirming the importance of both in-domain and in-language training.

Italian. Tarasconi et al. (2020) evaluate different BERT-based approaches for three business problems in the processing of case law contents for electronic publishing purposes: identification of legal references in the text, new content classification based on relevance, hierarchical labeling of text according to predetermined topics. In the first case, the aim is to identify in a judgment the references to specific laws or

⁴³ <https://github.com/lilytang2017>

⁴⁴ <https://www.bger.ch/it/index.htm>

⁴⁵ <https://www.fedlex.admin.ch>

⁴⁶ <http://legislation.di.uoa.gr/>

other judgments. Judgments come from the Italian Highest Courts of Appeal. Best performances are obtained using a fine-tuned version of m-BERT for NER purposes. In the second case, the identification of the potential relevance of a document aims to select the ones to be eventually published. Documents are mostly gathered from judgments from the Italian Highest Courts of Appeal, but also T.A.R. Administrative Regional Tribunal, Italian Constitutional Court and EU courts. The task is addressed using BERT for binary classification, although best results are obtained using a Random Forest model because of the use of hand-crafted features. In the third case, the goal is to assign each document of the Italian Highest Courts of Appeal, with a set of topics belonging to a publisher's proprietary resource. The task is addressed using a fine-tuned version of m-BERT for extreme multi-label classification. The high-quality of legal data, collected over the years by publisher's managers and decision-makers, allows to successfully experiment with supervised methods.

LamBERTa Tagarelli and Simeri (2022) is the first BERT-based framework for law article retrieval as a prediction problem, focusing on the modeling, learning and understanding of the Italian Civil Code (ICC). To this purpose, LamBERTa fine-tunes a pre-trained Italian BERT on the ICC and is designed to answer legal questions. The task is conceived as an sequence classification task with a high number (i.e., hundreds or thousands) of classes, which correspond to the number of articles in the ICC, resp. a within-book portion of it, that is used to train a LamBERTa global model, resp. book-specific model. Also, the task is a few-shot learning one, since there are few per-class examples to train a model, which are extracted from each article of the ICC according to one of several schemes of *unsupervised training-instance labeling* defined by the authors. These schemes adopt different strategies for selecting and combining portions from each article to derive the training set, while sharing the requirements of generating a minimum number of training units per article. LamBERTa models have been assessed through single-label as well as multi-label evaluation tasks, based on six different types of queries, which include jurisprudential sentences associated with the ICC articles, and annotations about the interpretation of the meanings and law implications associated to the articles. Also, Simeri and Tagarelli (2023) focus on an investigation of the injection of out-of-vocabulary legal terms in LamBERTa models' tokenizer and of the impact of domain-adaptive pre-training of LamBERTa models on article retrieval performance.

Licari and Comandè (2022) contribute with ITALIAN-LEGAL-BERT, which is the result of a further pre-training of a pre-trained Italian BERT on a corpus extracted from the National Jurisprudential Archive (pst.giustizia.it), a repository containing millions of legal documents, such as decrees, orders, and civil judgments, from Italian courts and courts of appeal. ITALIAN-LEGAL-BERT was evaluated on named entity recognition, sentence classification, and sentence similarity tasks.

Japanese. As previously described, many efforts have been made by researchers in the attempt of modeling TLMs on the Japanese version of COLIEE tasks, especially on statute law retrieval, entailment and question answering.

Korean. Yoon et al. (2022) use a KoBERT-based version of BERT2BERT, a model composed of BERT on both the encoder and decoder side, as well as KoBART to

generate abstractive summaries of Korean legal judgments. The dataset is a collection of legal precedents gathered from the *Korean Court Comprehensive Legal Information* site.⁴⁷ Both the models show good results on the task, with KoBART performing better than BERT2BERT.

Romanian. Masala et al. (2021) specialize BERT models for Romanian juridical domain.⁴⁸ To this end, the models are pre-trained from scratch using *RoJur*, a large corpus containing civil and criminal cases documents published by Romanian civil courts, and the *whole word masking* technique. In *RoJur*, each judgment is composed of the description of the parties, brief of the arguments, legal reasoning and final verdict. The resulting model, dubbed *JurBERT*, is then evaluated using two datasets: *RoBanking*, extracting from *RoJur* common types of cases relating to banking domain, and *BRDCases*, a collection of cases involving the Romanian BRD bank. The downstream task is to predict if the final verdict in a legal case is in favor of the defendant or the plaintiff, so it is addressed as binary classification.

Spanish. Gutiérrez-Fandiño et al. (2021a) train a RoBERTa model on a large legal corpora obtained from a collection of different Spanish datasets, including Legal-ES (Samy et al., 2020). The resulting model is dubbed *RoBERTalex*, and has been compared with m-BERT and a Spanish RoBERTa (Gutiérrez-Fandiño et al., 2021b). Since there is no domain-specific evaluation dataset available, the models are tested on general-domain tasks (e.g., NER, classification), on which RoBERTalex obtains good performance.

3.2.6 Multilingual and cross-lingual methods

Unlike English and few high resource languages, many other languages have often been characterized by low resource data, sometimes resulting in limited or absent benchmarks, which negatively affect the amount of pre-training data, and hence of performed of TLMs. To overcome this issue, one common approach is to develop *multilingual language models*, which are pre-trained using large amounts of unlabeled data from multiple languages, under the assumption that low resource languages can benefit from high resource languages due to shared vocabulary and semantic relatedness aspects. Multilingual TLMs have been indeed proposed in the past three years, such as m-BERT and XLM-RoBERTa, mainly differing from each other in terms of number of languages involved, architecture components, pre-training objective functions and corpora. A comprehensive survey on multilingual TLMs is recently provided in Doddapaneni et al. (2021). In the following, we shall instead keep our focus on some representative studies of multilingual and cross-lingual TLM-based legal learning.

Aydemir et al. (2020) propose to use m-BERT along with the BERT-*large* cased and uncased models to address the COLIEE-2020 Task 3. Results have shown that

⁴⁷ <https://glaw.scourt.go.kr/wsjo/intesrch/sjo022.do>

⁴⁸ <https://huggingface.co/readerbench>

the multilingual BERT model yields better performance than BERT-*large* cased. However, for Task 4, results based on multilingual BERT appeared not be satisfactory as for Task 3, even compared to the competition baseline.

Niklaus et al. (2021) experiment with several BERT-based methods on binary classification of the judgment outcome, using Italian, German and French legal cases from the Federal Supreme Court of Switzerland (FSCS). Among them, there is also a hierarchical version of BERT similar to Chalkidis et al. (2019a), where a BERT (monolingual or multilingual) encoder produces fact embeddings that are used as input for a BiLSTM to get the document representation. Another variant of BERT (monolingual or multilingual) proposed in Niklaus et al. (2021) to deal with long document is *Long BERT*, where a maximum length of 2048 tokens is reached replicating four times the original 512 positional encodings. Results unveil that monolingual models generally outperform the multilingual counterparts, with the overall best results obtained by hierarchical BERT.

Niklaus et al. (2022) evaluate cross-lingual, cross-domain (i.e., cross-legal areas), cross-regional and cross-jurisdiction transfer learning of several TLMs on the LJP task. An augmented version of the SJP corpus (cf. Section 2.4.2) is provided. Since most SJP documents are around 2048 tokens, Hierarchical BERT models (Niklaus et al., 2021; Chalkidis et al., 2019a) are used to encode up to 2048 tokens for each document. For the cross-lingual transfer learning, monolingual BERT-based models (German-BERT, Camembert, and UmBERTo), and the multi-lingual XLM-RoBERTa are compared on three scenarios: fine-tuning the models for a specific language (i.e., monolingual fine-tuning), fine-tuning the models across languages (i.e., cross-lingual fine-tuning), and fine-tuning across languages but excluding the target language (i.e., zero-shot cross-lingual fine-tuning). For the first and the second scenarios, two versions of the training set are used, one containing only the documents in the original language contained in the SJP corpus and the other including also the machine-translated versions of the documents. Concerning the monolingual fine-tuning of monolingual models and XLM-RoBERTa, results have shown that monolingual models obtain better performance w.r.t. XLM-RoBERTa, with even better results than Niklaus et al. (2021). Moreover, the augmentation of the data through machine-translation seems to further improve the performance. For cross-lingual fine-tuning, a standard fine-tuning of XLM-RoBERTa is compared with an adapter-based fine-tuning of the model (Houlsby et al., 2019; Pfeiffer et al., 2020), which consists on the addition of adapter layers to the model and training them along with the parameters of the normalization layers. Results unveil the beneficial effect of including adapters in XLM-RoBERTa, regardless of the addition of machine-translated documents, and adapters also improve performance for XLM-RoBERTa fine-tuned with zero-shot cross-lingual fine-tuning. As regards cross-regional transfer evaluation, the SJP documents are divided w.r.t. several regions, then XLM-RoBERTa is fine-tuned in three settings: fine-tuning w.r.t. a specific region with data augmentation, fine-tuning across all the regions without machine-translated data augmentation and fine-tuning across all the regions with machine-translated data augmentation. Results show that in most cases a model fine-tuned on the same region of the target is outperformed by zero-shot models (i.e., models

fine-tuned on another region). However, cross-regional models obtain better results than regional-specific models, with adapter-based models obtaining top results in most cases. As regards cross-domain transfer learning evaluation, fine-tuning is conducted over three settings: domain-specific fine-tuning with data augmentation, cross-domain fine-tuning without data augmentation and cross-domain fine-tuning with data augmentation. Results demonstrate that cross-domain models outperform domain-specific models in most cases. Unlike the cross-regional transfer analysis, models fine-tuned on the same domain of the target outperform the zero-shot models. Finally, Niklaus et al. (2022) perform a cross-jurisdiction transfer learning evaluation, by adding to the SJP corpus, concerning the Swiss laws, the ILDC corpus, which contains Indian laws. They consider two fine-tuning scenarios: fine-tuning XLM-RoBERTa on only the machine-translated Indian cases (zero-shot fine-tuning) and fine-tuning the models with original SJP training set, machine-translated SJP cases and machine-translated ILDC cases (further augmented fine-tuning). Results show that zero-shot models perform poorly, but the further-augmented models outperform the results of the cross-lingual models evaluated in the cross-lingual transfer learning analysis. Again, adapter-based models obtain highest performance in most cases.

Shaheen et al. (2021) evaluate m-BERT and m-DistilBERT in large-scale multi-label text classification task using a zero-shot cross-lingual transfer learning scheme and a joint learning scheme. In the first case, models are built using an English training set although performance is tested on French and German test sets. In the second case, models are trained using all the three languages. To this purpose, the multilingual version of JRC-Acquis (Steinberger et al., 2006) and an extended version of EURLEX57K (Chalkidis et al., 2019b) are exploited. From the experimental results, m-BERT outperforms m-DistilBERT in zero-shot transfer learning scheme in both datasets. As regards the joint learning scheme, m-BERT and m-DistilBERT are compared with monolingual BERT-based methods using the English test set. The multilingual models obtain about 96.83-98.39% of the performances of monolingual models in JRC-Acquis, while in EURLEX57K better results are obtained by m-BERT. Comparing the performance in zero-shot and joint learning scheme, m-BERT (resp. m-DistilBERT) in zero-shot scheme achieves about 86% (resp. 79%) of the performance in the joint learning scheme on French and German test sets. The zero-shot results indicate that multilingual models are able to achieve transfer learning from English to French language more easily than from English to German.

Avram et al. (2021) propose a tool for multi-label classification of legal documents on 22 languages. For each language, BERT-based methods are fine-tuned using the JRC-Acquis (Steinberger et al., 2006) and the *Publications Office of the European Union* (OPOCE) corpora, manually labelled with almost 7K EuroVoc⁴⁹ descriptors. The choice of BERT models for each language is conducted prioritising pre-trained models on legal domain corpora, otherwise according to the following order: models pre-trained on a specific language, models pre-trained on monolingual Wikipedia, multilingual models. As a result, Legal-BERT (Chalkidis et al., 2020b) is used for the English language, m-BERT (Devlin et al., 2019) for the Maltese language,

⁴⁹ <https://data.europa.eu/data/datasets/eurovoc>

WikiBERT models (Pyysalo et al., 2021) for Bulgarian, Lithuanian, Latvian, Slovak and Slovene, and monolingual BERT models for the remaining 14 languages. From the evaluation results, m-BERT obtain the worst performance, probably due to the low number of Maltese documents as well as due to the use of a multilingual model. In contrast, it is interesting that WikiBERT models obtain satisfactory scores, even better than Legal-BERT.

Nguyen et al. (2021a) introduce *ParaLaw Nets* (Nguyen et al., 2021b) to address the COLIEE-2021 Task 5. This is a family of pre-trained models that rely on sentence-level translation information to reduce language ambiguity and increase performance in legal tasks. The core idea in these models is that, in the translation process, the most correct meaning of a sentence will be captured by the model. To this end, the model is pre-trained on sentence-level cross-lingual tasks, while the COLIEE task is used in the fine-tuning phase. According to the type of pre-training task, ParaLaw Nets are divided in *NFSP* (Next Foreign Sentence Prediction) and *NMSP* (Neighbor Multilingual Sentence Prediction) models. The former formulates the pre-training task as binary classification, the latter as multi-label classification. The pre-training input consists of sentence pairs, where a sentence can be in its original language or in the target language. As previously described in Section 3.2.2, the NFSP task is basically a Next Sentence Prediction task where the input is composed of sentences pairs expressed in different languages. In the NMSP models, training data also includes sentence pairs with the same language. The labels in the NMSP task correspond to four training data generation schemes, namely random sampling, normal order, reverse order, and non-contiguous. In both tasks, m-BERT and DistilBERT's architectures are used respectively as the base and the distilled version for ParaLaw Nets. The models are trained with Japanese-English legal data, but the approach can be generalized to all language pairs. The data used to pre-train ParaLaw Nets is provided by Japanese Law Translation website.⁵⁰ The data used in the fine-tuning phase come from the Japanese Civil Code and COLIEE. The authors use also augmentation strategies in the fine-tuning phase, which mainly consist in the statement negation of the original sentences. From the competition results, it is observed that the base model using NFSP obtains the highest score, while the base model using NMSP reaches the third position, proving the effectiveness in exploiting cross-lingual information.

Trautmann et al. (2022) explore the potential of legal prompt engineering on zero-shot GPT-based models for multilingual LJP. In particular, they evaluate mGPT, GPT-J-6B,⁵¹ i.e., a 6 billion parameter version of GPT trained on the Pile dataset (Gao et al., 2021), and GPT-NeoX-20B on the ECHR and FSCS datasets. A prompt template is defined by mapping the LJP task in a question-answer form, where the question requests to detect whether or not there are violated articles in the document. Results have shown that the proposed prompting methods outperform simple unsupervised baselines, but perform worse than supervised (task-adapted) models, thus leaving much room for improvement especially in terms of defining

⁵⁰ <https://www.japaneselawtranslation.go.jp/>

⁵¹ <https://github.com/kingoflolz/mesh-transformer-jax>

templates suitable to different corpora and languages (as well as dealing with long documents without truncating at 2048 tokens).

3.2.7 Dealing with explainability and interpretability issues

As for any sophisticated neural network models, the TLMs' behavior at both learning and inference phases is not straightforward to fully understand. This clearly may limit their applicability, or even increase fear in legal actors or general public that who may want to use such AI-based tools, especially for high societal impact fields like law. It is therefore demanding to resort to techniques that can provide explainable justification for the models' decisions, or that can make their prediction outcomes interpretable.

Post-hoc explanation. One approach is to infer post-hoc explanations. Within this view, some works such as Savelka and Ashley (2021) and Tagarelli and Simeri (2022) provide insights into the informativeness of the attention weights produced by a BERT model for selected use cases, through the lens of the interactive visualization tool *bertviz* (Vig, 2019).⁵² The general objective is to inspect the formation of complex inter-token relationships and the corresponding distinctive attention patterns that most influence the final representation of a given sentence in input and, therefore, such as to justify the model's outcome for that input. A different perspective is taken in Simeri and Tagarelli (2023), where the explanation is accomplished by approximating a BERT-based classifier locally by an interpretable linear model, by means of a technique known as LIME - Local Interpretable Model-Agnostic Explanations (Ribeiro et al., 2016). For a given input sentence, LIME explains a classifier's behavior "around" that query instance, by weighing perturbed versions of the input by their proximity to the original query instance, then observing the associated predictions by the underlying classifier to determine which of those changes will have most impact on the prediction of the original query.

Wehnert et al. (2022) address interpretability using KERMIT to encode symbolic syntactic parse trees of queries and articles in addition to BERT representation of input sentence, thus injecting further linguistic knowledge. KERMIT is specifically designed to include syntactic interpretations in deep neural networks. The KERMITviz architecture (Ranaldi et al., 2021) enables the visualization of which part of the sentence is used during the inference step.

Malik et al. (2021) evaluate a number of explainability methods as a post-prediction step of the case decision prediction task. In this regard, legal experts are asked to mark the sentences they consider as explanations for the judgments (from a portion of the ILDC test set) and assign each explanation sentence with a score reflecting the importance of the explanation for the judgment. A hierarchical system composed of XLNet and BiGRU on top is chosen as the base model on which to experiment an explainability method based on *occlusion* and inspired from Zeiler

⁵² <https://github.com/jessevig/bertviz>

and Fergus (2014) and Li et al. (2016a). Documents are divided into chunks of 512 tokens, each chunk embedding is masked at a time, then the masked input is given to the BiGRU component of the model. A chunk explainability score is computed as the difference between the output probabilities of the prediction calculated on the masked input and the unmasked input. The sentences that explain the final decision correspond to the chunks with positive scores. Similarly, each sentence of selected chunks is masked at a time and supplied to the XLNet component of the model. The difference between the logits calculated on the masked input and the original logits of the prediction represent the explanation score, so that the top-k sentences for each chunk are selected as explanations. The analysis of the explanations given by the occlusion method indicate that most of the relevant information for the judgment is located at the end of the document; however, explanations selected by the occlusion method are shown to be significantly different from explanations given by legal experts.

Early explanation. A different approach is to produce explanations during the data modeling or learning process. Liu et al. (2021b) consider the interrelation between charges and the court view sections in legal documents. The court view is regarded as charge explanation since it contains supporting information for a charge and is charge-discriminative (i.e., strongly charge-dependent). Within this view, the JPGM method is introduced to jointly predict charges based on fact descriptions and generate court views. The key idea is to predict a group of similar charges that may lead to confusion. To this purpose, charge-discriminative keywords are defined for each charge, then an attention mechanism is involved to select the best matching keywords for a charge predicted by a classifier, finally the generation module provides the court view based on the fact description and the best keywords. The generated court view and fact description are also fused to refine the previous classifier prediction. JPGM has been shown to perform better than baselines based on CNN, GRU, vanilla Transformer, and graph neural networks.

In the context of alleged violation prediction, Chalkidis et al. (2021c) propose to regularize the extraction of rationales, as the paragraphs of the input that support the decisions, by constraints that reward the model if its decisions are based on concise rationales. The rationale constraints include *sparsity*, *continuity*, *comprehensiveness* and *singularity*. The first two constraints encourage to select a small number of paragraphs that sufficiently justify the allegation and to prefer contiguous paragraphs, respectively. The comprehensiveness constraint requires to use a mask to ideally contain all the paragraphs that support the correct decision, whereas the reverse mask should contain the irrelevant paragraphs; this way, the output of the model based on the mask (i.e., the estimated probabilities on the allegations) should be better than the output of the model based on the reverse mask. Moreover, the singularity constraint requires that the selected mask should be not only better than its reverse version, but also w.r.t. any other mask. The effect of such constraints are evaluated w.r.t. *faithfulness* and *rationale quality*, given a sparsity threshold, where the latter is calculated comparing the predicted rationales with gold annotations, and faithfulness measures how much the rationales reflect the reasoning of the model, by computing

the difference between the predicted probabilities obtained by the model over the whole text in input and the predicted probabilities obtained considering only the (complement of) extracted rationales. The singularity constraint is shown to improve faithfulness and rationale quality in relation to both silver and gold rationales; by contrast, continuity appears not to be beneficial for the task at hand.

Santosh et al. (2022) start from the observation that LJP models on ECtHR cases tend to be confused by distracting factors in the text that might originate from the corpus construction, the case distribution, or spurious correlations with the outcome. In this regard, an expert-informed deconfounding method based on adversarial training is introduced to prevent a model from being influenced by distractors recognized by ECtHR experts. To this purpose, Santosh et al. (2022) train and evaluate BERT models on LexGLUE (ECtHR Task A and B) and ECHR benchmarks; in particular, a BERT variant of hierarchical attention networks Yang et al. (2016) is chosen as base model, where input texts are segmented with a greedy sentence packing strategy and encoded using Legal-BERT@aub.

To evaluate the alignment with expert rationales, the expert relevance assessments provided for ECtHR by Chalkidis et al. (2021c) are used. The ability of the model to identify correct rationales at the paragraph level is evaluated through an interpretability technique that determines the importance score for each paragraph by measuring the impact of a particular input token on the final prediction. A token-level focus score is calculated using the integrated gradients (Sundararajan et al., 2017), then paragraph-level scores are obtained aggregating the token scores in the paragraph. The top-k paragraphs according to the paragraph scores are compared with the golden paragraph rationales. Results have demonstrated that the deconfounding process effectively improves the model alignment with expert rationales in ECHR benchmarks; however, such improvements are marginal and there is still a large gap between paragraphs suggested by models and what legal experts have annotated as relevant.

To address a legal case matching task, Yu et al. (2022b) emphasize the importance of taking into account the difference between the roles of sentences corresponding to rationales and other sentences in a legal case, as well as distinguishing rationales that are in favor or against the matching decision. Each sentence of a given pair of legal cases is assigned one of the following labels: not a rationale, a key circumstance, a constitutive element of a crime or a focus of disputes. The goal is to extract aligned and misaligned rationales, to assign a matching label for the legal case pair (not matching, partially matching or matching) and to provide the set of sentences explaining the reasons for the label. The problem to extract aligned and misaligned rationales is formulated as an optimal transport problem, where the probability of the cross-sentence coherency (pro and con rationale pairs) is computed to provide evidence for the matching. The problem is guided by an affinity matrix, which reflects both semantics and legal feature relations between cross-case sentences. The affinity matrix and the optimal transport, based on both sentence embeddings and a manually-labeled (noisy) alignment matrix, are learned by the inverse optimal transport (IOT) process. The IOT process consists of solving a bi-level optimization problem, in which the affinity matrix is the upper-level variable and the optimal

transport is the lower-level variable. The optimal transport is then fitted to the alignment matrix and used to help the generation of the explanations. To this purpose, a Chinese T5-PEGASUS (Su, 2021) model for each matching label is fine-tuned to get the label-specific candidate explanations. Finally, the model performs the matching prediction based only on the extracted rationales and the label-specific explanations. In this way, the noisy sentences of the legal cases pair are filtered out and not involved in the matching prediction, reducing the number of input sentences to be processed. The proposed IOT-Match model is fine-tuned and tested on eCAIL and ELAM datasets against legal case matching competitors. The quality of the extracted rationales and the generated explanations is verified by conducting an empirical analysis on faithfulness (described above) and *plausibility*, which measures how the model explanations are convincing to humans. In particular, to assess the plausibility of rationales, resp. explanations, the ones produced by the IOT-Match and competitors have been compared with those generated by humans, unveiling that the rationales, resp. explanations, by IOT-Match are more consistent with human annotations. Moreover, as concerns the faithfulness of rationales, IOT-Match has been evaluated with and without rationales, eventually finding that the extracted rationales play an important role for the task of legal case matching.

As previously described in Section 3.2.6, Feng et al. (2022) propose an event extraction process from the fact description of criminal cases to constrain the prediction of LJP models on CAIL2018. A hierarchy is defined for the events of legal cases, based on Chinese law articles. Events are detected from specific triggers, i.e., words expressing the occurrence of the event. Each event has a role in the fact description, and the extracted events and constraints on such events are used to help the learning process of the EPM model for the LJP task. The context representation of the fact description is used to query all law article candidates and to determine the most relevant semantics in the articles. The context representation of the fact description and the most article relevant semantics are given as input for the CAIL sub-tasks. The base model is augmented with a hierarchical event extraction layer to identify event triggers and roles while jointly performing the LJP task. The judgment prediction is thus inferred through detected event features. The model is guided by *event-based constraints*, in order to search for a single trigger and its related roles, and *cross-task consistency constraints*, in order to take into account dependencies between the sub-tasks, since each article establishes the charges and the range of penalty terms.

3.3 Methods for Legal Generation Tasks

Since the launch of ChatGPT in November 2022, there has been a worldwide renewed interest towards generative language models, which has inevitably impacted on the legal domain as well. In the following, we overview a few works focusing on GPT-based models that have very recently appeared in the literature.

II and Katz (2022) evaluate the performance of a zero-shot GPT-3 (text-davinci-003) model on the multiple choice component of the US Bar Exam. The questionnaire is typically divided into eight categories, seven of which regard specific law areas and one is used to experiment with the test design. Each category contains 25 questions and 4 candidate answers for each question. Different types of prompts were assessed to get the best answers from the model, and the best strategy turned out to be asking the model to sort the top three multiple-choice answers for the question. Results show that the model is not able to pass the entire exam with its first choices, however it greatly outperforms baselines based on random choices. Moreover, it reaches the average passing range (58%-62%) in two categories. When considering the top two answers, the model passes the average passing range in all categories and exceeds the average results obtained by human examiners in 5 out of 7 categories, with an overall average score on all categories of 71%, in contrast to the average score of 68% obtained by humans. Results get even better when considering the top three answers, with an overall average score of 88%.

Blair-Stanek et al. (2023) evaluate several prompting approaches on GPT-3 (text-davinci-003) to perform statutory reasoning on the SARA benchmark. In particular, they consider dynamic few-shot prompting, chain-of-thought prompting and zero-shot prompting. In the few-shot scenario, the four most similar training cases are provided for each test case; the prompt pattern for the training case consists of the text of case labeled with the premise, the hypothesis and the answer (entailment or contradiction), whereas the prompt pattern of the test case contains only the premise and the hypothesis. In the zero-shot prompting, no training cases are used. In the chain-of-thought prompting, a prompt consists of ten training cases, each having the same labels as for the few-shot prompting, followed by a chain-of-thought that explains the reasoning leading to the conclusion (i.e., entailment or contradiction). The same prompt pattern is used for the test cases. Moreover, the prompts can be augmented by adding the statute(s) required to solve the entailment reasoning, since GPT-3 may also have been trained on the US tax code. Following Kojima et al. (2022), if the model does not explicitly indicate the result of the reasoning (entailment or contradiction), the model is prompted again with the original prompt augmented with the model's answer and adding a clear answer imposition string at the end, i.e., *"Therefore, the answer (Entailment or Contradiction) is"*. Optionally, in zero-shot and few-shot prompting, the model is forced to explain the chain-of-thought reasoning by adding the phrase *"Let's think step by step"* (Kojima et al., 2022). Results show that the model is highly sensitive to the prompt patterns, with some of the prompts outperforming the previous state-of-the-art BERT-based model (Holzenberger and Durme, 2021). Also, the chain-of-thought imposition improves the performance but not systematically, whereas the worst prompt combination is, as expected, the zero-shot setting without the text of the statute(s). However, examining the chain-of-thought reasoning, it turns out that the model has a certain knowledge of the US tax code, yet imperfect, meaning that the model tends to refer to the wrong part of the statutes, even when the statute is provided in the prompt.

Yu et al. (2022a) test the performance of GPT-3 (text-davinci-002) on the Task 4 of COLIEE 2021, adopting the following approaches: zero-shot/few-shot setting

(with and without the chain-of-thought reasoning) and fine-tuning. In the zero-shot setting, the prompt pattern consists of instructions, the premise-hypothesis pairs from COLIEE dataset and the question phrase “*True or False?*”. In the few-shot setting, hypothesis-answer demonstration pairs of previously evaluated bar exam questions are given to the model in a 1-shot, 3-shot and 8-shot manner. Results show that the worst performing prompt in the zero-shot setting reaches the best model of COLIEE 2021, which is outperformed when a more instructive prompt is provided to the model. In the few-shot setting, all the k -shot combinations exceed the COLIEE 2021 winner, with 3-shot and 8-shot reaching the same performance and surpassing the 1-shot. Following Kojima et al. (2022), the authors consider another configuration adding the chain-of-thought reasoning imposition (“*Let’s think step by step*”) in the prompt and query again the model with its first response and the answer imposition (“*Therefore, the hypothesis is (True or False)*”). Moreover, they also fine-tune the model on the COLIEE 2021 training set in two completion settings: providing only the binary entailment answer or providing also the explanations. In the second setting, the explanations are either pseudo-explanations (i.e., for each premise the most relevant sentence for hypothesis) or explanations created by the model (i.e., for each hypothesis-premise-answer triplets, the model is prompted to generate explanations, with a specific prompt pattern). Results show that fine-tuning with pseudo-explanation surpasses fine-tuning with its GPT-3 explanation, but the use of explanations underperforms the fine-tuning without explanations. This suggests that allowing the model to reason independently might lead to better results. However, the best results are achieved using legal reasoning prompts (Burton, 2017) under the zero-shot setting. This approach aims to lead the model to “think like a lawyer”, encouraging it to consider, for example, the facts and circumstances that lead to the legal case, to find the rules governing the issue, to apply the rules to the fact and to determine the entailment task.

Nguyen (2023) introduces *LawGPT 1.0*, a GPT-3 based model that provides legal assistance in a conversational manner. LawGPT 1.0 is obtained by fine-tuning GPT-3 on a large legal corpus and is evaluated on several tasks, including the generation of legal documents, legal question-answering and provide legal advice. Results reveal its competitiveness w.r.t. existing virtual legal assistants.

Choi et al. (2023) focus on evaluating ChatGPT on four law school exams at the University of Minnesota courses of Constitutional Law, namely Federalism and Separation of Powers, Torts, Employee Benefits and Taxation. The exams consist of 95 multiple choice questions and 12 essay questions. The model is questioned with a uniform set of prompts (i.e., without adapting the prompts to a specific course or question). For multiple choice questions, three prompting approaches are evaluated: simple prompting, which requires to give only the answer, the chain-of-thought prompting (described in Blair-Stanek et al. (2023)) and the rank-order prompting, which consists in ranking the top three choices (like in II and Katz (2022)). ChatGPT was able to pass all the exams, although reaching a low average score; in general, it performed better on the essay questions than on the multiple choice questions. On the essays, in some cases it was as good as or even better than the average performance of human students, but in other cases it strongly failed, particularly when essay

questions regard specific cases or theories explained in the course. On the multiple-choice questions, it performed much worse on questions involving numbers and better in questions involving relatively uniform legal rules across the jurisdictions. Based on these results, the authors suggest some guidelines for constructing prompts to help the model generate better outputs, such as specifying at the end of the prompt the tone of the writing of the legal essay, or a word range.

3.4 Chapter Review

We investigated the use of TLMs for legal AI-based tasks. We conducted a detailed study of the approaches proposed in the literature in this area, and we categorized the problems that received a particular attention from the scientific community into three macro categories, namely legal search, legal document review, and legal outcome prediction; we noticed, however, that such categories are apparently interleaved and interrelated, and their definition is mainly intended for the sake of presentation. Retrieval, entailment and question answering have traditionally been the most frequently considered legal problems, also due to the role played by the COLIEE competition which has represented an important venue to foster the development of AI-based approaches in the legal domain, including those relying on TLMs. But equally important are a number of other tasks, ranging from named entity recognition to judgment prediction, from abstractive/extractive summarization to rhetorical role labeling. Yet, the various tasks are often interrelated, since a common approach to address them is to reformulate a legal-specific task in terms of a more general, machine-learning task, mainly focusing on classification and similarity. It is worth noticing, however, that legal text similarity and classification can be challenging for many reasons. For instance, there might be multiple classes to be assigned with a single case, as a case can span multiple areas of law. Moreover, the categorization of a case can vary depending on the particular court handling it. Yet, there is no universally agreed-upon set of areas of law clearly defined into a taxonomy (Mistica et al., 2021).

TLMs impact on legal tasks. By examining the literature, there is evidence of how TLMs have been able to push forward the state-of-the-art in a variety of tasks for the legal domain. Again, a representative case corresponds to the COLIEE competition, where we notice a growing and successful use of TLMs. Nonetheless, the competition as well as other venues have shown that non-TLM methods, including the traditional vectorial space models (e.g., TF-IDF or BM25), are still useful especially when combined with TLMs, particularly as a data pre-processing or filtering step. This generally serves a twofold purpose: to exploit both the lexical modeling of traditional techniques and the semantic knowledge of TLMs (e.g., Nguyen et al. (2021a)) and to perform an initial selection of possible candidates through the ranking score of traditional techniques and then re-ranking such results through TLMs (e.g., Althammer et al. (2021)).

It does not come to surprise that, being the first and most popular among the TLMs, BERT is widely involved in the existing approaches, often along with its early variants, such as RoBERTa, DistilBERT, DeBERTa. Yet, recent works tend to consider more advanced architectures, depending on the downstream task, which have shown to be very competitive, such as ELECTRA and XLM-RoBERTa as encoder-only models, XLNet and the GPT family of models as decoder-only models, T5 and BART as encoder-decoder, along with task-specific and long range models such as Longformer, BigBird, DPR and PEGASUS.

Moreover, a key aspect that determines the effectiveness of a TLM-based framework compared to other deep learning approaches is the type of training adopted to build the models. In the legal domain, while most studies have focused on task-adaptive fine-tuning, there has also been an increased interest in performing domain adaptation through further pre-training or pre-training from scratch a TLM, again starting from BERT models (e.g., Chalkidis et al. (2020b); Holzenberger et al. (2020); Zheng et al. (2021)). Song et al. (2022) carry out an empirical evaluation on the effectiveness of domain-specific pre-training for the legal domain, on a number of datasets and tasks including binary classification, multi-label classification, multiple choice question answering, summarization, and information retrieval. Results have shown that domain-specific pre-trained models can lead to 1%-5% higher performance than general domain pre-trained models, but on condition that the datasets involved are very close to the pre-training corpora, thus concerning the same legal sub-domain.

Regardless of the type and size of TLM and its training, the current trend for best-performing frameworks is to exploit techniques that are extrinsic to the particular language model, such as data augmentation, data enrichment, and ensemble strategies. Data augmentation is performed in several ways, for example through back translation (e.g., Rabelo et al. (2020)), focusing on the logical mismatches between articles and questions (e.g., Yoshioka et al. (2021)), negating the statements in the original sentences (e.g., Nguyen et al. (2021b)), or retrieving top- k irrelevant articles for a query according to a similarity score (e.g., Wehnert et al. (2021)). In the latest editions of COLIEE, several works use previously released data to obtain more training samples for the tasks at hand. Also, a few works have adopted ensemble strategies, which combine the results of independent systems to boost the overall performance; for example, the ensemble system of Nguyen et al. (2021a) reached the second position in the COLIEE-2021 Task 3, the one by Shao et al. (2020a) ranked first in the COLIEE-2020 Task 3, as well as the ensemble method proposed in Rosa et al. (2021) achieved the first position in the COLIEE-2021 Task 2. Data enrichment is typically obtained through the use of taxonomies ((Tziafas et al., 2021; Chalkidis et al., 2021a)) and thesauri ((Kim et al., 2021)).

Resource availability on the benchmarks. The need to devise the above mentioned strategies for improving the performance of a model also arises from the awareness that free-access legal resources are often limited or partially available. According to Song et al. (2022), one of the most significant challenges in legal NLP is the lack of large-scale high-quality datasets, due to the costs of the annotation

processes that require knowledge of the legal domain. In effect, although there exists a significant body of legal corpora that have been used for training and evaluation of TLMs for legal tasks, as we have analyzed in Section 2.4.2, the current landscape of legal benchmarks is still not comparable in size with the largest NLP and information retrieval datasets; in fact, apart from few exceptions corresponding to multi-task benchmarks (e.g., LexGLUE), most legal benchmarks and datasets have size of thousands or tens of thousands of documents. Moreover, we believe that attention should be paid to the maintenance of existing benchmarks, so as to keep them updated on important changes in the legislative status of a jurisdiction. Yet, there is room for building larger or new freely available benchmarks to evaluate legal tasks in new contexts (e.g., online social media companies and online trading companies involved in the Web3), as well as contexts that require controlling specific types of bias and/or specific ethical aspects concerning the diversity or disparity of legal norms within and across different countries.

Creating larger and more representative legal corpora depends on a number of aspects that include not only those strictly pertaining the selected evaluation goals for a particular task, but also factors relating to the language resources to be involved (e.g., open/free access availability, various forms of bias), the target audience with their different levels of expertise on the domain (e.g., lawyers, courts, law firms, citizens), as well as ethics-related and privacy-related aspects to be considered, to mention a few.

Resource availability on the model training. The limited availability of legal data for a particular task might prevent an adequate training of the model, thus making it unable to properly internalize the meaning of legal texts and to generalize the knowledge learned in sufficient detail for successfully dealing with unknown input. Indeed, it has been shown that it is not guaranteed that a model pre-trained on a legal corpus can significantly improve upon its corresponding general-domain pre-trained model fine-tuned on the target task, in all situations. Besides the previously mentioned Song et al. (2022) about the affinity between the data for the downstream task and the pre-training data, according to Wang et al. (2020a), the advantage of domain-specific pre-training is also related to the size of the data used for the downstream task: that is, the benefit gained through domain-adaptive pre-training is likely to be more significant when the downstream task appears to be more low-resource. This is also confirmed in Gururangan et al. (2020), where another form of pre-training is highlighted and named as task-adaptive pre-training, i.e., (unsupervised) training the model on a smaller but directly task-relevant corpus. This form of pre-training has shown to be not only competitive w.r.t. domain-adaptive pre-training but also beneficial when combined with it for improving the performance on the downstream task. However, the above findings were not proved for the legal domain, which certainly opens to opportunities for further research.

In addition, by also taking into account models' implementation technicalities, other difficulties can be identified. One is the imbalance between positive and negative examples, for which solutions include the filtering of possible candidates (e.g.,

Nguyen et al. (2021a)), oversampling methods (e.g., Shao et al. (2020a)), and the generation of artificial examples (e.g., Rosa et al. (2021)).

Resource availability on the jurisdiction language. Data availability issues are even more exacerbated when moving from a high-resource language, such as English, to low-resource languages. Several works have indeed been concerned with the legal domain for poor-resource language families, such as Romance, Slavic, Germanic, Uralic, but also oriental languages such as Chinese and Japanese. For such languages, the use of TLMs has enabled a breakthrough in many cases. For example, in Serras and Finger (2022), the TLM-based method improves upon statistical baselines; in Sun et al. (2021), it reaches leading positions; in Masala et al. (2021), all the baselines are consistently outperformed; in Papaloukas et al. (2021), TLMs prove to be superior to vector-space-model based learning techniques and RNN-based methods, in particular when they are domain-adapted. However, for low-resource languages, TLMs can perform worse than others (e.g., Lage-Freitas et al. (2022)), or large TLMs may suffer in performance in contrast to smaller counterparts (e.g., Douka et al. (2021)). In some cases, when benchmarks are very limited or even absent, an evaluation on the specific domain is not practicable (e.g., Gutiérrez-Fandiño et al. (2021a)). One common approach to overcome this issue is to develop multilingual and cross-lingual language models, under the assumption that low-resource languages can benefit from high-resource languages due to shared vocabulary and semantic relatedness aspect. Although there is evidence that the use of multilingual TLMs has helped to improve the state-of-the-art, there is certainly room for improvement (e.g., in Avram et al. (2021)).

Linguistic issues. Even more challenging is developing a model that can incorporate the many linguistic nuances and subtleties of legal documents already in the early stages of its construction, such as during the tokenization or masking processes. This appears to be critical especially for the case law data: in fact, while law statutes and articles are usually written in a language that should be as much understandable as possible to non-experts as well, legal cases and judgments can be particularly tricky to understand, even for humans. A related issue is the language mismatch between statutes, describing legal concerns in form of abstraction, and law cases, describing real facts; for instance, Savelka and Ashley (2022) focus on the interpretation of statutory terms, by investigating on how a particular term has been explained and applied in the past, thus allowing for the lawyers the construction of arguments which support or counter particular interpretations. This is a complex scenario, since there may be little lexical overlap between statutes and cases, thus making it more difficult to understand the legal text and extract key concepts from the case law to be retrieved in the statutes. In this regard, the quality of vocabulary can play an important role. The legal language is full of specific terms with a precise meaning, which are in many cases not included in general-purpose vocabularies. Few works address this issue by injecting legal words in the vocabulary (e.g., in Tagarelli and Simeri (2022); Simeri and Tagarelli (2023)), in order to prevent that the TLM tokenizer will break out out-of-vocabulary terms that are found in a target legal corpus. Clearly, an enriched vocabulary cannot be enough for a model to deeply learn syntactic, lexical,

and semantic patterns of the legal language. More specifically, modeling syntactic patterns as predicates, legal reasoning methodologies can provide a significant support. Legal reasoning is in fact one of the primary challenges identified in Zhong et al. (2020), since it should in principle adhere to well-defined rules.

In addition to all these aspects that are pertinent to the legal domain, there are also general language peculiarities that still pose challenges for TLMs, such as recognition of pronominal forms, anaphora-related issues, understanding various forms with negations, etc. An example is how to distinguish texts that look very similar to each other but actually have a logical or semantic mismatch (Yoshioka et al., 2021).

Document length. Compared to the earliest approaches, whereby the limitation on the maximum number of token conditioned the ability of TLMs in handling long documents, the current trend is to learn a more conservative, “lossless” model, which can elaborate a text at paragraph-level (e.g., considering embeddings of entire paragraphs rather than individual words, as in Shao et al. (2020c)), or that it can directly process longer documents, such as Longformer (e.g., Xiao et al. (2021)), or DPR (e.g., Althammer et al. (2021)). Another option can be to filter out noisy sentences with sentence extraction techniques ((Yu et al., 2022b)) or to perform a span-level approach ((Koreeda and Manning, 2021)), in which the document is divided into overlapping contexts containing spans. On the other hand, a different perspective is to incorporate a summarization step into the language understanding process with the aim of providing a concise meaningful version of the input text. Some of the discussed works indeed involve a summarization task (e.g., Alberts et al. (2020); Rossi and Kanoulas (2019); Kim et al. (2021)), others increase the maximum input length (e.g., Mamakas et al. (2022)) or apply hierarchical attention patterns (e.g., Chalkidis et al. (2022a)), but TLMs have surely the potential to be further developed for achieving enhanced performance.

However, to significantly improve their ability related to long range modeling, TLMs should cope with larger computational resources to process, model, store, and manipulate long passages of text, which also impact on advanced ability to reason coherently across long-range dependencies. Addressing these challenges will likely involve advancements in model architecture, training methodologies, memory management, and computational efficiency.

Structures at document-level and corpus-level. Another important aspect is to exploit explicit structure levels within and/or across legal documents. For instance, legal source citations within the text of articles or cases could be managed by replacing the citation with the cited article, as in Rabelo et al. (2020), although a much more promising approach would be to model and mine legal citation networks. As discussed in Locke and Zuccon (2022), citations play a very important role in legal decisions: since the decisions of judges must be in accordance with the doctrine of precedent, which establishes that lower courts must observe the decisions of higher courts, they usually explain the reasons for their decisions by proving the agreement with the previous decisions of the binding authority. Thus, modeling and learning

from feature-rich legal citation networks is desirable, however their potential in retrieval and entailment tasks has not been fully explored yet.

Bhattacharya et al. (2020) consider both the logical subdivision of statutes and their citation information to measure similarity of legal case documents. To this purpose, the authors introduce *Hier-SPCNet*, a precedent citation network augmented with the information about the hierarchy of the statutes (e.g., an act is typically structured in parts, chapters, topics, sections). A statute is therefore modeled as nodes of different types, each representative of the structural levels, and two types of edges, the one reflecting the hierarchy of the nodes and the other one indicating the citations between cases or between a case and a statute. Adding this structural information shows to lead to better document similarity estimation than competitors based on precedent citation network only. A combination of textual features and legal citation networks is also proposed in Paul et al. (2022a). Given an heterogeneous network with nodes representing statutes and legal fact descriptions and edges representing hierarchical structures of the statutes as well as citations between statutes and facts, the goal is to predict which statute is relevant to a newly introduced fact, i.e., if a link exists between the statute and the new document. The text of a statute or fact is considered as an attribute of the node. Two separate encoders play the role of encoding the attributes and the structural information given by the network, where attribute encoding is performed through hierarchical attention network (Yang et al., 2016), and structural encoding is obtained through metapath schemes (Fu et al., 2020). Results show that the exploitation of textual and graphical features can lead to significant improvements in performance compared to state-of-art competitors (BERT-based included) for the task. Also in the work of Louis et al. (2023), the structure of statutes is regarded as a key point for the success of neural models in statutory article retrieval tasks. A graph-augmented dense retrieval model is defined to exploit the topology of the statutes to expand the article information. The resulting knowledge-rich cross-article representations contributes significantly to the improvement of performance of the dense statute retriever. Wang et al. (2022) state that the document structure, and in particular the relations among the participants of a litigation, is essential to recognize and classify legal cases of different categories yet with similar topics. To this purpose, four relation graphs are defined, each modeling different types of relations between participants (i.e., plaintiffs and defendants), namely the relations between the participants and the matters of the dispute, the actions performed by the participants, the topics related to the participants, and the relations between facts and third parties. Each graph consists of nodes representing the document, the participants, and a set of nouns and verbs selected from the text and whose relations to participants express either matters, actions, keywords, or facts and third parties. The graphs are then aggregated in one graph by merging the common nodes, then it is provided in input to a graph attention network to get the document representation for the classification task. Results on real-world Chinese documents regarding twenty semantically-similar disputes show that the proposed model outperforms all the considered baselines, included TLMs like BERT, RoBERTa and Lawformer.

A related aspect is also a debate on the notions of legal similarity. In fact, while generally two documents would be similar if legal experts evaluate them as similar in their contents, Bhattacharya et al. (2020) point out that two legal cases should be regarded as similar if they jointly cite a common statute or precedents, and also if they cite different statutes or precedents which are structurally similar in their proposed Hier-SPCNet network.

Interpretability and knowledge injection. Two further challenges regard interpretability aspects and knowledge modeling (Zhong et al., 2020). In Section 3.2.7, we have discussed about the existence of a number of approaches recently developed to improve explainability and interpretability of TLM-based methods, which can broadly be categorized into post-hoc explanation methods and early explanation methods. Besides that, it is important to define specific legal requirements on the interpretability of machine learning models applied to private and public decision making, as discussed in Bibal et al. (2021). Also, Branting et al. (2021) highlight the trade-off between explanation quality and representation effort, stating that a key requirement for explainable systems is the ability to obtain useful and comprehensible predictions with low costs in terms of development, testing, and maintenance at scale. In this respect, knowledge injection can play a key role in improving not only the language understanding ability of TLMs but also their explainability and interpretability — recently, the term *augmented language models* has been introduced in Mialon et al. (2023) to also refer to those approaches that aim to enhance TLMs with reasoning skills and other external tools. Liu et al. (2021c) introduce the use of causal graphs to ensure explainability requirements for a task of similar charge disambiguation. The proposed model, named *GCI*, aims to capture explainable discriminative nuances among confusing charges through building a causal graph to detect keywords from the charges, where nodes are obtained from the charges and by clustering similar keywords, and edges (i.e., the causal relationships) are learned using a causal discovery algorithm. The graph is sampled in more graphs that are refined by estimating the strength of the relationships. *GCI* decides which charge is more suitable for a case by extracting keywords from fact description and mapping the case according to the graph. This approach has been integrated into LSTM models to assess its potential and improve interpretability, where in particular, causal strength constraints are included into the attention weights of the neural network. Knowledge modeling mainly refers to properly utilizing the legal knowledge, and in this regard, an increasing number of works carry out a domain-adaptation pre-training process of a TLM in the attempt of better internalizing legal knowledge in the model.

Unfortunately, as we previously mentioned, the ability of successfully modeling the legal language and capturing peculiar patterns to the domain is strongly related to the exploitation of large, possibly multiple, legal data sources. Actually, the amount of existing legal data is considerable but many resources are only available to large companies or court sections. For example, it is unlikely that the whole history of a section of a court can be recovered, which would be very useful for many tasks like court profiling or case recommendation to courts, or even to characterize the language

and creation style of arguments by lawyers, which can be supportive in the preparation of a lawyer's pleading. In Francesconi (2022), it is pointed out that the knowledge available in the Semantic Web is essential for the AI applied to legal domain, since it provides knowledge models for a top-down approaches (legal knowledge representation, legal reasoning, planning, explainability), and data for a bottom-up approach (argument mining, rule-based/case-based systems, legal information retrieval and discovery). Gan et al. (2021) suggest to inject legal knowledge into deep neural networks by including first-order logic rules, motivated by the fact that logic rules yield models with inductive inclination and, thus, can alleviate the dependency of deep neural networks on high amounts of training data, besides making them more interpretable because of the presence of the rules. A model for LJP on private loan cases is proposed as a co-attention network followed by a symbolic module, where the co-attention network exploits the relations between claims and fact descriptions and provide the probability distribution for judgments, and the symbolic module adapts the distribution according to logic rules to prevent outputs that violate the law. Moreover, non-differentiability of first-order rules is ensured by associating continuous real-values to the outputs of logic rules with mapping functions. The rules are then injected with a reward mechanism, i.e., the output of the co-attention network is increased, resp. decreased, if the facts in the text satisfy, resp. violate, the conditions. Results show that the gain in injecting legal knowledge into TLMs such as BERT and RoBERTa increases as the size of the training set decreases.

Ethical aspects. Another critical point in the use of AI-based technologies for the legal domain is related to *ethics*. Specific recommendations for the legal NLP community are provided in Tsarapatsanis and Aletras (2021) according to three ethical parameters: the importance of academic freedom, the diversity of ethical and legal norms and the threat of moralism. In Zhong et al. (2020), it is reminded that the purpose of AI is not to replace humans in legal matters but only to provide support in decision-making processes, but it is also pointed out that applying AI systems to law can inadvertently lead to ethical problems, particularly *bias* of different types such as gender and racial discrimination. Being aware of such issues and trying to alleviate them is an emergence for developing next-generation (legal-focused) TLMs. Henderson et al. (2022) highlight the difficulty in filtering toxic and private information in data used to train a model, since its removal could affect the meaning of the text; moreover, privacy expectations can be different based on the specific country. In this regard, filtering rules should be designed to reflect the standards developed by legal and administrative experts. Chalkidis et al. (2022c) measure the fairness of BERT-based models on three categories, namely demographics, regional and legal topic: for the first category, the goal is to evaluate if a model performs worse because it is biased by factors like gender, age, race, language, legal status, whereas for the other two categories, the goal is to assess if a model performs differently on cases associated with courts of specific regions, or in a specific field of law (e.g., it may perform better on criminal cases than on civil law cases). In this respect, some group disparity in performance is found as related to the defendant's state (Central European states versus the other European states), applicant's gender, language, legal

areas, and court's regions (e.g., Switzerland courts versus federation courts, and Beijing courts versus Sichuan courts); however, some group disparities can also be influenced by general factors based on the distribution of training data. Wang et al. (2021b) measure the judicial inconsistency related to attributes like gender, race, and region, as the average disagreement of the judgments (term penalty) given by LJP models, which are used as virtual judges. The judicial disagreement is defined as the standard deviation of virtual judges' results. Results on the CAIL data are produced according to region and gender: in the first case, seven virtual judges are trained on data belonging to seven provincial-level administrative regions, whereas in the second case, two virtual judges are trained, one for each defendant's gender. Gender and, especially, regional inconsistency are found in the legal system, with regional inconsistency varying over time; Moreover, judicial inconsistency seems to be negatively correlated with the severity of criminal charges.

Finally, a special remark should be made concerning all the hype of large generative language models after the launch of ChatGPT in November 2022. Despite limitations of ChatGPT and similar tools have soon been detected (e.g., possibility of writing plausible-sounding but incorrect, misinformed or nonsensical answers, sensitivity to tweaks to the input phrasing, verbosity, possibility of responding to harmful instructions or exhibiting biased behavior, etc.), we have witnessed a technological shift in the way we work: by using ChatGPT, people build websites and apps, write novels or technical papers, even pass college and university exams, from medical degree to law degree. This rise in "bad" actors abusing of generative contents has prompted to start developing countermeasures to debunk artificially generated text as well as to check data security requirements (e.g., on March 2023, Italian data-protection authority temporarily banned ChatGPT over concerns about breaching of existing EU privacy rules). This and much more opens to further challenges that are going to be a major point of interest for researchers involved in legal AI.

Chapter 4

Understanding and Summarizing Rights and Duties in European Constitutions through Transformer-based Language Models

A *constitution* is a fundamental document that serves as the supreme law of a country or an organization. It establishes the basic principles, rights, and rules that govern the functioning of the entity it applies to. When constitutions concern sovereign countries, they can be regarded as a proxy to reflect the unique historical, cultural and social contexts of their respective nations. Yet, amid this diversity, constitutions contain a backbone of themes of universal importance, such as core principles of democracy and human rights. In this respect, our research interest is in understanding and unveiling shared constitutional elements among countries that transcend national boundaries. We believe this holds profound practical and academic implications. First, it can enable policymakers and legal professionals to more deeply and quickly identify areas of agreement, reducing the time and effort required for decision-making and policy formulation. Second, in a federation of countries or a supranational political/economical union of states, it can promote legal alignment, efficiency and transparency among countries, making a fundamental step towards a more integrated and cooperative legal landscape.

In this context, we aim to (i) advance research on the understanding of similarities in constitutional texts concerning rights and duties and (ii) take a first step towards learning a summarization of European constitutional texts that share a number of rights-and-duties-relevant topics, hereinafter referred to as *RD-relevant topics*. Our rationale for this focus is twofold. First, Europe per se represents a highly relevant case in point, telling a rich and centuries-old history involving all European populations; moreover, there has been an increased interest in processing European law documents, thanks also to a publicly available access to vast resources of the European Union (Chalkidis et al., 2021a). Second, rights and duties are at the core of a society, shaping the relationship between individuals and the state, thus representing a thematic framework that establishes the boundaries of governmental authority, safeguards individual liberties, and ensures a just and equitable society.

RQ0: *How much European countries agree (or differ) w.r.t. specific topics within the realm of citizen rights and duties? how TLMs are able to discern the differences*

between topics? Can the European constitutions be regarded sufficiently similar to each other to justify a summarization task over the European constitutions?

RQ1: *Does a generative TLM currently have adequate knowledge of the European constitutions? If so, what are the RD-relevant topics that the generative TLM is acquainted with?*

RQ2: *How well is a generative TLM able to produce a coherent summary capturing the common constitutional law across all countries while keeping the main distinctive country-specific topical signals?*

To the best of our knowledge, we are the first to address the above research questions by leveraging TLMs. To answer **RQ0**, we employed TLMs-based methods for evaluating the degree of similarity among portions of European constitutions that discuss RD-relevant topics (Greco and Tagarelli, 2023). To answer **RQ1**, we assessed its awareness of the topics covered in European constitutional texts; without providing the model with the actual constitutional texts as context, we analyzed its capability to recognize and articulate on the main RD-relevant topics of the European constitutions (Greco and Tagarelli, 2024b). Finally, to answer **RQ2**, we defined a two-stage process of summarization, i.e., within- and across-country, for each of the RD-relevant topics, based on suitable choices of prompt engineering and incremental summary update. We assessed the model’s ability to produce meaningful summaries across the constitutions of different countries, thus capturing a transnational constitutional context for each of the topics concerning the European citizens’ rights and duties. Ultimately, these summaries might serve as the basis for developing a unified European constitutional block relevant to the citizens’ rights and duties (Greco and Tagarelli, 2024b).

4.1 Related Works

Researchers have made strides in utilizing AI techniques to analyze European human rights laws, notably focusing on the European Convention on Human Rights (ECHR). As discussed in Section 3, Chalkidis et al. (2019a) introduced the ECHR dataset, which contains allegations of violated provisions within the Convention. Each case includes a list of facts as well as a score awarded by the Convention to reflect the case’s significance in the evolution of case law. Furthermore, each instance is linked to a specific provision of the Convention that was violated. The authors tested a BERT-based model on the ECHR dataset to perform legal judgement prediction, case importance prediction and alleged violation prediction tasks. In the context of large-scale multi-label text classification of EU legislative documents, Chalkidis et al. (2019b) applied a number of methods, including BERT, to investigate which portions of the documents in EURLEX57K are informative. Moreover, LexGLUE (Chalkidis et al., 2022b) offers a platform for multi-label classification of allegations regarding violations of ECHR provisions. In particular, it includes multi-label classification of allegations regarding violations of the ECHR provisions. The dataset has been used

to test models on article violation prediction (Chalkidis et al., 2019a) and alleged violation prediction (Chalkidis et al., 2021c). Chalkidis et al. (2022b) evaluated a series of Transformer-based models, in their original versions or domain-adapted. Also, MultiEURLEX (Chalkidis et al., 2021a) provides annotated EU laws translated into multiple languages, enhancing accessibility and analysis across diverse linguistic contexts. The authors assessed monolingual BERT models pre-trained in Romance, Slavic, Germanic, and Uralic languages, as well as multilingual models (mT5 and XLM-RoBERTa) for cross-lingual legal text

Surprisingly, the current literature on the application of AI-based NLP to the processing and understanding of constitutions is quite limited. In Bayrak (2022), the authors carry out a comparative analysis between US constitution and a number of constitutions of four geographic regions (Africa, Asia, Europe, and Middle East), with the aim of detecting differences and similarities w.r.t. rights of citizens and the relationship between the major institutions of their governments; however, the study employs basic techniques used in text mining, discarding any reference to context-free as well as contextualized deep language models. Moreover, no insights into the European countries' constitutions are provided.

We aim to fill this gap in the literature by conducting a similarity analysis of the European countries' constitutions with a focus on the rights and duties of citizens, which are essential to define and protect the status of citizens as individuals and as members of the society. We believe our work can pave the way for further developments on AI-based solutions for NLP tasks involving the constitutions.

This is supported by the evidence that similarity search is broadly employed in the legal AI as an essential means to address more complex tasks, such as statutory article retrieval (Louis et al., 2023), legal case retrieval (Shao et al., 2020c), document review (Shaghaghian et al., 2020), legal judgement prediction (Aletas et al., 2016), summarization (Shukla et al., 2022), and many others. Moreover, the use of TLMs for performing similarity tasks on legal contexts is a prevailing trend in recent years. For instance, in Aumiller et al. (2021) such models are trained on a topic similarity task to predict the coherence among topics and to detect topical changes on legal texts. In Ostendorff et al. (2021), legal and general-purpose TLMs are compared for legal document recommendation addressed as a similarity task.

On the other hand, manual summarization methods would clearly struggle to cope with the vastness and complexity of constitutional texts, thus opening opportunities for an exploration of automated alternatives. Efforts to summarize European legal documents have seen significant progress with the introduction of specialized datasets and methodologies. EUR-Lex-Sum (Aumiller et al., 2022) stands out as a comprehensive multi- and cross-lingual dataset, comprising pairs of legal documents and summaries in various European languages. Similarly, Klaus et al. (2022) focused on extractive summarization of European regulatory documents, creating the EUR-LexSum corpus tailored to facilitate comprehension and compliance for non-jurists and organizations navigating complex regulatory landscapes. The authors tested BERT-based models addressing the summarization task as a binary classification of the sentence representations of the documents. Also, Zmiycharov et al. (2021) evaluated different extractive and abstractive summarization techniques

on EU legislation documents, with T5 achieving better results than other methods, including those designed for summarization tasks (e.g., PreSumm) as well as models fine-tuned on legal texts such as Legal-BERT. In contrast to existing research efforts and following the wave triggered by generative AI, we want to leverage the reasoning and summarization capabilities of generative TLMs from the decoder-only autoregressive family, a subset of the broader category of TLMs. These models have demonstrated remarkable proficiency in generating coherent natural language responses when provided with appropriate prompts, being able to interpret the instructions specified therein and execute them effectively (Liu et al., 2023a). Legal actors, such as law professionals and firms, have expressed interest in exploiting these models to enhance legal search and document review tasks, achieving scalability and cost-efficiency.

4.2 Data

Data collection and structure. We retrieved the texts of the European constitutions from the portal www.constituteproject.org (Elkins et al., 2014), which provides free and public access to constitutions from various countries around the world. The resources for the website come from the *Comparative Constitutions Project*. The documents in the Comparative Constitutions Project were originally labeled according to a number of topics they identify on the constitutions. The topics are organized on three levels, whereby the labels correspond to the most specific topics (i.e., the third-level topics). Hence, given a label, it is possible to identify the topic hierarchy to which it belongs. We use the same labeling system to structure our dataset for the similarity analysis at multiple levels of depth. Specifically, we narrowed down on the first-level topic “rights and duties” and on the European countries. We therefore conducted our similarity analysis of texts associated with third-level topics (hereinafter *micro-topics*, for short) and second-level topics (hereinafter *macro-topics*, for short) referring to rights and duties.

Table 4.1 shows the macro-topics and an overview of the corresponding micro-topics. Overall, European constitutions encompass 9 macro-topics and 111 micro-topics. Note that the same portion of a constitution can be assigned to multiple micro-topics, but also the same micro-topic can be associated to several, not necessarily contiguous parts of a constitution, in which case the texts are concatenated. In summary, the resulting dataset consists of text portions of constitutions, with each portion being hierarchically assigned the country name, the micro-topic and the macro-topic.

Data cleaning and chunking. Each text corresponding to a particular combination of country and micro-topic may span from one sentence to multiple parts of a constitution. Transformers generally have a maximum limit on the number of tokens they can process. To handle this, we divide the text into chunks so that the input does not exceed the 512 tokens limit imposed by BERT and BERT-based models.

Table 4.1: RD-relevant topics and their descriptions, grouped by the 9 macro-topics concerning Rights & Duties

TOPIC ID	TOPIC NAME	TOPIC ID	TOPIC NAME
(M1) Physical Integrity Rights			
<i>Country coverage: 28.2±11.5</i>			
slave	Prohibition of slavery	infoacc	Right to information
cruelty	Prohibition of cruel treatment	press	Freedom of press
cappun	Prohibition of capital punishment	petition	Right of petition
life	Right to life	privacy	Right to privacy
torture	Prohibition of torture	opinion	Freedom of opinion/thought/conscience
corppun	Prohibition of corporal punishment	(M7) Legal Procedural Rights	
(M2) Social Rights		<i>Country coverage: 22.3±11.2</i>	
<i>Country coverage: 23.9±10.1</i>		debtors	Rights of debtors
edcomp	Compulsory education	devlpers	Right to development of personality
env	Protection of environment	overthrw	Right to overthrow government
finsup3	State support for the disabled	prerel	Right to pre-trial release
safework	Right to safe work environment	juvenile	Privileges for juveniles in criminal process
provwork	Right to work	presinoc	Presumption of innocence in trials
conright	Protection of consumers	wolaw	Principle of no punishment without law
childwrk	Limits on employment of children	miranda	Protection from self-incrimination
standliv	Right to reasonable standard of living	jury	Jury trials required
edfree	Free education	excrim	Extradition procedure
finsup1	State support for the elderly	amparo	Right to amparo
finsup4	State support for children	fairtri	Right to fair trial
remuner	Right to just remuneration	falseimp	Protection from false imprisonment
shelter	Right to shelter	couns	Right to counsel
strike	Right to strike	dueproc	Guarantee of due process
health	Right to health care	habcorp	Protection from unjustified restraint
achighed	Access to higher education	appeal	Right to appeal judicial decisions
leisure	Right to rest and leisure	speedtri	Right to speedy trial
jointrde	Right to join trade unions	examwit	Right to examine evidence/witnesses
finsup2	State support for the unemployed	trilang	Trial in native language of accused
water	Right to water	expost	Protection from ex post facto laws
scifree	Right to enjoy the benefits of science	vicright	Protection of victim's rights
(M3) Economic Rights		doubjep	Prohibition of double jeopardy
<i>Country coverage: 29.6±8.7</i>		pubtri	Right to public trial
busines	Right to establish a business	evidence	Regulation of evidence collection
intprop	Provisions for intellectual property	(M8) Enforcement	
transfer	Right to transfer property	<i>Country coverage: 15.3±16.1</i>	
occupate	Right to choose occupation	hr	Human rights commission
proprght	Right to own property	amparo	Right to amparo
exprop	Protection from expropriation	ombuds	Ombudsman
freecomp	Right to competitive marketplace	inalright	Inalienable rights
(M4) Citizen Duties		(M9) Equality, Gender and Minority Rights	
<i>Country coverage: 17.0±11.5</i>		<i>Country coverage: 17.7±11.9</i>	
milserv	Duty to serve in the military	equalgr2	Equality regardless of nationality
taxes	Duty to pay taxes	equalgr13	Equality regardless of financial status
work	Duty to work	equalgr5	Equality regardless of language
(M5) General Duties		indpolgr1	Indigenous right to vote
<i>Country coverage: 16.0±11.3</i>		indpolgr6	Indigenous right to self governance
abide	Duty to obey the constitution	equalgr11	Equality regardless of creed or belief
binding	Binding effect of const rights	equalgr16	Equality regardless of parentage
(M6) Civil and Political Rights		equalgr8	Equality regardless of age
<i>Country coverage: 27.0±13.6</i>		equalgr3	Equality regardless of origin
assoc	Freedom of association	equalgr10	Equality regardless of skin color
express	Freedom of expression	rightres	Restrictions on rights of groups
voteun	Claim of universal suffrage	equalgr1	Equality regardless of gender
assem	Freedom of assembly	equalgr4	Equality regardless of race
fnfam	Right to found a family	socclas	Mentions of social class
freemove	Freedom of movement	selfdet	Right to self determination
libel	Right to protect one's reputation	equal	General guarantee of equality
nomil	Right to conscientious objection	equalgr15	Equality regardless of political party
civmar	Provision for civil marriage	equalgr12	Equality regardless of social status
freerel	Freedom of religion	asylum	Protection of stateless persons
arms	Right to bear arms	indpolgr2	Indigenous right to representation
citren	Right to renounce citizenship	cultright	Right to culture
dignity	Human dignity	matequal	Provision for matrimonial equality
marriage	Regulation of marriage	equalgr6	Equality regardless of religion
childpro	Rights of children	equalgr9	Equality for persons with disabilities
acfree	Right to academic freedom	indcit	Citizenship of indigenous groups
		equalgr7	Equality regardless of sexual orientation

The chunking process was carried out so as to keep as many sentences together as possible and ensuring not to break sentences and paragraphs. In any case, just a few instances ended up to exceed 512 tokens and the splitting consisted of mostly 2 or 3 chunks. When possible, a long portion was chunked on the basis of the constitution's structure (e.g., if the portion is the concatenation of parts from different articles, the subdivision was carried out keeping together the sentences from the same article). The chunks were associated with the same country, micro-topic and macro-topic. In general, an instance of the dataset corresponds to one country, but if the text was divided into chunks, there is one instance per chunk. The resulting dataset size is about 2580 instances. In contrast, more recent generative models can handle a considerable number of input tokens, which eliminates the need for the chunking step for the work discussed in Section 4.4.

Moreover, a step of *anonymization* was carried out to debias the lexical analysis from specific terms, while preserving the essential meaning in the sentences. In particular, we introduced generic identifiers to replace particular occurrences in the text, such as: names of persons (e.g., royals, secretaries who drew up documents) with “person”, inhabitants (e.g., Italians) with “European people”, countries (e.g., Italy) with “geo-political European entity”, locations (e.g., the Athos peninsula mentioned in the Greek constitution) with “location”, organizations (e.g., the United Nations mentioned in the Croatia constitution) with “organization”. Analogously, all legal references were replaced with a special token (*law_ref*).

4.3 Methods for Constitutional Text Understanding

4.3.1 Methodology

Lexical vs Semantic similarity across European countries. Firstly, we assess lexical and semantic differences among European countries on the parts of constitutions that share the same micro-topic. This is useful for understanding how much European countries agree (or differ) with respect to a specific theme of interest.

Multi-level semantic similarity analysis. A significant portion of our research efforts has been devoted to an extensive and multi-faceted examination of the similarity between texts extracted from European constitutions. The aim is to assess how well Transformer-based models can capture the closeness of texts that share the same topic, but more importantly how well they can discern differences between texts that deal with different topics (albeit discussing rights and duties) and whether and to what extent they are able to detect subtle nuances of texts from different but similar topics. In particular, we compare our selected Transformer-based models to conduct the following analysis tasks:

- **Topic similarity of texts having the same micro-topic:** the input corresponds to the instances concerning the same topic. The purpose is to compare the language models to assess the ability to detect similarities between countries.
- **Topic similarity of texts having the same macro-topic:** the input consists of all the instances that are associated with micro-topics belonging to the same macro-topic. The purpose is to compare the various language models to assess the ability to discern similar but not identical topics.
- **Topic similarity of texts across all the macro-topics:** the input consists of all portions of the countries dealing with micro-topics from all macro-topics. The purpose is to conduct an overall assessment on the generic topic of rights and duties. Specifically, given a micro-topic as a query, we evaluate the ability of the models to assign higher similarity scores to instances related to the query and, conversely, to assign lower scores with respect to instances related to other micro-topics. Ideally, the lower scores given to instances from the other micro-topics should still reflect whether or not they belong to the same macro-topic of the query, i.e., the scores given to instances from the same macro-topic should be higher than the scores given to instances related to other macro-topics.

4.3.2 Experimental Evaluation

Settings. Topic similarity is measured as cosine similarity throughout all the conducted tests. For the lexical analysis, the term-frequency inverse-document-frequency term relevance function (TF-IDF) is adopted to get the document embeddings. In this case, since the sparse vectorial space poses no limit to the input length, the text was not divided into chunks. On the other hand, lemmatization¹ and stemming² operations were performed as pre-processing steps.

For the semantic analysis, the text embeddings are obtained by each of our models, using the following implementations. `bert-base-uncased` is selected as the domain-general BERT. The selected legal-specific models are developed by Chalkidis et al. (2020b) and available on Huggingface,³ namely

- `legal-bert-base-uncased`, a BERT model pre-trained from scratch on legal corpora, which is referred to as `legal-bert-sc` in the original paper;
- `legal-bert-500k`, a BERT model further pre-trained on legal corpora, which is referred to as `legal-bert-fp` in the original paper;
- `bert-base-uncased-echr`, which is `legal-bert-fp` fine-tuned on ECHR cases;
- `bert-base-uncased-eurlex`, which is `legal-bert-fp` fine-tuned on EurLex.⁴

¹ <https://spacy.io/api/lemmatizer>

² <https://www.nltk.org/howto/stem.html>

³ <https://huggingface.co/nlpueb/legal-bert-base-uncased>

⁴ <https://eur-lex.europa.eu/>

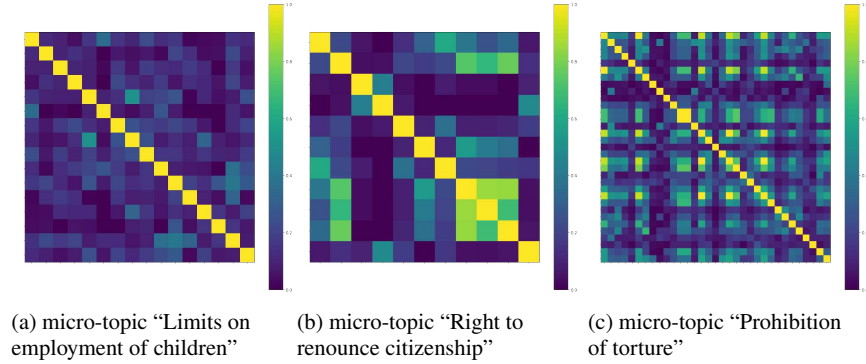


Fig. 4.1: Lexical similarity using TF-IDF

We did not consider other available legal models in Chalkidis et al. (2020b) since they are specific to U.S. law, while our study is focused on European constitutions. Using all the aforementioned models, we obtain the sentence embeddings applying mean pooling strategy on top of the contextualized token embeddings.

Finally, we consider sentence-Transformer models since they are highly applicable to similarity-related tasks. Currently, several models are available through the `sentence-transformers` library⁵ and they are ranked according to the quality of sentence embeddings, based on the performances achieved on different tasks and domains.⁶ Based on the ranking, we chose models that achieve good performance, but at the same time have a manageable size and a maximum length of 512 tokens. We therefore opted for the following models:

- `gtr-t5-large`,⁷ based on T5 Raffel et al. (2020) and fine-tuned for semantic search,
- `all-mpnet-base-v1`,⁸ based on MPNet model Song et al. (2020) and fine-tuned on different use-cases,
- `all-distilroberta-v1`,⁹ based on a distilled RoBERTa model Liu et al. (2019) and fine-tuned on different use-cases.

Lexical vs Semantic similarity. We first compute and analyze heatmaps of the similarity scores between TF-IDF embeddings produced for texts related to the same micro-topic. Each heatmap entry refers to an instance of the dataset dealing with the selected micro-topic and corresponds to a particular country. Note that, in all heatmaps shown throughout this research, lighter colors correspond to higher similarity scores.

⁵ <https://www.sbert.net/index.html>

⁶ https://www.sbert.net/docs/pretrained_models.html

⁷ <https://huggingface.co/sentence-transformers/gtr-t5-large>

⁸ <https://huggingface.co/sentence-transformers/all-mpnet-base-v1>

⁹ <https://huggingface.co/distilroberta-base>

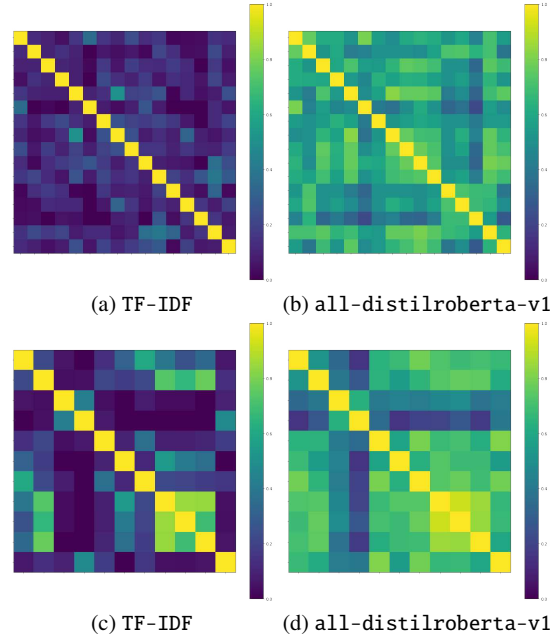


Fig. 4.2: Heatmaps corresponding to (a-b) micro-topic “*Limits on employment of children*” and (c-d) micro-topic “*Right to renounce citizenship*”.

Fig. 4.1 shows representative examples corresponding to selected micro-topics; for the sake of readability, we removed the country labels from the heatmaps. We notice different situations depending on the micro-topic, but in general, there is a low similarity among countries, as it is evident in, e.g., Fig. 4.1a. In some cases, a number of countries show strong lexical similarity, such as in Fig. 4.1b and 4.1c in which there are peaks of high scores and also peaks of maximum similarity.

In Table 4.2, we report some examples of the most similar texts according to the similarity scores w.r.t. the micro-topics discussed in Fig. 4.1. We can notice that for the micro-topics “*Right to renounce citizenship*” and “*Prohibition of torture*” the texts associated with the highest scores are structurally similar or almost identical.

Fig. 4.2 compares the heatmaps obtained by pairwise similarity of lexical-based embeddings and pairwise similarity of semantic-based embeddings (generated by all-distilroberta-v1), corresponding to the micro-topics “*Limits on employment of children*” and “*Right to renounce citizenship*”. It can be noticed that the lexical and the semantic heatmaps have markedly different scores, with the former having significantly lower scores than the latter. In particular, in Fig. 4.2(a-b), we can notice that all-distilroberta-v1 reveals some similarity matches between countries which are absent in TF-IDF. We show some examples in Table 4.3. In Fig. 4.2(c-d), the heatmaps have a similar shape but, again, the semantic model associates significantly higher scores than the lexical model. By comparing the two models, it

Table 4.2: TF-IDF most similar examples w.r.t. the micro-topics (a) “Limits on employment of children”, (b) “Right to renounce citizenship”, and (c) “Prohibition of torture”. Given the query text (highlighted in bold), the list of the most similar candidates in descending order of score is shown.

Micro-topic	Similar examples
Limits on employment of children	Italy 1947 (rev. 2020): “ Working women are entitled to equal rights and, for comparable jobs, equal pay as men. Working conditions must allow women to fulfil their essential role in the family and ensure appropriate protection for the mother and child. The law establishes the minimum age for paid labour. The Republic protects the work of minors by means of special provisions and guarantees them the right to equal pay for equal work. ” Malta 1964 (rev. 2016): “ <i>Minimum age for paid labour. The minimum age for paid labour shall be prescribed by law.</i> ”
Right to renounce citizenship	Russian Federation 1993 (rev. 2014): “ A citizen of the Russian Federation may not be deprived of his (her) citizenship or of the right to change it. ” Serbia 2006: “ <i>A citizen of the Republic of Serbia may not be expelled or deprived of citizenship or the right to change it.</i> ” Slovakia 1992 (rev. 2017): “ <i>No one may be deprived of the citizenship of the Slovak Republic against his will.</i> ”
Prohibition of torture	Albania 1998 (rev. 2016): “ No one may be subjected to torture, cruel, inhuman or degrading punishment or treatment. ” Czech Republic 1993 (rev. 2013): “ <i>No one may be subjected to torture or to or to cruel, inhuman, or degrading treatment or punishment.</i> ” Kosovo 2008 (rev. 2016): “ <i>No one shall be subject to torture, cruel, inhuman or degrading treatment or punishment.</i> ” Moldova (Republic of) 1994 (rev. 2016): “ <i>No one shall be subject to torture or other cruel, inhuman or degrading punishments or treatments.</i> ” Portugal 1976 (rev. 2005): “ <i>No one shall be subjected to torture or to cruel, degrading or inhuman treatment or punishment.</i> ” Montenegro 2007 (rev. 2013): “ <i>No one can be subjected to torture or inhuman or inhuman or degrading treatment.</i> ”

can be inferred that strong similarities are captured by both, but the semantic model is able to detect more adequately the common focus of the texts. High scores are, indeed, expected since the texts discuss the same micro-topic. By contrast, the lexical model often has very low scores even on texts of the same micro-topic, consequently it also fails to differentiate texts of the same micro-topic from texts of different micro-topics.

Results at micro-topic level. As the lexical model revealed to be unable to adequately detect similarities between texts about the same micro-topic, hereinafter we

Table 4.3: Example pairs for the micro-topic “Limits on employment of children” that are regarded as similar by all-distilroberta-v1 but dissimilar based on TF-IDF embeddings

Moldova (Republic of) 1994 (rev. 2016) “All employees shall have the right to social protection of labour. The protecting measures shall bear upon the labour safety and hygiene, working conditions for women and young people, the introduction of a minimum wage per economy, week-ends and annual paid leave, as well as difficult working conditions and other specific situations. The exploitation of minors and their involvement in activities, which might be injurious to their health, moral conduct, or endanger their life or proper development shall be forbidden.”
Montenegro 2007 (rev. 2013): “Youth, women and the disabled shall enjoy special protection at work. A child shall be guaranteed special protection from psychological, physical, economic and any other exploitation or abuse.”
Albania 1998 (rev. 2016): “Every child has the right to be protected from violence, ill treatment, exploitation and use for work, especially under the minimum age for work, which could damage their health and morals or endanger their life or normal development.
Croatia 1991 (rev. 2013): “Children may not be employed before reaching the legally determined age, nor may they be forced or allowed to do work which is harmful to their health or morality.”

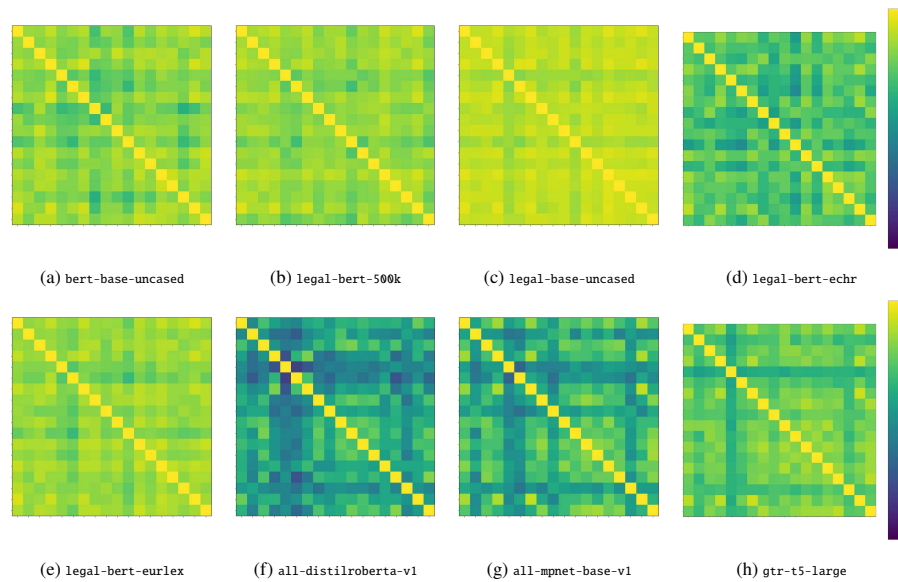


Fig. 4.3: Micro-topic “Duty to pay taxes”.

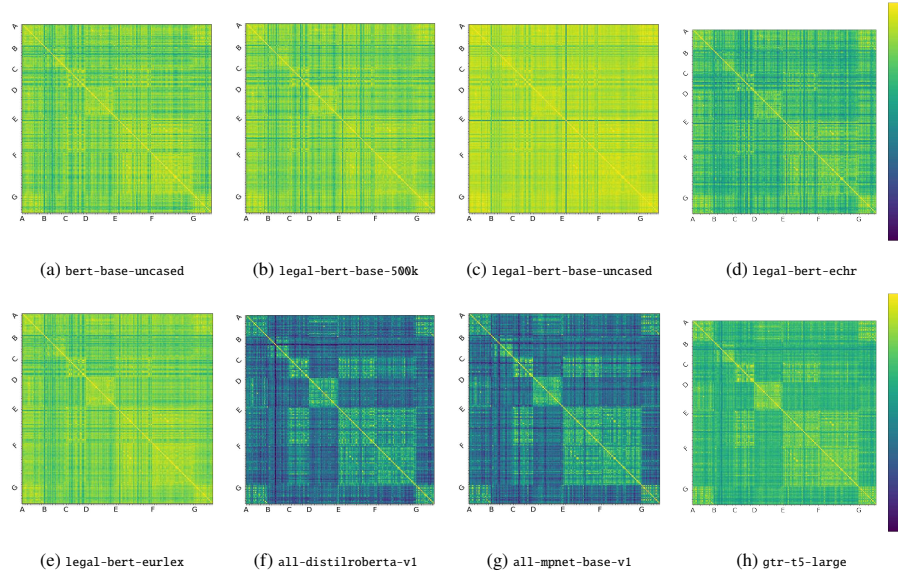


Fig. 4.4: Macro-topic “Economic Rights”: A - *Right to establish a business*; B - *Provisions for intellectual property*; C - *Right to transfer property*; D - *Right to choose occupation*; E - *Right to own property*; F - *Protection from expropriation*; G - *Right to competitive marketplace*

focus on semantic similarity only, in the attempt of identifying the best model in capturing commonalities and differences between micro-topics.

A first step is at the micro-topic level, that is, evaluating which model is able to associate the highest scores for texts belonging to the same micro-topic. Fig. 4.3 shows heatmaps of all the considered Transformer-based models (BERT, Legal-BERT models, and Sentence-Transformers) for the micro-topic “Duty to pay taxes”. In general, they all provide high scores, but legal-bert-base-uncased exhibits extremely high similarities. Among all, legal-bert-echr and the Sentence-Transformers provide some slightly lower scores. The models exhibit a similar pattern across all micro-topics; In Table 4.4 we provide an overview of the models’ behavior. More precisely, for each macro-topic, we provide the average values of the mean, minimum, maximum, and median similarity scores calculated on the same micro-topics. For instance, for the macro-topic “Physical Integrity Rights”, bert-base-uncased provides, on average, a mean similarity score of 0.834 on texts related to the same micro-topic, while all-distilroberta-v1 provides an average maximum similarity score of 0.878 for the macro-topic “Civil and Political Rights”. We observe that legal models have the highest values and the smallest range between the minimum and maximum scores for most macro-topics. This may be due to an over-specialization knowledge of legal domain, leading to high scores for texts having strong legal concepts in common, or conversely to a poor ability to recognize differences.

Table 4.4: Statistics on similarity scores within the same micro-topic, for each macro-topic.

Model	Mean	Min	Max	Median	Mean	Min	Max	Median	Mean	Min	Max	Median
	M1				M2				M3			
bert-base-uncased	0.834	0.670	0.976	0.842	0.850	0.720	0.944	0.859	0.836	0.670	0.968	0.845
legal-bert-uncased	0.876	0.773	0.976	0.884	0.912	0.842	0.967	0.917	0.899	0.798	0.983	0.907
legal-bert-500k	0.820	0.644	0.965	0.828	0.865	0.751	0.951	0.872	0.841	0.677	0.971	0.853
legal-bert-eurlex	0.829	0.668	0.969	0.833	0.854	0.745	0.942	0.859	0.836	0.693	0.965	0.844
legal-bert-echr	0.747	0.538	0.946	0.750	0.769	0.601	0.906	0.779	0.744	0.516	0.933	0.751
all-distilroberta-v1	0.612	0.327	0.899	0.606	0.569	0.283	0.820	0.568	0.564	0.226	0.883	0.570
all-mpnet-base-v1	0.629	0.336	0.923	0.625	0.612	0.337	0.860	0.614	0.590	0.239	0.917	0.592
gtr-t5-large	0.787	0.654	0.933	0.781	0.764	0.625	0.903	0.763	0.760	0.583	0.939	0.757
	M4				M5				M6			
bert-base-uncased	0.808	0.658	0.926	0.822	0.855	0.736	0.945	0.857	0.853	0.724	0.954	0.861
legal-bert-uncased	0.899	0.835	0.954	0.903	0.903	0.824	0.958	0.911	0.906	0.835	0.970	0.915
legal-bert-500k	0.831	0.708	0.926	0.842	0.851	0.755	0.933	0.858	0.856	0.738	0.952	0.866
legal-bert-eurlex	0.825	0.713	0.926	0.829	0.845	0.762	0.916	0.850	0.853	0.743	0.949	0.860
legal-bert-echr	0.704	0.496	0.874	0.717	0.747	0.594	0.892	0.755	0.763	0.582	0.914	0.771
all-distilroberta-v1	0.486	0.188	0.803	0.480	0.569	0.327	0.815	0.566	0.609	0.331	0.878	0.614
all-mpnet-base-v1	0.542	0.279	0.740	0.537	0.590	0.407	0.749	0.581	0.633	0.361	0.893	0.637
gtr-t5-large	0.740	0.615	0.852	0.743	0.751	0.637	0.874	0.746	0.781	0.652	0.914	0.779
	M7				M8				M9			
bert-base-uncased	0.857	0.739	0.956	0.866	0.856	0.685	0.979	0.865	0.873	0.755	0.964	0.880
legal-bert-uncased	0.891	0.802	0.969	0.898	0.908	0.803	0.985	0.916	0.919	0.844	0.974	0.924
legal-bert-500k	0.834	0.695	0.948	0.845	0.864	0.732	0.978	0.871	0.874	0.764	0.960	0.879
legal-bert-eurlex	0.850	0.736	0.949	0.857	0.849	0.704	0.971	0.856	0.875	0.766	0.961	0.879
legal-bert-echr	0.744	0.552	0.914	0.751	0.785	0.581	0.943	0.797	0.808	0.654	0.935	0.815
all-distilroberta-v1	0.556	0.277	0.837	0.554	0.569	0.238	0.869	0.584	0.595	0.324	0.860	0.595
all-mpnet-base-v1	0.594	0.311	0.866	0.602	0.565	0.269	0.854	0.566	0.602	0.329	0.877	0.594
gtr-t5-large	0.758	0.619	0.899	0.759	0.756	0.573	0.926	0.755	0.784	0.659	0.918	0.780

Results at macro-topic level. Considering the aforementioned behavior of the models on each macro-topic, we explore whether the high scores of legal models can be ascribed either to a robust understanding of the micro-topic or to a potential inability to discern subtle semantic distinctions.

For this purpose, we assess whether different micro-topics belonging to the same macro-topic are indistinguishable, that is, whether the models can distinguish related but not identical topics. In Fig. 4.4, we show the heatmaps corresponding to the various micro-topics belonging to the macro-topic “*Economic Rights*”. Note that each heatmap reports both the similarities between texts belonging to the same micro-topic (which are concentrated on the main diagonal) and the similarities between texts belonging to different micro-topics. For the sake of readability, we replace the names of the countries with letters corresponding to their respective micro-topics; when an explicit label is missing, it is inferred that the entry of the heatmap is associated to the preceding label.

It can be noticed that legal models (and even bert-base-uncased) are unable to perceive different degrees of similarity, which should be higher for texts of the same micro-topic and lower between texts of different micro-topics. On the contrary, the Sentence-Transformers (particularly all-distilroberta-v1 and all-mpnet-base-v1) are able to distinguish the different micro-topics much more

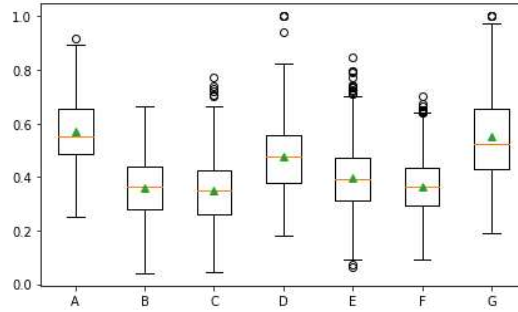


Fig. 4.5: Boxplots on the micro-topic “*Right to establish a business*” vs the other micro-topics of the same macro-topic (“*Economic Rights*”) using `all-mpnet-base-v1`. A - *Right to establish a business*; B - *Provisions for intellectual property*; C - *Right to transfer property*; D - *Right to choose occupation*; E - *Right to own property*; F - *Protection from expropriation*; G - *Right to competitive marketplace*.

clearly. However, there are micro-topics that are nearly indistinguishable even for the Sentence-Transformers. Examining these challenging micro-topics, we observe that they often involve remarkably similar concepts. For example, in Fig. 4.4g, there is a clear difficulty in distinguishing the micro-topics C, E, and F, which, however, have in common the aspect of addressing matters pertaining to property rights. Similarly, the micro-topic A and G are practically indiscernible even for the Sentence-Transformers.

The above is also evident in the boxplots in Fig. 4.5. In this case as well, the micro-topics A and G encompass a closely related concept, namely aspects related to business. Once again, the behavior of the models is consistent across all macro-topics, with `all-distilroberta-v1` and `all-mpnet-base-v1` showing the best results. To provide an example, Fig. 4.6 shows the boxplots of `all-mpnet-base-v1` and `legal-bert-uncased` for the micro-topic “*Prohibition of slavery*” against all the micro-topics of its macro-topic (“*Physical Integrity Rights*”). It can be observed that, in the case of `all-mpnet-base-v1`, the boxplot corresponding to the texts of the micro-topic under examination (the first one from the left) has a higher mean compared to the other boxplots, which represent the similarity scores of texts from the “*Prohibition of slavery*” topic compared to texts from other micro-topics of its macro-topic. On the contrary, the boxplots of `legal-bert-uncased` demonstrate that the model does not perceive substantial differences between texts with different micro-topics compared to texts with the same micro-topic.

Overall results on Rights and Duties. Another crucial aspect concerns the models’ capability to distinguish between micro-topics that belong to the same macro-topic, as opposed to micro-topics from different macro-topics. Given a micro-topic

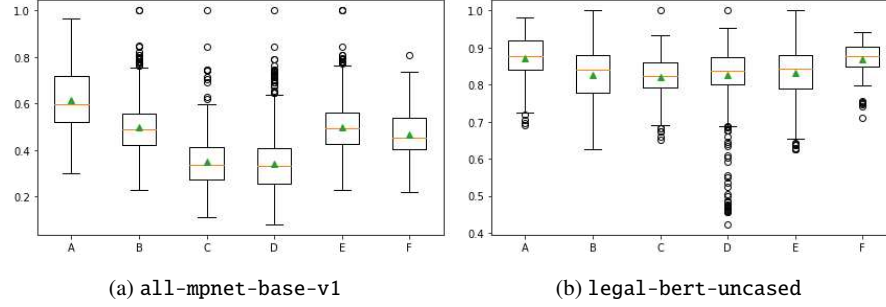


Fig. 4.6: Boxplots on the micro-topic “*Prohibition of slavery*” vs the other micro-topics of the same macro-topic “*Physical Integrity Rights*”. A - *Prohibition of slavery*; B - *Prohibition of cruel treatment*; C - *Prohibition of capital punishment*; D - *Right to life*; E - *Prohibition of torture*; F - *Prohibition of corporal punishment*.

as a query, we shall compute similarity scores for all micro-topics. This can be seen as an overall evaluation of the Rights and Duties topic.

Our analysis can be grouped into three categories: (1) similarity between texts of the same micro-topic as the query, (2) similarity between texts of different micro-topics but within the same macro-topic as the one associated to the query micro-topic, and (3) similarity between texts with both a different micro-topic and macro-topic w.r.t. the query. The expected behavior for the models is to assign high similarity scores for the first category and low scores for the other two categories, but the scores associated with the second category should be higher compared to the scores associated with the third category.

Table 4.5 summarizes the analysis conducted on all micro-topics, providing an overall mean value across the three categories. For instance, `bert-base-uncased` obtains, on average, a mean similarity score of 0.853 on texts related to the same micro-topic, a mean similarity score of 0.799 on texts related to different micro-topics but having the same macro-topic, and a mean similarity score of 0.762 on texts related to different micro-topics and different macro-topics. The difference between the values of the first and second categories (column $|\mu - M|$), is of 0.053, for the second and the third categories (column $|M - M'|$) is of 0.037, and for the first and the third categories (column $|\mu - M'|$) is of 0.091. Once again, it can be observed that the generic `bert-base-uncased` and the legal models are the least effective in distinguishing between different micro-topics. Even `gtr-t5-large` demonstrates limitations in this regard, particularly in distinguishing between the second and third category. Among all the models, `all-mpnet-v1` achieves the most favorable results, demonstrating the largest disparity in all scenarios (columns $|\mu - M|$, $|M - M'|$ and $|\mu - M'|$), followed by `all-distil-roberta`.

As an illustrative example, we show the boxplots of all models for the micro-topic “*Prohibition of slavery*” in Fig. 4.7. The differences across the boxplots are more

Table 4.5: Overall evaluation of the models. The *micro-topic* column represents the average (μ) over the mean values of similarity scores between texts belonging to the same micro-topic. The *macro-topic* column represents the average (M) over the mean values of similarity scores between texts belonging to the same macro-topic but not the same micro-topic. The *other macro-topics* column represents the average (M') over the mean values of similarity scores between texts belonging to different macro-topics and different micro-topics. The remaining columns correspond to the difference between the values of the columns micro-topic and macro-topic ($|\mu - M|$), the difference between the values of the columns macro-topic and other macro-topics ($|M - M'|$), and the difference between the values of the columns micro-topic and other macro-topics ($|\mu - M'|$).

model	micro-topic (μ)	macro-topic (M)	other macro- topics (M')	$ \mu - M $	$ M - M' $	$ \mu - M' $
bert-base-uncased	0.853	0.799	0.762	0.053	0.037	0.091
legal-bert-uncased	0.905	0.871	0.852	0.033	0.019	0.052
legal-bert-500k	0.853	0.798	0.766	0.055	0.031	0.087
legal-bert-eurlex	0.853	0.796	0.756	0.057	0.039	0.096
legal-bert-echr	0.765	0.687	0.643	0.078	0.044	0.122
all-distilroberta-v1	0.576	0.399	0.313	0.177	0.086	0.263
all-mpnet-base-v1	0.603	0.414	0.323	0.189	0.090	0.279
gtr-t5-large	0.769	0.682	0.643	0.086	0.039	0.126

evident with `all-mpnet-base-v1`. Indeed, the first boxplot (related to the first category) exhibits the highest and most uniform values, the second boxplot (related to the second category) is sufficiently lower than the first one but higher than the others (related to the third category).

4.4 Methods for Constitutional Text Knowledge and Summarization

4.4.1 Methodology

We introduce a novel application of abstractive summarization on a multi-source collection of texts from the constitutions of European countries dealing with RD-relevant topics. This is a challenging testbed, given not only the intricacies and subtleties of the legal sublanguage but also because of the intrinsic complexity of dealing with foundational laws of different countries which are thematically related yet diverse in many different aspects concerning their history, culture, ethnicity, religion, and more.

As previously mentioned, our model choice to tackle the stated RQs corresponds to `gpt-3.5-turbo`¹⁰ (hereinafter referred to as GPT-3.5T), which was accessed through OpenAI API based on authenticated requests. It should be noted that, being

¹⁰ <https://platform.openai.com/docs/models/gpt-3-5>

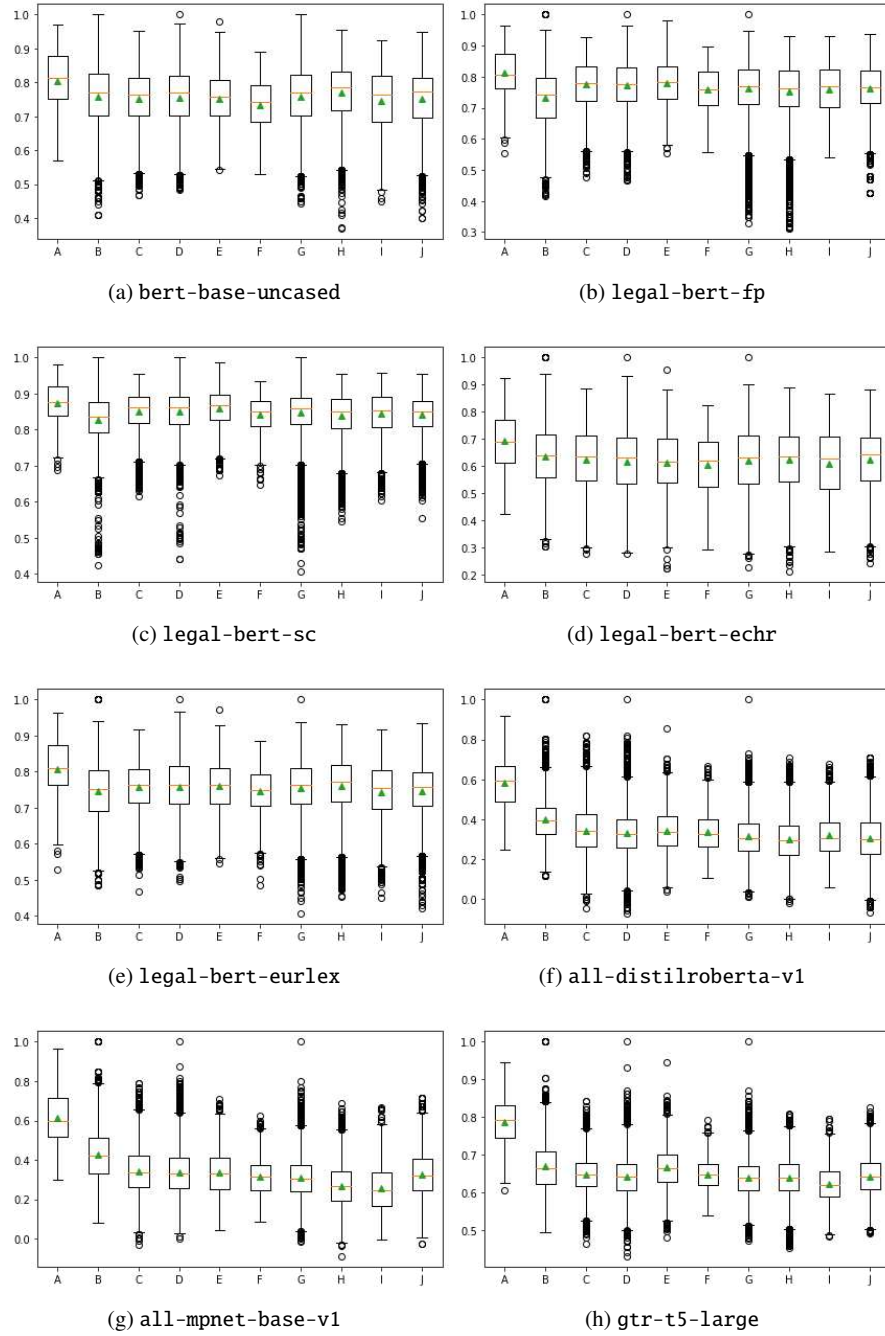


Fig. 4.7: Boxplots on the micro-topic “*Prohibition of slavery*” vs the parts of constitutions referring to its macro-topic (“*Physical Integrity Rights*”) and all the others macro-topics. A - *Prohibition of slavery*; B - *Physical Integrity Rights*; C - *Social Rights*; D - *Economic Rights*; E - *Citizen Duties*; F - *General Duties*; G - *Civil and Political Rights*; H - *Legal Procedural Rights*; I - *Enforcement*; J - *Minority Rights*.

a well-established and widely used TLM, we adopted GPT-3.5T as a case in point to demonstrate the capabilities and limitations of current TLMs to cope with the tasks corresponding to our RQs. This choice allows us to provide a comprehensive analysis ensuring that our findings are relevant and accessible to a broad audience. Additionally, GPT-3.5T offers a balanced trade-off between performance, computational resources, and cost-effectiveness (especially compared to GPT-4) making it an ideal candidate for the scope of our study.

To address our first research question (**RQ1**), we delved into the generative model's awareness of the themes woven into the European constitutions. For this purpose, we asked GPT-3.5T to describe as many topics as possible about each particular macro-topic (selected from our dataset's topic-hierarchy), according to its knowledge of the European constitutions. In this stage, no portion of an European country's constitution was introduced into the prompts to GPT-3.5T. Specifically, we provided the model with the following prompt for each macro-topic M :

Prompt

Describe the topic $\langle M \rangle$ based on what is written in the constitutions of European countries. Provide a list of as many arguments as possible that are related to this topic, with the associated description. The list must be structured as a number, followed by the argument name, followed by the description.

Given a macro-topic M to ask GPT-3.5T about, each item in the topic list returned by the model was compared to the actual subset of micro-topics for M (i.e., T_M), in order to find the best matching in terms of semantic similarity. To this aim, we computed the cosine similarity between the sentence embedding of a model topic and the sentence embedding of an actual micro-topic $t \in T_M$, which were produced by all-mpnet-base-v1,¹¹ a highly effective S-BERT model (Reimers and Gurevych, 2019).

To address **RQ2**, we devised a two-stage summarization task, which was applied to each macro-topic M . In Stage 1, we focused on each micro-topic $t \in T_M$ at a time, to select all the portions of a country's constitution that were originally labeled with t , with the purpose of generating a topic-specific summary for that country. To this aim, we asked GPT-3.5T to first extract keywords from the text (i.e., set of constitution portions for that country) then to build a summary upon those keywords. Given a country C and a micro-topic t , for Stage 1 we provided the model with the following prompt:

Prompt

Given the text $\text{TEXT}(C, t)$, detect keywords in the text and generate a compressed version of it based on the keywords. All keywords must appear in the compressed version of the text. Show in output the keywords and the compressed text in the following format:

Keywords: $\langle \text{list-of-keywords} \rangle$
Compressed Text: $\langle \text{compressed-text} \rangle$

¹¹ https://www.sbert.net/docs/pretrained_models.html

where `<list-of-keywords>` is the list of identified keywords and `<compressed-text>` is the generated compressed text.

In Stage 2, we requested the model to create a unified constitution summary for each particular $t \in T_M$ based on the country-level summaries produced in Stage 1. This is clearly more challenging than the previous stage because of token length limitations of TLMs. In particular, despite the GPT-3.5T’s context window of 16,385 tokens allows us to accommodate the length of the constitution portions pertinent to any micro-topic, a “full-batch” approach to build a unified summary over all countries might not be feasible in general without losing information. To overcome this issue, we devised an incremental approach, starting with the summary of largest constitution-portion and requesting the model one step at a time to incorporate the next country-level summary, in decreasing order of size of the original country portion-set. More specifically, we provided GPT-3.5T with the current summary and the next country-level summary as a concatenated text and asked the model to produce a new summary by avoiding redundant information while retaining the keywords identified in Stage 1:

Prompt

Given the text:
 Current-TopicLevel-Summary
 Current-CountryLevel-Summary
 Detect redundant information into the text and reformulate it in order to have all information without redundancies. Note that you must keep the following keywords:
 list-of-keywords-current-topiclevel-summary, list-of-keywords-countrylevel-summary
 Return only the final complete text, in the following format:
 Final Text: `<final-text>`
 where `<final-text>` is the generated text.

Model settings. Since we strive to maintain each final micro-topic-level summary as coherent as possible with the corresponding original texts in the constitutions, we set the GPT-3.5T *temperature* to 0. Also, we kept the *top-p* and *penalty* hyperparameters at their default values. Our choice was motivated to prevent disruptive effects on the delicate balance between diversity and coherence in generation, thus ultimately avoiding adding further complexity and making reproducibility easier to fulfill. Rather, we assigned a *role* to the model, instructing it to act as “*summarization expert*”: by means of this directive, we actually enhanced the model’s ability to effectively extract key aspects from the text and construct a summary that ensures the inclusion of essential information.

4.4.2 Experimental Evaluation

We assessed the model’s ability to produce meaningful summaries across the constitutions of different countries, thus capturing a transnational constitutional context for each of the topics concerning the European citizens’ rights and duties. Ultimately, these summaries might serve as the basis for developing a unified European constitutional block relevant to the citizens’ rights and duties.

Evaluation criteria. We resorted to a number of statistical criteria that allow us to assess the quality of a generated summary, according to (i) its input source (original text), (ii) its reference summary, and (iii) its intrinsic characteristics. To this purpose, we devised two groups of criteria, namely *non-TLM-based* and *TLM-based* criteria.

The group of non-TLM-based includes the following methods.

- **Compression ratio (CR)**, which measures the character-level compression percentage of the generated summary compared either to the original text or a reference summary.
- **Rouge metrics** *Rouge-1*, *Rouge-2*, *Rouge-L*, and *Rouge-LSum* (Lin, 2004b), which are the most widely used metrics for evaluating the quality of automatic summaries generated by NLP models, either to the original text or a reference summary. In this work, we report the F-measure score for each metric. Moreover, note that to balance precision and recall against the compression ratio, the β value of the F-measure is set as the reciprocal of the compression ratio, expressed in absolute terms.
- **Novelty (N)**, which is defined as the percentage of unique words in the summary that do not appear in the original text, or alternatively in a reference summary. This indicates the model’s tendency to introduce new words. In particular, given the set of words (excluding stopwords) of the input source (alternatively, reference summary) I and the generated summary S , the novelty is calculated as $N(I, S) = |S \setminus I| / |S|$.
- **Jaccard (J)** and **Dice (D)** coefficients to compare the overlap between the set of words (excluding stopwords) of the original source (alternatively, of the reference summary) I and the set of words of a generated summary S ; formally, they are calculated respectively as $J(I, S) = |S \cap I| / |S \cup I|$ and $D(I, S) = 2|S \cap I| / (|S| + |I|)$.
- Cosine-based *lexical similarity* between the **TF-IDF** vectors of the original source (alternatively, the reference summary) and the generated summary, respectively, obtained after stopword removal and stemming.

Our selected group of the TLM-based criteria is comprised as follows:

- Cosine-based *semantic similarity* between the **S-BERT** embeddings of the original source (alternatively, the reference summary) and the generated summary, respectively. The embeddings are obtained from the 512-token-sized **all-mpnet-base-v1** TLM, since it demonstrates a good trade-off between the quality of sentence embeddings and model size. Note that the limit on manageable tokens (512 tokens, the current limit for S-BERT models) led to the truncation

of text that exceeded this threshold (this occurred for 99 original texts and 74 generated summaries, out of 114 topics);

- **BLANC** (Vasilyev et al., 2020) scores which involve BERT to estimate the quality of a summary without a reference summary. More specifically, it measures the efficacy of a summary in supporting a TLM when performing the *cloze* task (masked token prediction) on a document. Such measure is proposed in two versions (Vasilyev et al., 2020): **BLANC-Help** (*BH*), which concatenates the summary to each sentence within the document, in order for the summary to be of assistance during inference, and **BLANC-Tune** (*BT*), which employs the summary for the fine-tuning phase, before analyzing the document. In both cases, the measure is calculated in terms of token unmasking accuracy, i.e. the difference in accuracy with and without the use of the summary. Note that BLANC has a limit on manageable tokens (512 tokens, as it employs BERT), which led to the truncation of summaries that exceeded this threshold (this occurred for 99 original text and 74 generated summaries, out of 114 topics).
- A **qualitative evaluation** by asking GPT-3.5T to rate a summary on a scale 1(worst)-5(best) according to five criteria (Adams et al., 2023): **Informative** - a summary is informative if it encapsulates the crucial details from the source, offering a precise and concise presentation; **Quality** - a summary has an high quality if it is understandable and comprehensible; **Coherence** - a summary is coherent if it demonstrates a sound structure and organization; **Attributable** - all the information in the summary are attributable to the source; **Overall Preference** - the summary should succinctly, logically, and coherently convey the primary ideas presented in the source.

Concerning the reference summaries, it is worth emphasizing that, since there are no publicly available reference summaries for (the RD-relevant portions of) the European constitutions to date, we were assisted by legal professionals, with confidence on constitutional topics, who manually created summaries for each macro-topic. We hereinafter refer to these as **human-generated reference summaries**.

We employed Python implementations available on NLTK for stemming,¹² Scikit-learn for TF-IDF modeling,¹³ HuggingFace for Rouge,¹⁴ and GitHub for BLANC.¹⁵

RQ1 evaluation. Table 4.6 reports the cosine similarity values computed by comparing the sentence embeddings of the actual topics and the ones corresponding to the topics revealed by the generative model when asked to describe a macro-topic. For each macro-topic M , the table shows the minimum, maximum, and mean similarity over all best matchings between a micro-topic in M and a micro-topic generated by the model when asked about M .

¹² <https://www.nltk.org/howto/stem.html>

¹³ https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html

¹⁴ <https://huggingface.co/spaces/evaluate-metric/rouge>

¹⁵ <https://github.com/PrimerAI/blanc>

Table 4.6: *Answering RQ1* – For each macro-topic, the average, the maximum and the minimum cosine-similarity between the sentence embeddings of generated topics and the sentence embeddings of the best-matching actual topics. The most similar and least similar topics are those corresponding to the maximum and minimum cosine-similarity, respectively, for each macro-topic.

M	max similarity	most similar topic	min similarity	least similar topic	avg similarity
M1	0.778	slave	0.667	corppun	0.717
M2	0.763	remuner	0.554	childwrk	0.637
M3	0.708	intprop	0.556	transfer	0.636
M4	0.875	taxes	0.613	work	0.730
M5	0.766	abide	0.533	binding	0.646
M6	0.867	privacy	0.496	civmar	0.669
M7	0.855	speedtri	0.449	excrim	0.701
M8	0.709	ombuds	0.545	amparo	0.569
M9	0.716	indpolgr2	0.482	equalgr8	0.587

Looking at the table, we observe relatively tight ranges of values over mid-high regimes of cosine similarity – in most cases, the average similarity between the description of the actual topics and the ones of the generated topics is above 0.63. This gives evidence of the model’s knowledge about the existing RD-relevant topics, and its high confidence about the actual meanings within the constitutions. This is further confirmed by inspecting how well the descriptions of the generated topics are aligned with those of the best-matching actual topics, as we report in Appendix A.1.

It is worth emphasizing that, since the model was not given instructions on the number of topics to identify for each macro-topic, we also inspected if the model over-/underestimated the number of actual topics. By denoting with $(n, n')_M$ the actual (n) and generated (n') micro-topic numbers for a given macro-topic M , we found $(6, 15)_{M1}$, $(21, 14)_{M2}$, $(7, 15)_{M3}$, $(3, 15)_{M4}$, $(2, 15)_{M5}$, $(24, 20)_{M6}$, $(22, 18)_{M7}$, $(4, 15)_{M8}$, $(26, 15)_{M9}$: note that, although 4 out of 9 times the model underestimated the micro-topic number, there was an overall marked opposite tendency, with 53 overestimated topics.

RQ2 evaluation. We present the obtained results to answer our **RQ2**, based on the following organization: we first qualitatively assess the country-specific summary consistency using UMAP visualizations, then we comparatively analyze the generated summaries with the source texts and with the reference summaries according to the various evaluation criteria.

We started with qualitatively investigating how the country-specific summaries were consistent with each other in relation to the specific micro-topic. In this respect, we leveraged Uniform Manifold Approximation and Projection (UMAP) technique (McInnes et al., 2018b) to visualize a 2D projection of the embeddings generated for

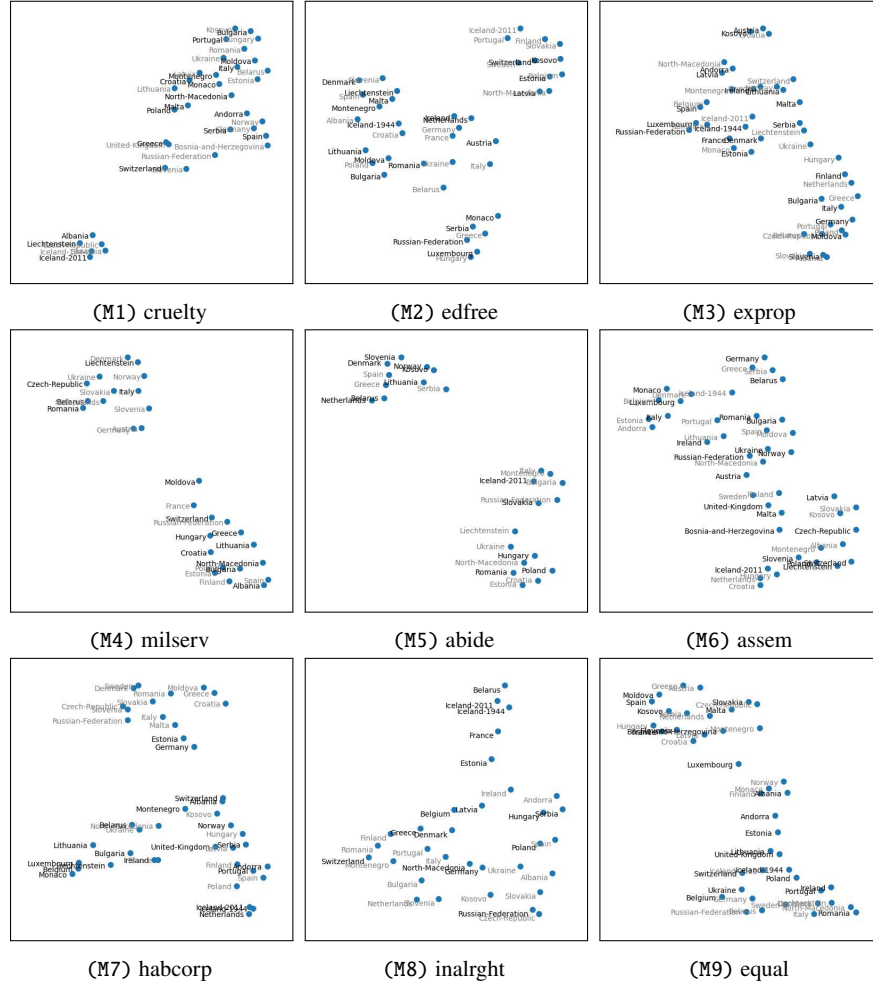


Fig. 4.8: Answering *RQ2 (Stage 1)* – For each macro-topic, the 2D UMAP visualization of the country-level summary embeddings related to the micro-topic with the highest number of participating countries.

the country-specific summaries.¹⁶ Fig. 4.8 shows for each macro-topic, the micro-topic with the highest number of involved countries. We notice cohesive groups of country-specific summary embeddings yet varying degree of separation among the

¹⁶ We configured UMAP with `n_neighbors` (i.e., the size of the local neighborhood UMAP will look at when attempting to learn the manifold structure of the data) set to 5, and `min_dist` (i.e., the minimum distance apart that points are allowed to be in the low dimensional representation) set to 0.1. This setup is chosen to balance local and global structure preservation given the relatively small-sized dataset—each micro-topic is discussed by not more than 40 constitutions on average.

Table 4.7: Answering **RQ2** – Summarization performance (mean and standard deviation values): generated summaries vs. input source texts.

M	CR (%)	Rouge-1	Rouge-2	Rouge-L	Rouge-LSum
M1	45.856±16.551	0.46±0.149	0.371±0.125	0.384±0.144	0.423±0.132
M2	57.250±21.169	0.554±0.164	0.439±0.138	0.486±0.154	0.514±0.153
M3	48.776±14.927	0.497±0.16	0.404±0.134	0.451±0.153	0.466±0.157
M4	65.227±2.607	0.668±0.032	0.524±0.053	0.588±0.052	0.614±0.063
M5	70.982±11.861	0.734±0.094	0.575±0.039	0.681±0.088	0.681±0.092
M6	49.611±19.141	0.491±0.190	0.394±0.188	0.451±0.188	0.465±0.186
M7	56.701±18.536	0.585±0.186	0.5±0.191	0.539±0.193	0.554±0.185
M8	65.125±33.277	0.657±0.332	0.584±0.337	0.622±0.335	0.631±0.327
M9	46.218±17.306	0.472±0.187	0.394±0.172	0.436±0.179	0.447±0.180
mean	56.194	0.569	0.465	0.515	0.533

M	N (%)	J	D	S-BERT	TF.IDF	BH	BT
M1	6.413±1.829	0.695±0.105	0.816±0.076	0.877±0.065	0.823±0.069	0.39±0.088	0.282±0.078
M2	7.553±5.002	0.638±0.132	0.771±0.104	0.909±0.053	0.904±0.043	0.29±0.081	0.301±0.073
M3	4.762±1.819	0.638±0.142	0.771±0.111	0.864±0.102	0.902±0.03	0.241±0.105	0.286±0.089
M4	8.154±3.209	0.714±0.049	0.833±0.033	0.922±0.063	0.91±0.061	0.338±0.077	0.351±0.053
M5	4.797±0.891	0.729±0.053	0.843±0.036	0.938±0.037	0.941±0.001	0.315±0.151	0.376±0.034
M6	8.922±10.416	0.588±0.163	0.727±0.141	0.88±0.091	0.899±0.084	0.245±0.117	0.296±0.092
M7	5.155±2.632	0.703±0.135	0.818±0.104	0.888±0.078	0.917±0.06	0.282±0.124	0.313±0.097
M8	4.563±1.853	0.717±0.222	0.820±0.158	0.796±0.210	0.882±0.085	0.332±0.252	0.356±0.211
M9	5.709±3.176	0.606±0.157	0.742±0.137	0.872±0.067	0.858±0.117	0.225±0.089	0.242±0.090
mean	6.225	0.670	0.793	0.883	0.893	0.295	0.311

countries across the topics. The UMAP visualizations reveal distinct 2D subspaces that indicate similarities and differences in the countries’ embeddings within the same micro-topic. For example, topics like *exprop* and *inalrht* correspond to more scattered patterns, indicating relatively higher diversity in the respective summary embeddings, whereas on *cruelty* and *abide*, more compact clusters emerge for some countries, suggesting a high agreement on those shared topics.

Table 4.7 shows the performance results of the comparison between generated summaries and corresponding input source texts, averaged over all topics in each macro-topic. Full details at micro-topic-level are reported in Appendix A.2, Tables A.1.

Compression ratio. Looking at Table 4.7, we observe that the per-macro-topic compression ratio varies from 46% to 71%, with a tendency to increase as the average number of involved countries (#C) decreases; note also that the latter, being generally lower than 43 for all macro-topics, generally indicates that only a subset of European constitutions contain portions that are relevant to a certain macro-topic, and that these subsets are of different size over the macro-topics. By delving into details at micro-topic-level (Table A.1), there are few cases that can be regarded as anomalies, whereby the compression ratio is either higher or lower than the per-macro-topic average. In the former case, this might be attributed to a difficulty of the model to fully eliminate redundant concepts or due to the imposition to preserve all the information (e.g., the micro-topic *amparo* is covered by only one country,

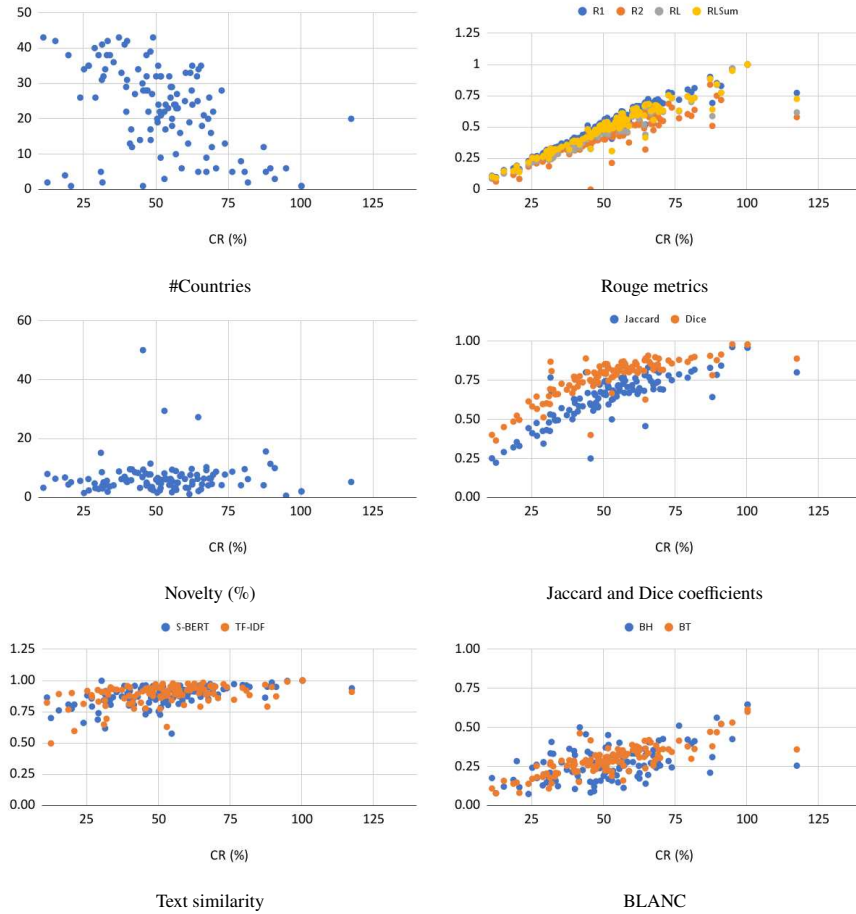


Fig. 4.9: Answering **RQ2** – Generated summaries vs. input source texts: Compression ratio (x-axis) versus each of the assessment criteria, for all RD-relevant topics.

so all the included information is preserved by the generative model). On the other hand, too low CR values (e.g., in *freerel* and *indcit*) might be attributed to excessive redundancy of concepts in the constitution portions of the involved countries. These extreme cases correspond nonetheless to a small subset: out of 114 topics, only 9 topics have CR exceeding 80% and 12 topics have CR below 30%.

Novelty and textual similarity. Novelty of the generated summaries is relatively low yet non-negligible, ranging from 4% to 9% over the per-macro-topic averages. From the micro-topic-level results (Table A.1), we also spot few exceptions having about 30% (*debtors*, *scifree*) or even 50% (*arms*) of novelty, which correspond to a very low coverage of countries whose constitutions discuss those topics.

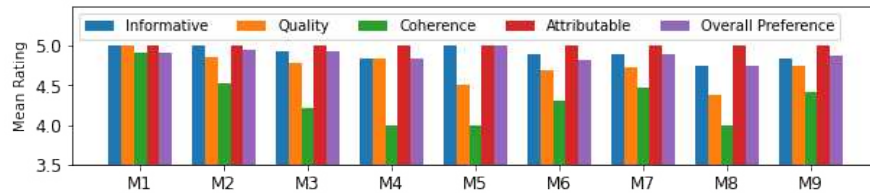


Fig. 4.10: Qualitative summary evaluation by GPT-3.5T

Being within mid-high regimes, the values of Jaccard and Dice similarity coefficients indicate a strong word-set sharing between the generated summaries and the original constitution-portions, which also couples with a very high lexical TF.IDF-vector-based similarity and semantic similarity based on S-BERT embeddings (for instance, the micro-topic *health* corresponds to a Jaccard similarity of 0.689, a Dice coefficient of 0.816, a lexical similarity of 0.975, and a semantic similarity of 0.896.) This indicates the high degree of textual similarity maintained in the summaries, reinforcing the reliability and faithfulness of the generated summaries to the original texts.

Rouge and BLANC scores. Table 4.7 also shows good indicators of quality from the Rouge metrics.¹⁷ In particular, Rouge-1 scores are between 0.46 and 0.73, indicating that the generated summaries captures many of the important words present in the input texts. In addition, Rouge-2 scores between 0.37 and 0.58, and Rouge-L scores between 0.38 and 0.68, suggest a moderate degree of similarity at the word-pair level and of the original sentence structure and sequence, respectively, reflecting a coherent and well-ordered summary that preserves the context, important expressions and the logical flow of information from the original text. (Looking at the micro-topic-level results in Table A.1, we also spot topics corresponding to Rouge peaks, such as *acfree* with an Rouge-1 score of 0.777, an Rouge-2 score of 0.686, and an Rouge-L score of 0.754, reinforcing the effectiveness of the summarization in maintaining the original content’s structure and meaning.)

Analogously, in several cases the observed BLANC scores (BT and BH) are higher than the typical threshold of 0.3 for indicating a good summary (e.g., the micro-topic *conright* has a BH score of 0.416 and a BT score of 0.348). This is likely due to the constraint we imposed to the model in generating summaries based on terms extracted from the original text, thus remaining as much faithful as possible to the constitution texts, which should be regarded as a desired requirement in any legal application of these summaries.

Correlation between compression ratio and evaluation criteria. Fig. 4.9 shows scatter plots to compare the compression ratio to each of the assessment criteria for

¹⁷ As expected, given the comparison with the input texts, Rouge scores tend to be higher compared to the domain-general abstractive/extractive summarization leaderboard scores (<https://paperswithcode.com/task/abstractive-text-summarization>).

Table 4.8: Answering **RQ2** – Summarization performance (mean and standard deviation values): generated summaries vs. reference summaries.

M	CR (%)	Rouge-1	Rouge-2	Rouge-L	Rouge-LSum
M1	113.643±25.254	0.659±0.049	0.391±0.04	0.284±0.082	0.437±0.032
M2	109.028±43.278	0.571±0.099	0.268±0.051	0.337±0.074	0.449±0.083
M3	101.451±30.919	0.591±0.055	0.279±0.066	0.344±0.076	0.48±0.057
M4	116.094±19.046	0.574±0.05	0.279±0.051	0.353±0.055	0.432±0.058
M5	123.026±23.385	0.577±0.001	0.305±0.032	0.404±0.067	0.492±0.013
M6	89.781±36.19	0.529±0.177	0.268±0.104	0.316±0.113	0.436±0.149
M7	115.593±34.255	0.562±0.1	0.271±0.057	0.324±0.055	0.444±0.086
M8	130.224±41.428	0.535±0.097	0.253±0.019	0.328±0.091	0.445±0.067
M9	102.117±38.436	0.539±0.129	0.261±0.081	0.318±0.086	0.432±0.11
mean	111.217	0.571	0.286	0.334	0.450

M	N (%)	J	D	S-BERT	TF-IDF
M1	33.51±5.812	0.514±0.064	0.677±0.056	0.926±0.016	0.69±0.127
M2	44.431±6.327	0.391±0.046	0.561±0.047	0.905±0.049	0.71±0.099
M3	46.543±8.563	0.366±0.081	0.531±0.086	0.892±0.045	0.78±0.076
M4	50.002±1.474	0.36±0.011	0.529±0.011	0.889±0.054	0.626±0.12
M5	52.358±0.594	0.346±0.023	0.514±0.025	0.92±0.032	0.711±0.106
M6	47.494±15.622	0.341±0.115	0.497±0.153	0.867±0.151	0.679±0.214
M7	47.834±8.888	0.362±0.054	0.529±0.059	0.892±0.062	0.644±0.104
M8	49.899±3.729	0.361±0.018	0.531±0.019	0.774±0.091	0.6±0.071
M9	44.324±9.817	0.374±0.068	0.541±0.078	0.874±0.073	0.653±0.16
mean	46.266	0.379	0.546	0.882	0.677

all RD-relevant topics. It can be noted that the Rouge, Jaccard, Dice and BLANC values increase as the compression ratio increases, which is not surprising since the closer the summary is to the original text in terms of length, the more likely there is a high overlap between the two texts. By contrast, we observe an inverse correlation between compression ratio and number of countries, since the more countries are involved, the larger and more redundant the overall text of the micro-topic of interest is likely to be. Novelty and textual similarity instead appear to be independent of the compression ratio.

Qualitative evaluation. Fig. 4.10 shows the distribution of rates provided by the generative model on the summaries produced at the end of Stage 2: for each macro-topic, each of the self-evaluation scores (i.e., informativeness, quality, coherence, attributability, and overall preference) is averaged over all unified summaries corresponding to the topics of that macro-topic. Note that this stage of evaluation was carried out separately from the summary generation process, since the model was asked to rate the summaries compared to the corresponding source texts by presenting them in a random order and without disclosing the originator of the summaries.

From the figure, we observe generally high values (at least 4 out of 5) for all criteria, on each macro-topic. In particular, attributability has consistently the maximum score (5), informativeness and overall preference are always close to 5, while quality and coherence are never below 4.5 and 4.0, respectively.

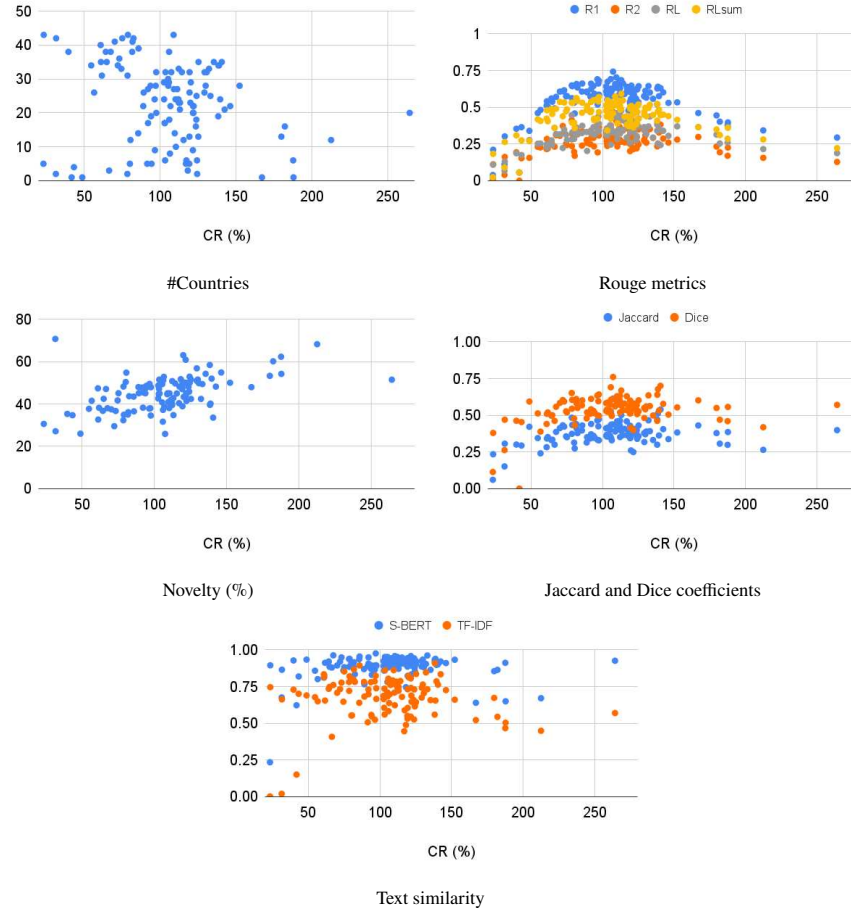


Fig. 4.11: Answering **RQ2** – Generated summaries vs. reference summaries: Compression ratio (x-axis) versus each of the assessment criteria, for all RD-relevant topics.

We extended our analysis to the comparison of the generated summaries with the human-generated reference summaries. To this purpose, we resorted to the same criteria as those used in the previous evaluation, with the exception of BLANC and self-evaluation by GPT-3.5T since they were used in absence of reference summaries. Topic-level details are provided in Appendix A.2, Table A.2.

Compression ratio, novelty, textual similarity and Rouge scores. Table 4.8 shows the macro-topic level averaged scores for each metric. In contrast to the comparison of the summaries with the original texts (Table 4.7), the average compression ratio is significantly higher and closer to 100, indicating that the size of the generated summaries is generally in line with the reference summaries. Instead, we notice an

inverse trend regarding the values of Novelty resp. Jaccard and Dice coefficients, which are increased, resp. decreased. This is actually not surprising, since the same text can in principle be abtractively summarized in alternatively valid summaries, each employing varying terms and syntactic structures. While semantic similarity is high on average, TF-IDF lexical similarity is generally decreased compared to Table 4.7. This confirms that there are structural and lexical differences between generated and reference summaries, although still being high semantic consistency. Regarding the Rouge metrics, similar average values are observed on Rouge-1 compared to Table 4.7, while there is a general decrease in values on Rouge-2, Rouge-L and Rouge-LSum. This suggests that, while the generated summary effectively captures key words or terms that are regarded as important in the reference summary, the generated summary is more faithful to the content and form of the original text w.r.t. the reference summary.

Correlation between compression ratio and evaluation criteria. Fig. 4.11 shows the scatter plots for each metric compared to the compression ratio. We notice that the trend of the number of countries w.r.t. the compression ratio is similar to that displayed in Fig. 4.9, although for several countries the generated summary appears to be longer than the reference summary (i.e., compression ratio above 100%). A moderately linear correlation can be noted w.r.t. the Rouge values until the compression ratio remains below 100%, otherwise a reverse trend occurs, with Rouge values decreasing as compression ratio increases. Regarding novelty, there is a rising trend by increasing compression ratio, and in general higher values compared to the corresponding plot in Fig. 4.9; in particular, the highest concentration occurs when the compression is between 100 and 150, mostly ranging between 40 and 60%, which confirms that the generated summary tends to use different words compared to the reference, as it is expected given that they are abtractive summaries. The Jaccard and Dice plots show that the word overlap tends to be much less affected by the compression ratio compared to the corresponding plot in Fig. 4.9, whereas no significant differences w.r.t. the comparison with the input texts are noted in the trends of text similarity.

4.5 Chapter Review

Our research investigates the application of TLMs to understanding and summarizing European constitutions, with a focus on the range of topics covered in the European constitutions that guarantee rights and duties (RD) to citizens.

Firstly, we present the first study providing lexical and semantic similarity analysis of the European constitutions. Our work unveils a number of major findings. In general, we have found that European constitutions share many RD-relevant topics. Despite for most micro-topics the lexical similarities are generally very low, it also happens that few pairs of European countries apparently address the same micro-topic in a similar manner; however, lexical analysis is not sufficient to capture the

similarities between countries at a fine grain. On the other hand, the semantic analysis unveils quite different behavior of the language models. In particular, it is evident that the legal models we consider are not directly applicable to similarity tasks, whereas the Sentence-Transformers demonstrate significantly better results, despite not being specifically trained on legal corpora. This is not actually surprising, since the constitutions are not written in a highly technical legal language, as they should be easily comprehensible even for non-experts. Secondly, the Sentence-Transformers are trained to generate sentence embeddings, which capture the semantic meaning and context of the texts rather than relying on specific legal terminology. As a result, the Sentence-Transformers can effectively capture the similarities and nuances of the constitutions. Among them, `all-mpnet-base-v1` has proven to be the best performer, although `all-distilroberta-v1` follows closely behind; by contrast, `gtr-t5-large` performs significantly worse, likely due to its specialization in semantic search and lack of fine-tuning for other use cases. Our work answers the preliminary research question (**RQ0**), finding evidence of the pre-requirement for addressing the subsequent two RQs. Our results confirmed that European constitution portions discussing RD-relevant topics are sufficiently similar to justify a summarization task.

The second research question (**RQ1**) required us to assess whether a generative language model, specifically GPT-3.5T, has adequate knowledge of European constitutions before challenging it with a summarization task. By comparing the descriptions generated by the model with the reference ones about the topics, the model proved to have substantial knowledge of the European constitutions, accurately describing relevant topics without being provided with any portion extracted from the constitutions.

Our third research question (**RQ2**) was addressed by first producing topic-specific summaries for the individual countries, then building a unified across-country summary for each topic. The resulting summaries were designed to capture the essence of the constitutions across different countries while maintaining the unique aspects of each country's constitution. The quality of the produced summaries was evaluated according to several criteria, including Rouge metrics and lexical/semantic similarity, both compared with the input sources and human-generated reference summaries. Let us wrap up main highlights through the macro-topics M2 (Social Rights), M6 (Civil and Political Rights), M7 (Legal Procedural Rights), and M9 (Equality, Gender and Minority Rights), which are exemplary cases given their broad coverage of topics and large number of countries involved; in particular, M2 covers a wide range of social rights, such as just remuneration, protection of consumers, and state support for various groups, M6 includes civil and political rights such as freedom of association, expression, and the right to privacy, M7 deals with legal procedural rights like the right to pre-trial release, principle of no punishment without law and requirement of jury trial, and M9 addresses minority rights, equality regardless of nationality, race, or gender. For each of such macro-topics, the summaries generated achieve high Rouge and BLANC scores as well as high lexical and semantic similarity, and moderate novelty, which indicate coherence and informativeness but also faithfulness to the original documents, maintaining a balance between country-

specific details and the essential principles of universal interest. thus providing a helpful assistance in understanding the European constitutional contents.

Overall, our research represents a first step towards assessing the implications of utilizing TLMs for understanding and summarizing European constitutions that address RD-relevant topics. We believe this can pave the way for further developments that aims to support a variety of tasks oriented to a more transparent, integrated and cooperative transnational legal landscape.

However, this study has some limitations that leave room for future improvement. In addressing **RQ0**, we notice that when the topics are highly similar, all the models faced difficulties in perceiving their differences. A fine-tuning phase on the constitution data would help recognize subtle nuances in meaning. Also, we are aware that there are many other legal models that could have been considered in the experimentation, and their inclusion could have provided further insights into real performances of legal architectures on similarity tasks. However, we opted to focus on a selected set of models that are widely recognized and representative of the current state-of-the-art in the field. Future research may include exploring a broader range of legal models to gain a comprehensive overview of their capabilities and limitations in detecting micro-topic similarities among constitutions.

For **RQ1** and **RQ2**, there exist various alternatives to our model choice not only looking at the OpenAI GPT family (e.g., GPT-4) or other commercially-licensed models but also to solutions based on open generative TLMs (e.g., Llama, Mistral). Either types of generative TLMs should be taken into account to get a much clearer picture of the capabilities of such models in understanding and summarizing European constitutions, thus providing effective support for computational tasks on European law.

Cost constraints had a twofold impact: preventing us from using more powerful OpenAI GPT models (GPT-4) and/or fine-tuning GPT-3.5T on European law corpora, but also from carrying out extensive prompt engineering to maximize the understanding and summarization capabilities of GPT-3.5T. A well-crafted prompt is instrumental in guiding the model's behavior and obtaining meaningful outputs, therefore further work should be dedicated to an enhanced prompt engineering phase.

We believe that incorporating various European law resources (e.g., Official Journal of the European Union, European case-law, summaries of legislation, etc.) might strongly enhance the capability of the model of reasoning about European constitutions and related law data.

Country-focused evaluation is missing. It would be interesting to understand, especially in the stage of across-country summarization, whether and to what extent one or more countries could lead a “primary” role in building an overall encompassing summary for each (macro-)topic. Moreover, further investigation would be needed to delve into the generated summaries' quality aspects that closely deal with the legal sublanguage used in the European constitutions as well as with those (country-specific) terms that were anonymized.

Our study in this work is focused on European countries' constitutions and their relations with the rights and duties theme. Nonetheless, the proposed methodol-

ogy can straightforwardly be extended to other countries/continents and/or thematic subjects in the constitutional context.

Finally, although our model was prompted equally for all countries, aspects related to fairness, equity and legal coherence of summaries w.r.t. the involved countries have not been checked.

PART II: AI-BASED NLP METHODS AND APPROACHES FOR LEXICAL ENTAILMENT

The study of entailment has long been recognized as a critical endeavor in NLP and related fields, as it is fundamental to several tasks such as classification, summarization, question answering, and machine translation. Given their remarkable success in solving the aforementioned tasks, TLMs have indeed reshaped the landscape of language understanding (Chang et al., 2024; Min et al., 2024); nonetheless, to achieve even more sophisticated capabilities in interpreting human communication through verbal nuances, unveiling entailment recognition into these models represents a demanding challenge for their development.

This has a number of motivations, which can be summarized as follows. Verbs hold a central role in language, serving as the foundational and semantic framework for conveying sentence meaning. Understanding their semantic nuances has been a long challenge due to the considerable variability and relatively looser cohesion of verb meanings compared to nouns (Gentner and France, 1988). Correctly inferring entailment relations between verbs is indeed essential for a TLM to be robust in understanding meanings and logical connections within sentences. By exploring how TLMs interpret and handle entailment among verbs, we can also gain insights into their decision-making processes, unveiling their strengths and weaknesses, and also contributing to model interpretability and refinement. As previously mentioned, handling verb entailment is crucial in various NLP tasks: on the one hand, improving models' comprehension of these relations can enhance the performance of these applications, but on the other hand, identifying and addressing biases or misinterpretations in entailment relations is also important in refining models to handle diverse linguistic contexts and minimize errors in real-world applications.

Chapter 5

Transformer-based Language Models for Lexical Entailment Recognition

Exploring the relationships between verbs can be addressed by considering the concept of *lexical entailment* (Fellbaum, 1990; Geffet and Dagan, 2005). Focusing on lexical entailment is paramount for several reasons. Verbs are crucial for expressing actions and relationships between entities, making it essential to properly capture their nuances. Grasping these relationships helps in deciphering sentence meanings and how verbs work together. Verbs often have polysemous and context-dependent meanings, and capturing entailment relations among verbs involves addressing challenges like aspect, modality, tense, and fine-grained semantic differences. Also, certain domains or specialized fields heavily rely on precise verb relations, such as in legal, scientific, or technical language.

A related aspect, as recently noted in Putra et al. (2023) for textual entailment in general, is an emergence for developing evaluation datasets and benchmarks for assessing the performance of models in entailment tasks. In this regard, we recognize the primary role played by *lexical databases*, which are meaningful resources for semantically exploring verb relations. In this work, we will focus on two widely used resources, namely WordNet (Miller et al., 1990) and HyperLex (Vulić et al., 2017).

Our study aims to unveil the actual capabilities of TLMs in recognizing lexical entailment relations (Greco et al., 2024a). By focusing on instances of verb pairs corresponding to entailment relations provided by lexical databases, we conduct a thorough evaluation based on eight TLMs, with emphasis on Open TLMs. We define different prompting strategies with different context details, both under a zero-shot and a few-shot setting, for asking a TLM to answer about a question relating to the entailment between any two verbs. Our primary evaluation goal is to understand to what extent TLMs are able to recognize lexical entailment between verbs by measuring their compliance with well-grounded, manually curated linguistic resources.

5.1 Related Works

Sainz et al. (2023b) address word sense disambiguation in terms of textual entailment to understand if BERT and RoBERTa can discriminate between different senses in a variety of domains. Tseng et al. (2023) train a mT5 model for generating Chinese word glosses, and raise the need for models to rely on semantic vectors.

Dealing with negation represents an additional challenge. Chen et al. (2023) assess negative commonsense knowledge of TLMs. Experiments carried out on Flan-T5 (Chung et al., 2024), GPT-3, Codex (Chen et al., 2021), Instruct GPT, and ChatGPT reveal behavior inconsistency among the TLMs. García-Ferrero et al. (2023) test the commonsense knowledge of open source TLMs (T5, Llama, Pythia (Biderman et al., 2023), Falcon, Vicuna) using both affirmative and negative sentences using various types of relations and negations. Results have shown that the TLMs excel in classifying affirmative sentences but fail in dealing with negative ones.

More generally, TLMs have been tested over various types of semantic relations. Lovón-Melgarejo et al. (2024) analyze the ability of BERT-based models and Sentence-Transformers to capture hierarchical semantic knowledge using WordNet-derived datasets. Oliveira (2023) apply BERT to capture synonyms, antonyms, hypernyms, and hyponyms in the Portuguese language. Hypernymy is also a focus of the study in Liao et al. (2023), which builds a dataset on the WordNet hypernyms and test several TLMs on hypernymy discovery, observing a consistent underperformance of TLMs when tasked with abstract concepts. Bai et al. (2022) assign words with a common WordNet hypernym into the same class, and train TLMs by gradually transitioning from predicting the class to predicting the token through a curriculum learning strategy. Also, Tikhomirov and Loukachevitch (2024) evaluate the use of TLMs with various prompts for hypernym prediction.

Other studies investigated how to leverage the relationships and definitions provided by WordNet to enhance the representation through generated embeddings and data augmentation (Loureiro and Jorge, 2019; Perçin et al., 2022).

Our work uniquely provides an analysis of how currently used TLMs can recognize verb entailment relations, in contrast with works with broader scopes like Lovón-Melgarejo et al. (2024) and Tikhomirov and Loukachevitch (2024). Our focus on verbs also differs from studies such as Sainz et al. (2023b), which consider an organization of relations into domain-specific and high-level concepts. Unlike Tseng et al. (2023), we do not focus on attempting to determine a specific sense for a sentence, instead our aim is to understand how TLMs answer to specific queries about semantic relations. Our methodology differs from Chen et al. (2023) and García-Ferrero et al. (2023), since they consider negative relations and not necessarily on verbs. Moreover, we focus on a representative set of recently developed open and commercially-licensed TLMs, while García-Ferrero et al. (2023) consider open models only and Chen et al. (2023) focus on models earlier than ChatGPT. Oliveira (2023) is limited to Portuguese and uses BERT to determine relations of hyponymy, hypernymy, synonymy, and antonymy. Unlike Liao et al. (2023), which train projection layers to learn WordNet hypernym relations, we approach the problem through a probing approach. Compared to Bai et al. (2022), which focuses on

reducing model perplexity, our interest is understanding how models specifically handle verb entailment relations.

5.2 Wordnet and Hyperlex Data

We resorted to two widely recognized and openly accessible lexical resources, which provided us with the means to address our research questions regarding the TLM awareness of verb entailment.

WordNet. To access WordNet verbs and entailments, we resort to the implementations provided by the NLTK¹ library. According to the logical organization in WordNet, verb entailments can be retrieved through two methods, namely *hyponyms()* and *entailments()*, where the former provides access to troponymies, and the latter includes the other kinds of entailment relations as described in Fellbaum (1990). Given a target verb synset, either method returns its associated set of hyponym/entailment synsets. To be consistent with the WordNet verb hierarchy which considers both direct and indirect hyponyms, we used the *hyponyms()* method recursively to get all hyponyms of a given synset.

To reduce the synset relations to lemma relations, a further step is to expand each pair of synsets as a set resulting from the Cartesian product of their respective lemma sets. Note also that each synset is provided with its definition (gloss), which transfers to each of its constituting lemmas.

By performing the above steps, we retrieved 114,490 lemma pairs based on *hyponyms()* and 2,352 lemma pairs based on *entailments()*. Therefore, the total of WordNet verb pairs that are **relevant** to the task under study is $114,490 + 2,352 = 116,842$. We then finally built our WordNet evaluation dataset by selecting as many verb-pairs that are **not relevant** to the task, by randomly picking 50% of them by rewiring the relevant pairs to obtain non-relevant pairs, and the other 50% from the complement set of WordNet verbs that are not involved in any type of entailment relation.

HyperLex. As mentioned in Sect. 2.4.1, HyperLex provides 380 verb pairs corresponding to different types of entailment relations: hyponymy, hypernymy, co-hyponymy, synonymy and antonymy. Among these, we notice that 169 verb pairs also appear in the set of WordNet pairs, six of which are labeled as synonyms and other six as hypernyms, while the remaining ones in HyperLex either are not present in entailment relations in WordNet or they are provided in reverse order; in particular, HyperLex has 71 reciprocal verb-pairs, of which 27 are shared with WordNet. Likewise in WordNet, we couple the set of 380 pairs (**relevant**) with as many verb pairs as **not relevant**, obtained by rewiring the relevant ones.

¹ <https://www.nltk.org/>

5.3 Methodology

Methods. Our study involves a representative selection of TLMs, which reflect various baseline architectures. Specifically, we use GPT-3.5 through the official OpenAI APIs, along with some of the most prominent *Open* TLMs (i.e., *open-source* or *open-weights*), namely Llama-3 in its 8B-parameter version, the 7B-parameter versions of Llama-2, Mistral, Falcon, Vicuna, Gemma (Team et al., 2024), and Intel NeuralChat.²

Prompts. We crafted three prompt schemes, which correspond to different types of instructions for the models. Also, for each prompt, we devised both a *zero-shot* scenario and a *few-shot scenario*, whereby we enhance the prompt with contextual examples.

We tailored the prompt setting depending on the underlying resource. When using verbs from WordNet, we augment the prompt by incorporating into it the definition of each verb lemma used in the prompt (i.e., the gloss of the corresponding synset from which the lemma is derived); we will use symbol $def()$ to denote a function returning the definition of a verb lemma. By contrast, for HyperLex, we do not augment the prompts since verb definitions are not originally provided within this resource. These two settings will enable us to investigate the impact of the presence/absence of verb definitions on the lexical entailment recognition capabilities of the examined TLMs.

Next we introduce our defined prompts (we shall specify by “[...]” the optionality of verb definitions). We begin with the *zero-shot* prompts we used as the first step of our experimental evaluation.

Direct Entailment. Our first type of prompt, dubbed as *Direct*, is devised to test the ability of a model to recognize (any) entailment relation between two verbs:

Direct Prompt

Given the verb v_1 [defined as $def(v_1)$] and the verb v_2 [defined as $def(v_2)$], what is the verb that entails the other?
Answer must be either of the form “X entails Y” or “there is no entailment”.

It should be noted that, through the *Direct* prompt, the model is required implicitly to first recognize the existence of entailment, and in this case, to decide which verb is the entailing verb, and which verb is the entailed verb.

Indirect Entailment. The *Direct* prompt requires the model to rely on its knowledge of the meaning of the word “entail”, thus explicitly framing an entailment recognition task. By contrast, our second type of prompt omits the use of “entail” and instead provides the model with a relational function that expresses a definition of entailment. We refer to this prompt type as *Indirect*:

² <https://huggingface.co/Intel/neural-chat-7b-v3-2>

Indirect Prompt

Relation F states that given two verbs X and Y , X and Y satisfy F if and only if when doing Y you are also doing X .

Given the verb v_1 [defined as $def(v_1)$] and the verb v_2 [defined as $def(v_2)$], what is X and what is Y ?

Answer must be either a pair (X,Y) or “relation F cannot be satisfied”.

Like for the **Direct** prompt, to answer the **Indirect** prompt the model needs to decide the correct order between entailing and entailed verbs, otherwise to recognize the case whereby no entailment relation holds for the two input verbs.

Reverse Entailment. Our third type of prompt is designed to allow us to determine whether the model can recognize the entailment relation in the presence of negation. That is, for each verb pair $\langle v_1, v_2 \rangle$ such that v_1 entails v_2 , we ask the model whether it holds true that “not v_2 entails not v_1 ”. We refer to this prompt type as **Reverse**:

Reverse Prompt

Given the verb v_1 [defined as $def(v_1)$] and the verb v_2 [defined as $def(v_2)$], answer YES if ‘not v_2 ’ entails ‘not v_1 ,’ NO otherwise.

Note that, despite the binary nature of its required response, the **Reverse** prompt poses a different challenge than the other two types in that the model is required to “reason” about a reverse entailment based on negation.

Few-shot settings. Our previously defined prompts are also used to perform in-context learning based on contextual examples. To this purpose, we define two few-shot scenarios, hereinafter referred to as *HyperLex-based few-shot examples* (HyperLex-FS) and as *Fellbaum-based few-shot examples* (Fellbaum-FS). In the former case, we select and introduce in a prompt four verb pairs that correspond to different difficulty levels based on the scores assigned by HyperLex to the verb pair. In the latter case, we introduce in a prompt one example for each of the four types of entailment relations described in Fellbaum (1990), namely *troponymy co-extensiveness*, *troponymy proper inclusion*, *backward presupposition* and *cause*. In both cases, the set of examples are fixed for all test verb-pairs, except when an example pair coincides with the test verb-pair (in which case, the example pair is replaced with another equivalent according to the particular strategy used).

Full details about our defined zero-shot prompts, few-shot prompts, and selection of the contextual examples are reported in Appendix B.

5.4 Experimental Evaluation

5.4.1 Model settings and deployment

As we leverage models that are aligned to human preferences, being them of *instruct* or *chat* type, we specified a “system prompt” to declare the model’s *role*. To this purpose, we set each model to be “*a linguistic expert who responds to user questions in a concise way.*”

A key aspect of our model deployment is the use of the *Guidance* library³ through its `select` method to achieve *constrained generation* (Liu et al., 2023b). This forces an TLM to produce *structured* outputs that adhere to the shape and contents required by our defined prompts (Sect. 5.3). This way, not only the models’ outcomes will be strictly pertinent to the admissible or valid responses, but also there is no constraint to set on the maximum number of tokens to generate. The model temperature was set to the minimum allowed value for any particular model, which is determined by the *Guidance* framework through its `select` method to achieve constrained generation. In reality, the temperature value is a number that is kept below 0.01, so as to avoid division by zero in the decoding probability computation. Likewise, we refrained from altering the *top-p* and *top-k* parameters from their default values of 50 and 1, respectively. Details on the experimental environment are reported in Appendix B.4.

5.4.2 Evaluation criteria

To validate the quality of the responses generated by the models, we resorted to two approaches: the first one based on standard statistical assessment criteria for a classification task, and the second one based on a *model’s self-evaluation*.

The former group includes *accuracy*, *precision*, *recall*, and *F1-score* calculated by comparing a model’s responses to the available ground-truth, for all models and prompt settings. Table 5.1 reports the definitions of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) for each of our defined types of prompt. Symbols $p^{\text{rel}}(X, Y)$ and $p^{\neg\text{rel}}(X, Y)$ are used to an input relevant and not-relevant verb-pair, respectively.

The model’s self-evaluation (Xiong et al., 2024) was carried out by asking the model to provide its response with a *confidence* score in [1..10], where higher scores correspond to higher degree of certainty by the model in deciding about entailment between two verbs. In the results, we shall report ***r-conf*** resp. ***nr-conf*** to denote the confidence of a model averaged over all relevant pairs resp. not-relevant pairs.

³ <https://github.com/guidance-ai/guidance>

Table 5.1: Description of the confusion matrix statistics for each of the three prompt types.

	Input	Answer
Direct		
TP	$p^{\text{rel}}(X, Y)$	X entails Y
TN	$p^{\neg\text{rel}}(X, Y)$	there is no entailment
FP	$p^{\neg\text{rel}}(X, Y)$	X entails Y, or Y entails X
FN	$p^{\text{rel}}(X, Y)$	there is no entailment
	$p^{\text{rel}}(X, Y)$	Y entails X
Indirect		
TP	$p^{\text{rel}}(X, Y)$	(X,Y)
TN	$p^{\neg\text{rel}}(X, Y)$	relation F cannot be satisfied
FP	$p^{\neg\text{rel}}(X, Y)$	(X,Y) or (Y,X)
FN	$p^{\text{rel}}(X, Y)$	relation F cannot be satisfied
	$p^{\text{rel}}(X, Y)$	(Y,X)
Reverse		
TP	$p^{\text{rel}}(X, Y)$	Yes
TN	$p^{\neg\text{rel}}(X, Y)$	No
FP	$p^{\neg\text{rel}}(X, Y)$	Yes
FN	$p^{\text{rel}}(X, Y)$	No

5.4.3 Results

We organize the result presentation into two subsections, which correspond to evaluation on WordNet data, resp. HyperLex data. We shall first discuss results obtained in a zero-shot scenario, then those in the few-shot scenarios. In the tables, bold resp. underlined values will correspond to the highest resp. second-highest scores per column.

Evaluation on WordNet data and zero-shot prompting. Table 5.2 reports the results obtained under the zero-shot setting on WordNet verb pairs. Let us first consider results corresponding to the **Direct** prompting. NeuralChat and Gemma emerge as the best-performing models, although according to different assessment criteria: NeuralChat excels in precision (0.95) and accuracy (0.62), followed by Llama-3, whereas Gemma outperforms the others in recall (0.89) and F1-score (0.62), with Vicuna as the second best. This indicates relatively more robustness by models like NeuralChat and Llama-3 against false positives and true negatives, while other models like Gemma and Vicuna behave better in terms of false negatives.

The above scenario changes when prompting the models through **Indirect**, i.e., with an explicit definition of entailment without using any cue word. The observed change regards not just the measurement scores (which correspond to higher F1 on average) but also the best performing models, which are now Llama-2, Mistral and Vicuna.

The **Reverse** prompting corresponds to a different scenario, whereby GPT-3.5 emerges as the most effective model according to all criteria, followed by Llama-3,

Table 5.2: Zero-shot prompting results on WordNet verb pairs.

	<i>A</i>	<i>P</i>	<i>R</i>	<i>F1</i>	r-conf	nr-conf
Direct						
gpt-3.5	0.144	0.151	0.154	0.152	8.93	8.12
llama-3	<u>0.571</u>	<u>0.862</u>	0.169	0.283	<u>9.94</u>	<u>9.95</u>
llama-2	0.422	0.363	0.205	0.262	8.00	8.00
mistral	0.536	0.604	0.210	0.311	8.97	9.13
falcon	0.352	0.263	0.165	0.202	10.00	10.00
vicuna	0.459	0.463	<u>0.515</u>	<u>0.488</u>	9.00	9.00
neural-chat	0.622	0.946	0.259	0.407	10.00	10.00
gemma	0.450	0.473	0.881	0.616	2.00	2.00
Indirect						
gpt-3.5	0.082	0.141	0.164	0.151	8.89	<u>9.36</u>
llama-3	0.334	0.392	0.605	0.476	<u>9.07</u>	<u>8.65</u>
llama-2	0.670	0.623	0.860	0.723	7.98	8.00
mistral	0.470	0.485	0.936	<u>0.639</u>	9.00	9.00
falcon	0.419	0.456	0.839	<u>0.591</u>	10.00	10.00
vicuna	<u>0.478</u>	<u>0.488</u>	<u>0.920</u>	0.638	9.00	9.00
neural-chat	0.360	0.418	0.718	0.529	10.00	10.00
gemma	0.421	0.456	0.820	0.586	2.00	2.00
Reverse						
gpt-3.5	0.629	0.577	<u>0.965</u>	0.722	9.29	9.14
llama-3	0.605	0.579	<u>0.767</u>	0.660	<u>9.74</u>	<u>9.67</u>
llama-2	0.539	<u>0.599</u>	0.236	0.339	7.99	7.99
mistral	<u>0.500</u>	<i>NaN</i>	<u>0.000</u>	<i>NaN</i>	<u>9.00</u>	<u>9.00</u>
falcon	0.527	0.515	0.909	0.658	10.00	10.00
vicuna	0.500	0.500	1.000	<u>0.667</u>	9.00	9.00
neural-chat	<u>0.616</u>	0.807	0.305	0.442	10.00	10.00
gemma	0.558	0.551	0.629	0.588	2.00	2.00

Vicuna, and NeuralChat. Despite the negation-based reversed form of the entailment relation to recognize, the good behavior of GPT-3.5 under **Reverse** contrasts with the disappointing results obtained under the other two prompts, which would suggest that GPT-3.5 is more suited to deal with the task at hand when prompted to provide a yes/no answer.

Interesting remarks can also be drawn from a comparison of the behaviors of the models w.r.t. their architectural commonalities. For instance, Vicuna can behave generally better than the Llama models in terms of recall and F1-score, despite sharing most of the same architecture with them, which might be explained as an effect of the fine-tuning of Vicuna that was carried out based on user-shared ChatGPT conversations collected from ShareGPT.com. Likewise, when using the **Direct** and **Reverse** prompts, NeuralChat outperforms Mistral despite being based on it, which can be attributed to the fine-tuning process using the higher-quality *SlimOrca* dataset (Mukherjee et al., 2023), and to a *Direct Preference Optimization* (DPO) phase,⁴ designed to better align the model with human preferences.

⁴ huggingface.co/datasets/Intel/orca_dpo_pairs

Table 5.3: HyperLex-FS prompting results on WordNet verb pairs.

	<i>A</i>	<i>P</i>	<i>R</i>	<i>F1</i>	r-conf	nr-conf
Direct						
gpt-3.5	0.318	0.240	0.169	0.198	8.13	7.37
llama-3	0.660	0.831	<u>0.402</u>	<u>0.542</u>	<u>9.71</u>	<u>9.30</u>
llama-2	0.358	0.256	0.149	0.188	8.01	8.00
mistral	<i>0.500</i>	<i>NaN</i>	<i>0.000</i>	<i>NaN</i>	8.75	8.86
falcon	0.444	0.436	0.385	0.409	10.00	10.00
vicuna	<u>0.572</u>	<u>0.574</u>	0.557	0.566	8.96	8.95
neural-chat	<i>0.500</i>	<i>NaN</i>	<i>0.000</i>	<i>NaN</i>	<i>10.00</i>	<i>10.00</i>
gemma	0.353	0.341	0.316	0.328	2.00	2.00
Indirect						
gpt-3.5	0.398	0.443	<u>0.796</u>	0.569	7.66	8.18
llama-3	<u>0.634</u>	<u>0.606</u>	<u>0.767</u>	<u>0.677</u>	<u>9.77</u>	<u>9.27</u>
llama-2	0.048	0.087	0.095	0.091	8.00	8.00
mistral	0.722	0.749	0.668	0.706	9.00	9.00
falcon	0.447	0.472	0.893	0.617	10.00	10.00
vicuna	0.357	0.416	0.709	0.524	9.00	9.00
neural-chat	0.330	0.389	0.594	0.470	10.00	10.00
gemma	<i>0.500</i>	<i>NaN</i>	<i>0.000</i>	<i>NaN</i>	<i>2.00</i>	<i>2.00</i>
Reverse						
gpt-3.5	0.664	0.613	0.890	0.726	7.58	7.60
llama-3	<u>0.617</u>	<u>0.958</u>	0.244	0.389	8.94	8.71
llama-2	0.539	0.602	0.231	0.333	8.00	8.00
mistral	<i>0.500</i>	<i>NaN</i>	<i>0.000</i>	<i>NaN</i>	6.78	7.20
falcon	0.557	0.565	<u>0.501</u>	<u>0.531</u>	<u>9.90</u>	<u>9.91</u>
vicuna	0.605	0.743	0.320	0.447	9.00	9.00
neural-chat	0.583	0.972	0.172	0.292	10.00	9.99
gemma	<i>0.500</i>	<i>NaN</i>	<i>0.000</i>	<i>NaN</i>	<i>2.00</i>	<i>2.00</i>

Another interesting remark regards the confidence expressed by the models in answering the prompts, which is consistently very high, or even equal to the maximum (10.0) for Falcon and NeuralChat. One exception is represented by Gemma which, despite being a reliable model under all promptings, consistently associates its answers with confidence 2, as it was inherently extremely cautious, or uncertain, on the task.

Evaluation on WordNet and HyperLex-FS prompting. Our results based on the HyperLex-FS strategy (Table 5.3) show that some models can better recognize verb entailments when supported by contextual examples. This holds especially for GPT-3.5 and Llama-3 under Direct and Indirect promptings (with percentage increase in F1 of 92% for Llama-3 under Direct and of 277% for GPT-3.5 under Indirect), but also Falcon, Vicuna and Mistral benefit from the HyperLex-FS strategy.

By contrast, such models as Gemma under Indirect and Reverse, Mistral under Direct and Reverse, and NeuralChat under Direct, are unexpectedly damaged by the use of the HyperLex-FS prompting, as it turns out to break their ability to properly answer. This might be ascribed to the increased complexity of the prompt,

Table 5.4: Fellbaum-FS prompting results on WordNet verb pairs.

	<i>A</i>	<i>P</i>	<i>R</i>	<i>F1</i>	r-conf	nr-conf
Direct						
gpt-3.5	0.426	0.358	0.187	0.245	8.03	7.33
llama-3	0.661	0.617	0.852	0.715	<u>9.29</u>	8.54
llama-2	0.290	0.271	0.248	0.259	8.00	8.00
mistral	<u>0.500</u>	<i>NaN</i>	<u>0.000</u>	<i>NaN</i>	8.85	8.87
falcon	0.442	0.436	0.392	0.413	10.00	10.00
vicuna	<u>0.597</u>	<u>0.598</u>	<u>0.593</u>	<u>0.595</u>	8.96	<u>8.95</u>
neural-chat	<u>0.500</u>	<i>NaN</i>	<u>0.000</u>	<i>NaN</i>	<u>10.00</u>	<u>10.00</u>
gemma	0.353	0.341	0.316	0.328	2.00	2.00
Indirect						
gpt-3.5	0.416	0.454	<u>0.832</u>	0.587	7.52	8.09
llama-3	<u>0.508</u>	<u>0.505</u>	<u>0.761</u>	<u>0.607</u>	<u>9.87</u>	<u>9.57</u>
llama-2	0.120	0.194	0.240	0.215	8.00	8.00
mistral	0.658	0.715	0.525	0.605	9.00	9.00
falcon	0.461	0.480	0.922	0.631	10.00	10.00
vicuna	0.458	0.473	0.734	0.576	8.99	9.00
neural-chat	0.414	0.452	0.813	0.581	10.00	10.00
gemma	<u>0.500</u>	<i>NaN</i>	<u>0.000</u>	<i>NaN</i>	<u>2.00</u>	<u>2.00</u>
Reverse						
gpt-3.5	0.659	0.626	<u>0.789</u>	0.698	7.62	7.62
llama-3	<u>0.617</u>	<u>0.942</u>	0.251	0.396	9.23	9.03
llama-2	0.526	0.523	0.585	0.552	8.00	8.00
mistral	<u>0.500</u>	<i>NaN</i>	<u>0.000</u>	<i>NaN</i>	7.07	7.38
falcon	0.522	0.512	0.947	<u>0.665</u>	<u>9.94</u>	<u>9.96</u>
vicuna	0.592	0.738	0.286	0.412	9.00	9.00
neural-chat	0.577	0.964	0.160	0.274	10.00	9.99
gemma	<u>0.500</u>	<i>NaN</i>	<u>0.000</u>	<i>NaN</i>	<u>2.00</u>	<u>2.00</u>

rather than its length, which however was ensured to be within the models' limits of maximum token length.

The higher complexity of the HyperLex-FS-based prompting w.r.t. the zero-shot scenario appears not to impact on the models' confidence, which remains equally high, with Falcon and (partly) NeuralChat confirming to be perfectly confident models.

Evaluation on Wordnet and Fellbaum-FS prompting. Let us now consider results corresponding to the Fellbaum-FS-based prompting which exploits examples of entailment relations in the Fellbaum taxonomy. Results are shown in Table 5.4.

One major remark is that the Fellbaum-FS strategy allows some models to further improve their ability of recognizing entailments. Compared to HyperLex-FS, significant improvements occur for the Llama models and Falcon, but also for GPT-3.5 and Vicuna under Direct and Indirect.

The Fellbaum-FS strategy turns out to be the best setting in absolute for Llama-3, GPT-3.5 and Vicuna under Direct, for Falcon, GPT-3.5 and NeuralChat under Indirect, and partly for Falcon, GPT-3.5 and Llama-3 under Reverse. Also, as already observed for the HyperLex-FS strategy, the performance of certain models

Table 5.5: Zero-shot prompting results on HyperLex verb pairs.

	<i>A</i>	<i>P</i>	<i>R</i>	<i>F1</i>	r-conf	nr-conf
Direct						
gpt-3.5	0.197	0.089	0.066	0.076	8.520	8.043
llama-3	0.405	0.409	0.426	0.418	9.075	8.476
llama-2	0.355	0.353	0.347	0.350	8.000	8.000
mistral	<u>0.533</u>	<u>0.562</u>	0.300	0.391	<u>9.351</u>	<u>9.228</u>
falcon	0.391	0.294	0.155	0.203	10.000	10.000
vicuna	0.346	0.403	<u>0.639</u>	<u>0.494</u>	9.000	9.000
neural-chat	0.579	0.885	0.182	0.301	10.000	10.000
gemma	0.466	0.482	0.932	0.636	2.000	2.000
Indirect						
gpt-3.5	0.024	0.045	0.047	0.046	8.556	8.664
llama-3	0.271	0.352	0.542	0.427	<u>9.325</u>	8.996
llama-2	0.504	0.636	0.018	0.036	6.449	6.281
mistral	<u>0.479</u>	<u>0.489</u>	0.958	0.648	9.000	<u>9.008</u>
falcon	<u>0.375</u>	0.429	0.750	0.545	10.000	10.000
vicuna	0.467	0.483	<u>0.934</u>	<u>0.637</u>	9.000	9.000
neural-chat	0.345	0.408	<u>0.689</u>	0.513	10.000	10.000
gemma	0.159	0.239	0.313	0.271	2.000	2.000
Reverse						
gpt-3.5	0.514	0.507	<u>0.995</u>	0.672	8.870	8.553
llama-3	0.625	0.673	0.487	0.565	<u>9.648</u>	<u>9.425</u>
llama-2	0.533	<u>0.727</u>	0.105	0.184	7.968	7.961
mistral	0.500	NaN	0.000	NaN	7.366	7.413
falcon	0.503	0.502	0.668	0.573	10.000	10.000
vicuna	0.500	0.500	1.000	<u>0.667</u>	9.000	9.000
neural-chat	0.689	0.919	0.416	0.572	10.000	10.000
gemma	<u>0.555</u>	0.672	0.216	0.327	2.000	2.000

is inhibited by the incorporation of usage examples in the prompt. Yet, the models' confidence values remain very similar to those observed for the HyperLex-FS strategy.

Evaluation on HyperLex data and zero-shot prompting. Looking at the results in Table 5.5, we notice a certain consistency with the zero-shot results on WordNet verb pairs (Table 5.2) in terms of best-performing models for each of the assessment criteria: Gemma and NeuralChat under the Direct prompting, Mistral and Llama-2 under the Indirect prompting, Vicuna and NeuralChat but also GPT-3.5 under the Reverse prompting. Considering all models and promptings, results are substantially comparable to those achieved on WordNet data; however, in several cases, the models' performances are lower than on WordNet data, which would be explained by the lack of verb definitions in the prompts that use HyperLex verb pairs.

Evaluation on HyperLex data and HyperLex-FS prompting. The HyperLex-FS based results (Table 5.6) offer a view which again resembles analogous results on WordNet data. In fact, relative improvements w.r.t. the zero-shot scenario occur for Llama-3, GPT-3.5 and Falcon under Direct and Indirect, and Vicuna under Direct, whereas most models cannot take advantage of the examples under Reverse.

Table 5.6: HyperLex-FS prompting results on HyperLex verb pairs.

	<i>A</i>	<i>P</i>	<i>R</i>	<i>F1</i>	r-conf	nr-conf
Direct						
gpt-3.5	0.329	0.207	0.121	0.153	8.504	7.839
llama-3	0.711	0.666	0.845	0.745	8.896	8.286
llama-2	0.504	0.516	0.126	0.203	8.000	8.000
mistral	0.500	NaN	0.000	NaN	9.000	9.000
falcon	0.414	0.406	0.371	0.388	10.000	10.000
vicuna	0.495	0.496	0.584	0.536	7.862	8.448
neural-chat	0.500	NaN	0.000	NaN	10.000	10.000
gemma	0.321	0.321	0.321	0.321	2.000	2.000
Indirect						
gpt-3.5	0.300	0.375	0.600	0.462	8.474	8.936
llama-3	0.557	0.537	0.826	0.651	10.000	9.947
llama-2	0.042	0.078	0.084	0.081	8.000	8.000
mistral	0.321	0.361	0.463	0.406	9.000	9.000
falcon	0.451	0.474	0.903	0.622	10.000	10.000
vicuna	0.204	0.290	0.408	0.339	9.000	9.000
neural-chat	0.229	0.195	0.174	0.184	10.000	10.000
gemma	0.500	NaN	0.000	NaN	2.000	2.000
Reverse						
gpt-3.5	0.501	0.501	0.929	0.651	7.898	7.799
llama-3	0.551	0.915	0.113	0.201	8.048	8.035
llama-2	0.499	0.499	0.997	0.665	8.000	8.000
mistral	0.500	NaN	0.000	NaN	7.755	7.903
falcon	0.501	1.000	0.003	0.005	9.961	9.966
vicuna	0.593	0.771	0.266	0.395	9.000	9.000
neural-chat	0.558	0.907	0.129	0.226	10.000	10.000
gemma	0.500	NaN	0.000	NaN	2.000	2.000

Evaluation on HyperLex data and Fellbaum-FS prompting. While no significant differences are observed in terms of the models’ confidence, the Fellbaum-FS based results (Table 5.7-bottom) allow us to draw remarks that are similar in some cases to the corresponding evaluation on WordNet data. Particularly, compared to HyperLex-FS based results, Vicuna and, to a less extent, GPT-3.5 performance increases under Direct and Indirect, whereas Llama-3 and Falcon benefit from Fellbaum-FS only under the Reverse and Indirect prompting, respectively. Mistral also takes advantage of the Fellbaum-FS strategy under the Indirect prompting. Overall, Llama-3 remains the best performing model on the few-shot scenarios.

5.4.4 Further experimental results

Impact of WordNet relation types. As we discussed in Sect. 5.2, WordNet verb pairs involved in entailment relations can be accessed through two methods, namely *hyponyms()* and *entailments()*.

Table 5.7: Fellbaum-FS prompting results on HyperLex verb pairs.

	<i>A</i>	<i>P</i>	<i>R</i>	<i>F1</i>	r-conf	nr-conf
Direct						
gpt-3.5	0.305	0.206	0.137	0.165	8.203	7.492
llama-3	0.620	0.572	0.947	0.714	8.567	8.038
llama-2	0.497	0.485	0.084	0.143	8.000	8.000
mistral	0.500	NaN	0.000	NaN	9.000	9.000
falcon	0.395	0.367	0.289	0.324	10.000	10.000
vicuna	0.543	0.540	0.587	0.562	7.228	7.790
neural-chat	0.500	NaN	0.000	NaN	10.000	10.000
gemma	0.321	0.321	0.321	0.321	2.000	2.000
Indirect						
gpt-3.5	0.317	0.388	0.634	0.482	8.602	8.923
llama-3	0.484	0.491	0.887	0.632	10.000	9.985
llama-2	0.138	0.216	0.276	0.243	8.000	8.000
mistral	0.347	0.392	0.553	0.459	9.000	9.000
falcon	0.462	0.480	0.924	0.632	10.000	10.000
vicuna	0.246	0.327	0.482	0.390	9.000	9.000
neural-chat	0.201	0.244	0.284	0.262	10.000	10.000
gemma	0.500	NaN	0.000	NaN	2.000	2.000
Reverse						
gpt-3.5	0.499	0.499	0.721	0.590	7.797	7.764
llama-3	0.605	0.955	0.221	0.359	8.496	8.473
llama-2	0.503	0.501	0.995	0.667	8.000	8.000
mistral	0.500	NaN	0.000	NaN	7.534	7.797
falcon	0.514	0.649	0.063	0.115	9.783	9.724
vicuna	0.559	0.846	0.145	0.247	9.000	9.000
neural-chat	0.532	0.853	0.076	0.140	10.000	10.000
gemma	0.500	NaN	0.000	NaN	2.000	2.000

In Table 5.8, we summarize results on our evaluation of the impact of the two methods on the TLMs’ responses, where the values reported in the table correspond to percentage values of the correctly identified verb pairs. Results show no significant differences in the percentage of correctly recognized verb relations for the best-performing models, for each prompt type and zero-/few-shot scenario. On average over all models, the set of correctly recognized verb relations of both types tends to be larger in the zero-shot scenario (around 33% for Direct, 68% for Indirect, and 60% for Reverse) followed by Fellbaum-FS (around 32% for Direct, 56% for Indirect, and 39% for Reverse).

Distribution of the lexicographers’ files over all verbs involved in entailment relations. WordNet verbs can also be categorized based on the 15 lexicographers’ files for verbs. In the NLTK WordNet library, these verb categories can be accessed through the *lexname()* function defined over synsets.

We carried out an additional statistical analysis focused on the distribution of the 15 lexicographers’ files over all verbs involved in entailment relations. We present our developed methodology as follows.

Table 5.8: Percentage of WordNet *hyponyms()*, resp. *entailments()*, based verb-pairs correctly recognized in entailment relations by the models, for each prompt type and strategy

	Direct		Indirect		Reverse	
	<i>Ent.</i>	<i>Hypo.</i>	<i>Ent.</i>	<i>Hypo.</i>	<i>Ent.</i>	<i>Hypo.</i>
Zero-shot						
gpt-3.5	18.35	12.45	17.40	15.30	96.65	96.65
llama-3	24.55	15.35	56.45	61.40	75.55	76.40
llama-2	21.65	20.60	52.00	49.15	19.40	24.80
mistral	22.55	20.75	92.40	93.60	0.00	0.00
falcon	17.50	16.70	82.15	82.90	85.75	91.45
vicuna	49.55	51.70	89.45	88.55	100.00	100.00
neural-chat	28.45	26.45	66.20	73.65	29.10	31.15
gemma	86.90	89.40	80.50	83.50	62.15	63.70
HyperLex-FS						
gpt-3.5	18.05	15.65	78.25	80.90	90.30	87.75
llama-3	42.40	38.50	73.95	72.60	26.30	24.95
llama-2	14.00	14.45	9.65	7.60	31.20	25.30
mistral	0.00	0.00	33.55	36.25	0.00	0.00
falcon	38.50	39.85	87.05	89.40	52.00	50.70
vicuna	54.65	55.90	72.50	70.20	36.55	34.00
neural-chat	0.00	0.00	64.90	59.35	15.05	17.65
gemma	31.65	31.50	0.00	0.00	0.00	0.00
Fellbaum-FS						
gpt-3.5	19.15	18.20	82.05	84.30	81.85	76.05
llama-3	85.25	86.55	74.85	75.65	31.10	25.65
llama-2	23.75	24.35	24.70	22.95	69.65	58.05
mistral	0.00	0.00	27.25	24.20	0.00	0.00
falcon	38.30	39.45	90.35	92.15	94.35	94.35
vicuna	56.20	58.75	65.75	67.15	35.00	29.00
neural-chat	0.00	0.00	83.25	80.65	18.10	16.35
gemma	31.65	31.50	0.00	0.00	0.00	0.00

For each model M, for each prompt type P (i.e., Direct, Indirect, Reverse), and for each relation in entailments(), hyponyms(), in both zero-shot and few-shot settings, we calculated:

- the distribution of *lexname* categories associated to all entailing verbs X, for which the model M correctly answered the prompt P on the entailment relation (X,Y).
- the distribution of *lexname* categories associated to all entailed verbs Y, for which the model M correctly answered the prompt P on the entailment relation (X,Y).
- the distribution of *lexname* categories involved in those verb pairs for which the model M correctly answered the prompt P on the entailment relation (X,Y), such that the *lexname* of X coincides with the *lexname* of Y.

For each of these distributions, we calculated the distribution entropy and inspected the most frequently occurring *lexname* categories.

The above analysis was analogously repeated for all verb pairs involving entailment relations that were not correctly recognized by a model. In Table 5.9, we present a summary of the results obtained on the zero-shot scenario — analogous results were observed on the two few-shot scenarios as well. Specifically, we report the following information: under the “top-3 lexnames” column, we report for each model and setting, the three *lexname* categories that are most frequently associated with either verb in a pair, over all prompting types, whereas the “avg. norm. entropy” column refers to the average (over all prompting types) of the normalized entropy values of the distributions of *lexname* categories. We also note that (results not shown) the 15 *lexname* categories are always represented in each of the distributions, regardless of the response type, which indicates that no subset of verb categories characterizes either the ‘correct’ or the ‘wrong’ answers.

Looking at the table at first glance, there is evidence of a small subset of *lexname* categories that consistently appear in the top-3 column; this is actually not surprising since those correspond to the most represented *lexname* categories among the verbs in WordNet, namely ‘change’, ‘contact’, ‘communication’, ‘motion’ and ‘social’, with more than 60% of the verb synsets falling into one of these categories. More importantly, by comparing the results corresponding to ‘correct’ and ‘wrong’ responses, the top-3 categories are mostly overlapping; this holds consistently for each model, regardless of the relation type and whether few-shots were used in the prompts. Also, for each model and relation type, the values of average normalized entropy (ranging within [0,1]) of the *lexname* distributions of ‘correct’ and ‘wrong’ responses are equally very high and similar to each other, thus hinting at a common trait of heterogeneity of *lexname* categories occurring in verb pairs corresponding to valid (either correct or wrong) responses.

Overall, based on this empirical evidence, we can conclude that the *lexname* categories cannot be regarded as predictors of a model’s performance in recognizing verb entailments; therefore, there is no evidence that TLMs fail to understand a particular subset of verbs in recognizing lexical entailments.

Preliminary experiments on TaxoLLaMA. At a late stage of writing of this work we came across TaxoLLaMA (March 2024) proposed in Moskvoretskii et al. (2024) as a lightweight fine-tune of LLaMA2-7b, which is designed to deal with multiple lexical semantics tasks with focus on taxonomy related tasks, including Taxonomy Enrichment, Hypernym Discovery, Taxonomy Construction, and Lexical Entailment tasks. We provide preliminary results concerning this model.

In Tables 5.10–5.13 we present the results of our evaluation of TaxoLLaMA on our WordNet and HyperLex datasets, with focus on its comparison with the Llama models previously included in our evaluation TLMs. Looking at the results from the tables, it stands out that TaxoLLaMA is not only outperformed by Llama-3 in nearly all cases, but also it might still behave worse than Llama-2 under certain conditions (e.g., zero-shot under Indirect on both datasets, HyperLex-FS under Reverse on WordNet data, and in other cases according to one or more assessment criteria). Interestingly, these findings contrast with the fact that TaxoLLaMA derives from a fine-tuning of Llama-2 on WordNet and related lexical taxonomies, and that in

Table 5.9: Distribution of the verb lexicographers’ files: zero-shot scenario

model	relation	outcome	top-3 lexnames	avg. norm. entropy
gpt-3.5	entailments()	correct	contact, motion, communication	0.89
gpt-3.5	entailments()	wrong	contact, motion, communication	0.91
gpt-3.5	hyponyms()	correct	change, contact, communication	0.90
gpt-3.5	hyponyms()	wrong	change, communication, contact	0.91
llama-3	entailments()	correct	contact, motion, communication	0.91
llama-3	entailments()	wrong	contact, change, communication	0.93
llama-3	hyponyms()	correct	change, contact, communication	0.91
llama-3	hyponyms()	wrong	change, contact, communication	0.91
llama-2	entailments()	correct	contact, motion, communication	0.91
llama-2	entailments()	wrong	contact, motion, communication	0.91
llama-2	hyponyms()	correct	contact, change, communication	0.89
llama-2	hyponyms()	wrong	contact, change, communication	0.90
mistral	entailments()	correct	contact, motion, communication	0.91
mistral	entailments()	wrong	contact, body, motion	0.91
mistral	hyponyms()	correct	change, communication, stative	0.90
mistral	hyponyms()	wrong	communication, contact, motion	0.93
falcon	entailments()	correct	contact, motion, communication	0.92
falcon	entailments()	wrong	contact, motion, communication	0.92
falcon	hyponyms()	correct	contact, change, communication	0.92
falcon	hyponyms()	wrong	contact, change, communication	0.90
vicuna	entailments()	correct	contact, motion, communication	0.94
vicuna	entailments()	wrong	motion, communication, contact	0.91
vicuna	hyponyms()	correct	change, communication, contact	0.89
vicuna	hyponyms()	wrong	change, communication, contact	0.90
neural-chat	entailments()	correct	communication, motion, contact	0.91
neural-chat	entailments()	wrong	communication, motion, contact	0.92
neural-chat	hyponyms()	correct	change, motion, contact	0.89
neural-chat	hyponyms()	wrong	communication, contact, change	0.88
gemma	entailments()	correct	motion, communication, contact	0.92
gemma	entailments()	wrong	change, contact, communication	0.91
gemma	hyponyms()	correct	contact, motion, communication	0.91
gemma	hyponyms()	wrong	change, contact, communication	0.93

Moskvoretskii et al. (2024) the model is said to show “*strong zero-shot performance on lexical entailment with no fine-tuning*”. Our results would hence indicate the need for a deeper investigation of TaxoLLaMA on verb lexical entailment tasks.

Accuracy and Confidence w.r.t. the average score in HyperLex. We investigated whether the accuracy and confidence of the models can vary w.r.t. how the HyperLex verb-pairs were associated with their scores. The HyperLex scores consist of the average of the scores assigned by human annotators, on a scale from 0 to 6, based on the degree to which a lexical entailment relationship holds for any two verbs.

Table 5.10: Comparison of TaxoLlama with Llama-2 and Llama-3 (cf. Table 5.2): Zero-shot prompting results on WordNet verb pairs.

	<i>A</i>	<i>P</i>	<i>R</i>	<i>F1</i>	r-conf	nr-conf
Direct						
llama-3	0.571	0.862	0.169	0.283	9.94	9.95
llama-2	<u>0.422</u>	<u>0.363</u>	<u>0.205</u>	0.262	8.00	8.00
taxollama	0.167	0.248	0.329	0.283	<u>9.00</u>	<u>9.00</u>
Indirect						
llama-3	<u>0.334</u>	<u>0.392</u>	<u>0.605</u>	<u>0.476</u>	9.07	8.65
llama-2	0.670	0.623	0.860	0.723	<u>7.98</u>	<u>8.00</u>
taxollama	<u>0.500</u>	<i>NaN</i>	<u>0.000</u>	<i>NaN</i>	<u>10.00</u>	<u>10.00</u>
Reverse						
llama-3	0.605	0.579	<u>0.767</u>	<u>0.660</u>	9.74	9.67
llama-2	<u>0.539</u>	0.599	0.236	0.339	7.99	7.99
taxollama	0.500	0.500	1.000	0.667	<u>9.00</u>	<u>9.00</u>

Table 5.11: Comparison of TaxoLlama with Llama-2 and Llama-3 (cf. Tables 5.3-5.4): Few-shot prompting results on WordNet verb pairs: (top) HyperLex-FS, (bottom) Fellbaum-FS.

	<i>A</i>	<i>P</i>	<i>R</i>	<i>F1</i>	r-conf	nr-conf
Direct						
llama-3	0.660	0.831	<u>0.402</u>	0.542	9.71	9.30
llama-2	<u>0.358</u>	0.256	0.149	0.188	8.01	8.00
taxollama	0.284	<u>0.343</u>	0.474	<u>0.398</u>	<u>9.10</u>	<u>9.04</u>
Indirect						
llama-3	0.634	0.606	0.767	0.677	9.77	9.27
llama-2	0.048	0.087	0.095	0.091	8.00	8.00
taxollama	<u>0.265</u>	<u>0.305</u>	<u>0.367</u>	<u>0.333</u>	<u>9.31</u>	<u>9.09</u>
Reverse						
llama-3	0.617	0.958	0.244	0.389	<u>8.94</u>	<u>8.71</u>
llama-2	<u>0.539</u>	<u>0.602</u>	<u>0.231</u>	<u>0.333</u>	8.00	8.00
taxollama	0.514	0.567	0.121	0.200	<u>9.01</u>	<u>9.00</u>

	<i>A</i>	<i>P</i>	<i>R</i>	<i>F1</i>	r-conf	nr-conf
Direct						
llama-3	0.661	0.617	0.852	0.715	9.29	<u>8.54</u>
llama-2	0.290	0.271	0.248	0.259	8.00	<u>8.00</u>
taxollama	<u>0.328</u>	<u>0.366</u>	<u>0.469</u>	<u>0.411</u>	<u>9.26</u>	9.10
Indirect						
llama-3	0.508	0.505	0.761	0.607	9.87	9.57
llama-2	0.120	0.194	0.240	0.215	8.00	8.00
taxollama	<u>0.276</u>	<u>0.343</u>	<u>0.489</u>	<u>0.403</u>	<u>9.11</u>	<u>9.00</u>
Reverse						
llama-3	0.617	0.942	<u>0.251</u>	<u>0.396</u>	9.23	9.03
llama-2	0.526	0.523	0.585	0.552	8.00	8.00
taxollama	<u>0.528</u>	<u>0.586</u>	0.194	0.291	<u>9.01</u>	<u>9.01</u>

Table 5.12: Comparison of TaxoLlama with Llama-2 and Llama-3 (cf. Table 5.5): Zero-shot prompting results on HyperLex verb pairs.

	<i>A</i>	<i>P</i>	<i>R</i>	<i>F1</i>	r-conf	nr-conf
Direct						
llama-3	0.405	0.409	0.426	0.418	9.075	<u>8.476</u>
llama-2	<u>0.355</u>	0.353	0.347	0.350	8.000	8.000
taxollama	0.259	<u>0.341</u>	<u>0.518</u>	<u>0.412</u>	<u>9.000</u>	9.000
Indirect						
llama-3	0.271	<u>0.352</u>	0.542	0.427	<u>9.325</u>	8.996
llama-2	0.504	0.636	<u>0.018</u>	<u>0.036</u>	6.449	6.281
taxollama	<u>0.499</u>	0.000	0.000	NaN	10.000	9.997
Reverse						
llama-3	0.625	<u>0.673</u>	<u>0.487</u>	<u>0.565</u>	9.648	9.425
llama-2	<u>0.533</u>	0.727	0.105	0.184	7.968	7.961
taxollama	0.500	0.500	1.000	0.667	<u>9.003</u>	<u>9.000</u>

Table 5.14, resp. Table 5.15, report a summary of the average accuracy *A* and average confidence *r-conf* obtained by the models in the zero-shot, resp. few-shot, scenario for each of the HyperLex subsets corresponding to bins of the score interval, i.e., $[0.0, 1.0)$, $[1.0, 2.0)$, $\dots$, $[5.0, 6.0]$, hereinafter referred to as *bin0*, *bin1*, $\dots$, *bin5*. As it can be observed from both tables, there is no evident trend that relates the models' accuracy and confidence to the score bins. One exception would be represented by Llama3 and GPT-3: the former shows an average accuracy of around 0.36 in *bin0* and *bin1*, before increasing in the subsequent bins for the zero-shot scenario, while the latter exhibits a similar behavior for the few-shot scenario, with an average accuracy below 0.4 in *bin0* and *bin1* for Fellbaum-FS, and around 0.32 in *bin0* for HyperLex-FS, before increasing in the other bins.

5.5 Chapter Review

We investigated the capabilities of eight generative recent TLMs in recognizing lexical entailment relations among verbs through differently devised prompting strategies and zero-/few-shot settings over verb pairs from two lexical databases, namely WordNet and HyperLex. Our evaluation of lexical entailment recognition over WordNet verb pairs has unveiled that the examined TLMs can deal with the task showing moderate effectiveness on average. While no absolute winner emerges among the models, they tend to better understand Indirect than Direct in the zero-shot prompting, although in general the models can deal with all three types with comparable results on average. Interestingly, the skills of some models, especially under Direct and Indirect, tend to be improved by using HyperLex-FS and especially Fellbaum-FS. Llama-3 and GPT-3.5 reveal to be the models that mostly benefit from the few-shot prompting strategies. In terms of self-confidence, all models but Gemma

Table 5.13: Comparison of TaxoLlama with Llama-2 and Llama-3 (cf. Table 5.7): Few-shot prompting results on HyperLex verb pairs: (top) HyperLex-FS, (bottom) Fellbaum-FS.

	<i>A</i>	<i>P</i>	<i>R</i>	<i>F1</i>	r-conf	nr-conf
Direct						
llama-3	0.711	0.666	0.845	0.745	<u>8.896</u>	<u>8.286</u>
llama-2	<u>0.504</u>	<u>0.516</u>	0.126	0.203	8.000	8.000
taxollama	0.370	0.423	<u>0.716</u>	<u>0.532</u>	9.032	9.023
Indirect						
llama-3	0.557	0.537	0.826	0.651	10.000	9.947
llama-2	0.042	0.078	0.084	0.081	8.000	8.000
taxollama	<u>0.159</u>	<u>0.223</u>	<u>0.274</u>	<u>0.246</u>	<u>9.140</u>	<u>9.033</u>
Reverse						
llama-3	0.551	0.915	0.113	0.201	<u>8.048</u>	<u>8.035</u>
llama-2	0.499	0.499	0.997	0.665	8.000	8.000
taxollama	<u>0.516</u>	<u>0.561</u>	<u>0.145</u>	<u>0.230</u>	9.056	9.052

	<i>A</i>	<i>P</i>	<i>R</i>	<i>F1</i>	r-conf	nr-conf
Direct						
llama-3	0.620	0.572	0.947	0.714	<u>8.567</u>	<u>8.038</u>
llama-2	<u>0.497</u>	<u>0.485</u>	0.084	0.143	8.000	8.000
taxollama	0.338	0.402	<u>0.663</u>	<u>0.500</u>	9.019	9.016
Indirect						
llama-3	0.484	0.491	0.887	0.632	10.000	9.985
llama-2	0.138	0.216	0.276	0.243	8.000	8.000
taxollama	<u>0.280</u>	<u>0.297</u>	<u>0.321</u>	<u>0.308</u>	<u>9.432</u>	<u>9.256</u>
Reverse						
llama-3	0.605	0.955	0.221	0.359	<u>8.496</u>	<u>8.473</u>
llama-2	0.503	0.501	0.995	0.667	8.000	8.000
taxollama	<u>0.533</u>	<u>0.531</u>	<u>0.561</u>	<u>0.545</u>	9.333	9.332

Table 5.14: Zero-shot prompting based accuracy and confidence over correct answers w.r.t. the HyperLex score bins.

	<i>A</i>						r-conf					
	<i>bin0</i>	<i>bin1</i>	<i>bin2</i>	<i>bin3</i>	<i>bin4</i>	<i>bin5</i>	<i>bin0</i>	<i>bin1</i>	<i>bin2</i>	<i>bin3</i>	<i>bin4</i>	<i>bin5</i>
gpt-3.5	0.333	0.344	0.363	0.352	0.384	0.393	8.667	8.290	8.607	8.167	8.630	8.501
llama-3	0.363	0.365	0.519	0.486	0.505	0.539	9.431	9.236	8.997	9.004	9.123	9.274
llama-2	0.196	0.188	0.163	0.129	0.153	0.155	8.000	8.000	8.000	8.000	8.000	8.000
mistral	0.461	0.396	0.378	0.386	0.437	0.438	9.000	9.000	9.000	9.000	9.000	9.000
falcon	0.549	0.521	0.548	0.548	0.497	0.525	10.000	10.000	10.000	10.000	10.000	10.000
vicuna	0.863	0.854	0.874	0.876	0.839	0.863	9.000	9.000	9.000	9.000	9.000	9.000
neural-chat	0.402	0.385	0.333	0.386	0.495	0.447	10.000	10.000	10.000	10.000	10.000	10.000
gemma	0.431	0.385	0.511	0.448	0.524	0.516	2.000	2.000	2.000	2.000	2.000	2.000
taxollama	0.500	0.490	0.511	0.529	0.492	0.516	9.000	9.016	9.000	9.000	9.000	9.000

Table 5.15: Accuracy and confidence over correct answers w.r.t. the HyperLex score bins on (top) HyperLex-FS and (bottom) Fellbaum-FS prompting.

	A						r-conf					
	bin0	bin1	bin2	bin3	bin4	bin5	bin0	bin1	bin2	bin3	bin4	bin5
gpt-3.5	0.324	0.552	0.548	0.519	0.595	0.607	8.568	8.014	7.987	7.781	7.910	7.970
llama-3	0.500	0.469	0.511	0.600	0.619	0.699	9.321	9.182	9.402	8.882	8.844	8.887
llama-2	0.441	0.375	0.415	0.386	0.389	0.429	8.000	8.000	8.000	8.000	8.000	8.000
mistral	0.235	0.135	0.119	0.167	0.140	0.160	9.000	9.000	9.000	9.000	9.000	9.000
falcon	0.412	0.427	0.459	0.457	0.392	0.438	10.000	10.000	10.000	10.000	10.000	10.000
vicuna	0.412	0.333	0.393	0.414	0.431	0.461	9.000	9.000	9.000	9.000	9.000	9.000
neural-chat	0.127	0.083	0.059	0.038	0.130	0.132	10.000	10.000	10.000	10.000	10.000	10.000
gemma	0.118	0.083	0.081	0.105	0.119	0.110	2.000	2.000	2.000	2.000	2.000	2.000
taxollama	0.490	0.354	0.348	0.405	0.365	0.352	9.208	9.167	9.000	9.152	9.123	9.100

	A						r-conf					
	bin0	bin1	bin2	bin3	bin4	bin5	bin0	bin1	bin2	bin3	bin4	bin5
gpt-3.5	0.314	0.396	0.511	0.476	0.558	0.534	8.167	8.290	8.010	7.840	7.948	7.935
llama-3	0.608	0.552	0.600	0.671	0.725	0.776	9.624	9.154	9.040	9.068	9.177	9.212
llama-2	0.490	0.448	0.444	0.429	0.444	0.475	8.000	8.000	8.000	8.000	8.000	8.000
mistral	0.235	0.219	0.185	0.152	0.185	0.174	9.000	9.000	9.000	9.000	9.000	9.000
falcon	0.402	0.438	0.400	0.448	0.418	0.438	10.000	10.000	10.000	10.000	10.000	10.000
vicuna	0.392	0.375	0.363	0.410	0.421	0.416	9.000	9.000	9.000	9.000	9.000	9.000
neural-chat	0.147	0.094	0.081	0.067	0.146	0.151	10.000	10.000	10.000	10.000	10.000	10.000
gemma	0.118	0.083	0.081	0.105	0.119	0.110	2.000	2.000	2.000	2.000	2.000	2.000
taxollama	0.696	0.510	0.496	0.500	0.479	0.521	9.210	9.278	9.167	9.219	9.237	9.167

exhibit high values, regardless of being successful in solving the task or degenerating to a constant answer, like it can happen for Mistral, Gemma and NeuralChat under certain conditions.

Our evaluation on HyperLex data has shown behaviors of the models that are close to those observed on WordNet data. Apart from the generally lower values compared to the corresponding zero- and few-shot scenarios on WordNet data, there is a certain consistency of the models in terms of their more favorable or unfavorable settings according to the performance criteria as well as in terms of their self-confidence.

Our results have shown evidence of both abilities and limitations that arise in the examined TLMs, which can be summarized as follows: (i) although at varying degree of effectiveness and under different conditions, the TLMs can tackle a task of lexical entailment recognition with moderately good results, however, perfectly solving the task remains an unmet challenge for all the examined TLMs; (ii) few-shot prompting can improve the models' performance in addressing the task; and (iii) providing models with examples of entailment relation based on the Fellbaum types represents the best few-shot prompting strategy.

This work is focused on *general-purpose* models, as to date they represent the most widely used family of models in various NLP tasks. Nonetheless, models specifically designed for natural language understanding or inference might provide important insights into how TLMs address lexical entailment recognition. Future work might extend the scope of our evaluation to such models.

Our findings are based on WordNet and HyperLex data on verb relations. Although they are invaluable lexical resources, they cannot be regarded as exhaustive

in capturing all nuances of verb entailment relations; for example, they miss the specificity of sublanguages associated with particular domains, such as those of scientific fields or the legal domain. It would be interesting to extend our study to lexical entailment relations that characterize specialized domains as well.

The *Guidance* framework requires full access to the models to enforce grammar constraints effectively, such as with the `select` method. This works well with Open TLM, while closed TLMs, like the examined GPT-3.5, are only accessible via remote APIs, and hence do not support full integration with *Guidance*. Therefore, we avoided using a fixed grammar based on the `select` method and allowed GPT-3.5 to generate answers while enforcing it to meet our required format. To ensure valid outcomes by GPT-3.5, we eventually parsed the generated answers to extract the actual responses.

PART III: AI-BASED NLP METHODS AND APPROACHES FOR PSYCHOSOCIAL DIMENSIONS

The study of psychosocial dimensions has gained increasing importance in AI research, especially in the field of NLP. These dimensions encompass various aspects of human behavior, values, and social interactions, and they directly influence individuals' mental well-being, moral judgments, and societal relationships. By examining linguistic cues and contextual elements within text data, TLMs can help to analyze and understand complex psychosocial phenomena. This integration offers significant advancements in fields such as healthcare, ethics, and policy analysis, where the capacity to decode underlying psychosocial factors enables more tailored and impactful interventions across diverse fields.

A critical component of psychosocial well-being is mental-health, as it affects both individual and collective health outcomes. In recent years, one of the most pressing healthcare needs is early identification and treatment of mental-health issues, as well-recognized by the inclusion of mental-health in the Sustainable Development Goals.⁵

Another foundational aspect of psychosocial dimensions are moral values, which trace back to the earliest human societies where ethical principles were formalized into laws and social norms. These values serve to help individuals distinguish actions considered right from those considered wrong. They instill a sense of responsibility and encourage behaviors that would support the well-being of the community (Schwartz, 1992) or increase the division between people within society (Brady et al., 2020). This makes morality a distinctive element of individual character, which influences decisions and human behavior that can lead to a sense of integrity and self-esteem (Rana and Solaiman, 2022), and to implication of what is a morally good or bad life (Hofmann et al., 2014; Ellemers, 2018; Kádár et al., 2019).

⁵ (<https://www.who.int/health-topics/mental-health>)

Chapter 6

Transformer-based Language Models for Mental-Health-related Tasks

The automated recognition of mental-health diseases from EHRs, clinical notes biomarkers, writings and online posts has gained attention recently in the NLP field (Skaik and Inkpen, 2021; Kumar et al., 2021; Ahmed et al., 2022). NLP approaches can indeed be applied to discover early signs of mental disease, by typically treating this task as a classification problem, such as sentiment analysis.

TLMs have been widely used for addressing health-related (González-Fernández et al., 2021) as well as mental-health issues. However, to the best of our knowledge, the use of TLMs for addressing mental-health issues has not been surveyed in depth so far. We aim to fill this gap in the literature by pursuing the following main objectives (Greco et al., 2023): (i) to detect all types of TLMs that have been used for text-based mental-health issue identification, (ii) to characterize the methods based on such models that have been developed in the past few years, according to the type of health problem involved and the data source domain, (iii) to understand how existing methods are able to effectively address some major mental-health issues, and hence unveil current limitations and rooms for improvement of methods based on TLMs.

Our presentation of the studies selected for this overview is organized into three subsections. Methods are mainly divided into two categories, i.e., domain-adapted training and non-domain-adapted training, where the latter includes both as-is use of a TLM in inference modality and task-adaptive fine-tuning modality. This organization enables us to unveil whether and what advantages domain-specific pre-training had on the mental-health domain against approaches that use TLMs pre-trained on a general domain language. In general, the role of TLMs in each approach is to address the specific task or to obtain contextual dense representations of the text. Approaches presented at the CLEF eRisk competition are discussed apart.

6.1 Methods

6.1.1 Non-domain-adapted methods

Malviya et al. (2021) deal with the early detection of depressive states on Reddit. This is managed as a binary classification problem, where a post is categorized as depressive or not. The dataset is available on Kaggle and consists of 6k training posts and 2k testing posts. Several TLMs are fine-tuned and evaluated on this task, in particular BERT, DistilBERT, RoBERTa, ELECTRA and XLNet. All of them reach very high performances on the validation set, in terms of accuracy (above 0.96) and F1-score (above 0.97).

Zhang et al. (2021a) concatenate word embeddings based on GloVe and the hidden outputs of a Transformer Encoder to obtain an input representation consisting of both semantic information and contextual information. These are then encoded by a BiLSTM, in order to capture long-term dependencies and context-aware information, and the resulting output is then used for the identification of suicide notes as a classification problem. They collected suicidal writings from existing corpora ((Schoene et al., 2019), about 1k samples) as well as neutral writings from Reddit posts (neutral subreddits, about 2k samples). The resulting corpus is split into 70% for training set, 15% for validation set and 15% for test set. The method achieves average F1 score of 0.95.

Chaurasia et al. (2021) deal with forecasting the mental-health of scholars, particularly depression and frustration, by monitoring the unpleasant phrases or offensive slang that they may intentionally use in their tweets. For each user, a bag-of-words representation of words expressing anger, frustration, depression, scholar, and stress and a LDA model representation of her/his tweets are first produced. Such representations are then independently provided in input to BERT to yield word embeddings that are averaged to finally obtain two sentence embeddings. The suspected frustrated scholar, i.e., the target user, is identified if the similarity score between the two sentence embeddings is above a certain threshold, for a number of her/his tweets. Those tweets are fed to a classification model to detect the emotion in the text. The authors tested a linear support vector machine (LSVM), random forest (RF) and multinomial naive Bayes (MNB) model as classifier. The task is to assign a tweet with an emotion across sadness, worry, neutral, happiness, joy, and relief.

Table 6.1: Main characteristics of the methods discussed in this work. Domain column reports the following abbreviations: SM (social media), SN (suicidal notes), LS (last statements), TM (text messaging), CN (clinical notes), BM (biomedical literature). The class labels are specified in the Task column (e.g., diagnosis vs. control), otherwise they can be inferred from the union of Task and Problem columns (e.g., if the task is binary-classification and the problem is depression, then the class labels are depressed vs. non-depressed).

Ref.	Method/Is Pre-Trained?	Domain	Problem	Data sources	Task
Malviya et al. (2021)	BERT, DistilBERT, RoBERTa, Electra, XLNet / No	SM	depression	Reddit	Binary-Classification
Zhang et al. (2021a)	TransformerRNN / No	SN, LS	suicide	suicide notes, Reddit, notes of sentenced to death	Binary-Classification
Chaurasia et al. (2021)	BERT / No	SM	depression, frustration	Twitter	Emotion Detection (sadness, worry, joy,...)
Murarka et al. (2020)	RoBERTa / No	SM	depression, anxiety bipolar disorder, ADHD and PTSD	Reddit	Problem Type Detection
Haque et al. (2020)	BERT, ALBERT, RoBERTa, XLNET / No	SM	suicidal ideation	Reddit	Binary-Classification
Jiang et al. (2020)	BERT-ATT / No	SM	borderline personality, depression, PTSD, anxiety, schizophrenia, eating disorder, obsessive compulsive disorder	Reddit	Binary-Classification (diagnosis vs. control) Problem Type Detection
Zeberga et al. (2022)	Bi-LSTM / No	SM	depression and anxiety	Reddit, Twitter	Binary- Classification
Matero et al. (2022)	RoBERTa / No	SM	depression	Facebook	Problem Severity
Ji et al. (2021)	mentalBERT / Yes mentalRoBERTa / Yes	SM TM	depression, suicidal ideation, stress anxiety, bipolar disorder	Reddit, Twitter, SMSs	Problem Type Detection
Meng et al. (2021)	BRLTM / Yes	CN	depression diagnosis	heterogeneous EHR data	Binary-Classification
Vajre et al. (2021)	psychBERT / Yes	BL	depression, anxiety, social anxiety,	Reddit, Twitter	Binary-Classification

Continued on the next page

Ref.	Method/Is Pre-Trained?	Domain	Problem	Data sources	Task
		SM	loneliness, mental illness, suicide		Problem Type Detection
Naseem et al. (2022)	PHS-BERT / Yes	SM	depression, suicide, stress, anorexia	Reddit, Twitter	Binary-Classification Problem Type Detection
Rawat et al. (2022)	ScANER / Yes	CN	suicidal ideations and attempts	EHR	Binary-Classification (SA vs SI) Problem Type Detection
Liu et al. (2021d)	Longformer / Yes	TM	mental or emotional distress	Shout	Problem Type Detection
Lokala et al. (2022)	GeM / Yes	SM	depression, anxiety, bipolar, PTSD	Reddit	Gender and Problem Type Detection
CLEF eRisk competition (2020, 2021)	BERT, RoBERTa, XLM-RoBERTa, SpanEmo DistilRoBERTa, S-BERT / No	SM	pathological gambling, self-harm, depression	Reddit	Early Detection of Signs of Pathological Gambling Early Detection of Signs of Self-Harm Measuring the severity of the signs of depression

The models are trained on a dataset containing labeled tweets (30k tweets as training set, 10k tweets as test set, and the target tweets of seven scholars as validation set). Results are evaluated using precision, recall and F1 scores for each class, and overall precision and accuracy scores. The RF classifier turns out to be the best classifier, with precision, recall and F1 scores higher than 0.9 for each emotion, overall precision and accuracy of 0.96.

Murarka et al. (2020) approach mental illness as a multi-class classification problem, focusing on depression, anxiety, bipolar disorder, attention deficit hyperactivity disorder, and post-traumatic stress disorder. The method is based on RoBERTa fine-tuned on 13 crawled subreddits and 17159 posts' body texts and headline texts. Five subreddits have a direct connection to the analyzed mental disorders. Results confirm the potential of social media to support therapeutic practices in the prediction of mental-health across two major user groups, i.e., those who either are looking for help online or are unaware of their condition. In fact, the performance of the LSTM model trained on posts is matched by the RoBERTa model trained on the titles only. The F1 score for posts and posts+titles for the RoBERTa model is 0.86 and 0.89, respectively.

Haque et al. (2020) are among the first to explore the idea of using TLMs for suicidal ideation detection. They compare BERT, ALBERT, ROBERTA, and XLNET models on the task of suicidal ideation detection, using Reddit data containing 3549 suicidal suggestive texts (from the SuicideWatch subreddit) and non-suicidal texts. In particular, the word-embeddings derived from TLMs are provided to a softmax layer for classification purpose, i.e., to predict whether or not a post contains suicidal ideation. RoBERTa turns out to be the best performer, with 0.84 accuracy, 0.8 precision, 0.91 recall, and 0.85 F1 score.

Jiang et al. (2020) address the identification of eight types of psychiatric disorders, namely schizophrenia, borderline personality disorder, post-traumatic stress disorder, eating disorder, major depressive disorder, general anxiety disorder, obsessive compulsive disorder, and bipolar disorder. To this end, more than 35k Reddit users are considered that include self-reported ones having mental-health disorders and a control user group (i.e., users with none of the aforementioned eight conditions). A REALM-like model and an attention-based model (which uses BERT representation of each post as input) are trained for a user-level classification task, to determine if a user has a certain diagnosis based on an aggregated representation of their posts. Moreover, a BERT model is fine-tuned to conduct post-level classification experiments. The training set is composed of 50k mental group posts and 50k control group posts, whereas 5k mental group posts and 5k control group posts are used as validation set, and analogously other 10k posts as test set. In addition to binary classifiers (diagnosis vs. control), an ensemble of such classifiers is built as a multi-label classifier for different diagnostic groups. For user-level binary classification, the authors consider two types of data settings, where the difference is the presence of mental-health sensitive posts (uncleaned version, UNCL) or not (cleaned version, CL). In the CL setting, the attention-model reaches best results, with accuracy and F1 scores higher than 0.8, while REALM-based model obtain accuracy and F1 scores higher than 0.96 for the UNCL setting. As regards the post-level results, the

BERT-based model reaches accuracy between 0.56 and 0.7 and F1 scores between 0.57 and 0.74. Finally, for the multi-label ensemble classification, the attention-based classifier achieves F1 scores between 0.19 and 0.21.

Zeberga et al. (2022) classify Reddit and Twitter posts to identify anxiety and depression using two approaches. The first one uses BERT to produce embeddings for a BiLSTM model as a classifier. The second one exploits knowledge distillation from a fine-tuned BERT to a smaller model, named *distilled_BERT*. Both models are trained on a depression-related dataset (95k posts) and an anxiety-related dataset (100k posts), and then evaluated on both depression- and anxiety-related test sets against baselines such as CNN, LSTM and SVM models; all competing models also used word2vec and GloVe embeddings alternatively to BERT. Results show that both the approaches outperform the baselines, especially *distilled_BERT*, with a 0.97 accuracy score in the depression-related dataset and a 0.98 accuracy score for the anxiety-related dataset.

Matero et al. (2022) consider the degree of depression that is determined by responses to online surveys. In particular, the study regards Facebook users who agreed to disclose their status changes between 2009 and 2011, and that answered questions about their personalities. A training set of 17k users and a test set of 986 users are obtained by selecting people who posted status updates of at least 1000 words. Here, experimental evaluation is focused on the effectiveness of the user representations derived from individual model's layers, as well as various combinations of representations from multiple layers, where a regularized linear regression algorithm is used as model of choice, with the mean aggregate of contextual embeddings as input. Also, a 10-fold cross-validation technique is employed to tune the regularization parameter, for each single layer and for a combination of layers, to determine which layers include in the final model. The models are evaluated using mean squared error (MSE) and disattenuated Pearson r (r_{dis}). Among the selected TLMs, i.e., BERT, RoBERTa, XLNet, and ALBERT, RoBERTa has shown to be the most effective with layer 19 (the sixth-to-last) performing at least as good as layer 23 when utilizing just one layer. As regards multiple layers, best results are achieved using layers 16, 18, 19, 22 and 24, with 0.554 MSE and 0.721 r_{dis} .

6.1.2 Domain-adapted models

Ji et al. (2021) release *mentalBERT* and *mentalRoBERTa*, the first pre-trained language models for mental disorder detection based on BERT and RoBERTa, respectively. The pre-training phase is conducted by further pre-training a checkpoint of the models on the specific downstream domains. The pre-training corpus is collected from Reddit and includes a number of subreddits for a total of about 13M sentences. The models are evaluated on various benchmarks, such as SWMH, containing posts from subreddits (34k, 8k, and 10k for the training, validation, and test set, respectively), T-SID, containing posts from Twitter with suicidal ideation (3k, 768, and 685 for the training, validation, and test set, respectively), UMD, containing posts from

Reddit with suicidal ideation (993, 249, and 490 for the training, validation, and test set, respectively), and SAD, which includes SMS-like sentences from different sources (e.g., chatbot-based conversations) expressing stressful states (5k, 617, and 685 for the training, validation, and test set, respectively). The two models are evaluated by recall and F1 score, and compared with the original BERT and RoBERTa as well as BioBERT (Lee et al., 2020) and ClinicalBERT (Huang et al., 2019) specialized on biomedical domain and clinical notes, respectively. In most cases, mentalBERT and mentalRoBERTa obtain the best scores (e.g., 0.89 recall and F1 by mentalRoBERTa when tested on T-SID dataset, and 0.64 recall by mentalBERT when tested on UMD), demonstrating that further pre-training on the mental-health domain can improve performance in classification tasks.

Meng et al. (2021) propose a novel pre-trained BERT-based model, named *BRLTM*, to deal with the prediction of depression diagnosis, focusing on patients affected of myocardial infarction (MI), breast cancer (BC), and liver cirrhosis (LC). The model is able to process five heterogeneous EHR data sources related to patients' visits. Like BERT, the pre-training phase addresses an MLM task on EHR sequences. In the fine-tuning phase, a classification head is added to perform the predictions. The pre-training, resp. fine-tuning, data include about 10k, resp. 3k, input sequences from patients with MI, 23k, resp. 5k, from patients with BC, and 11k, resp. 2k, from patients with LC. Also, the fine-tuning data is divided in 70% samples for training set, 20% samples for validation set and 10% samples for test set. Data are processed in a chronological order based on four prediction windows, i.e., two weeks, three months, six months, and one year before the time of depression diagnosis. As regards the input representation of the model, the token embeddings contain information from diagnoses, procedures, medications and clinical notes. The demographic information is added to each token in the sequences in terms of age and gender embeddings. Like BERT, position and segment embeddings are included. In order to encourage the support of such models in the health domain, interpretability aspects of *BRLTM* are also discussed, showing that the self-attention mechanism can quantify the inner relationships across the heterogeneous EHR data. The models are evaluated in terms of receiver operating characteristic (ROCAUC) and precision-recall area under curve (PRAUC). The model proves to reach high performance on the task for different prediction windows (e.g., 0.85 ROCAUC and 0.76 PRAUC for the breast cancer diagnosis in a two weeks window), demonstrating the superiority of bidirectional learning w.r.t. unidirectional language models as well as the power of the two-stage transfer learning approach (i.e., pre-training and fine-tuning phase) when limited amounts of data are available.

Vajre et al. (2021) propose *PsychBERT*, a BERT-based model pre-trained on mental-health texts taken from biomedical literature (40k PubMed articles related to mental illnesses) and social media platforms, specifically from Reddit and Twitter conversations discussing mental-health diseases as well as other non-mental-health topics. The data are annotated w.r.t. 6 labels (depression, anxiety, social anxiety, loneliness, mental illness, and suicide) by mapping the signs and symptoms of the Hierarchical Taxonomy Of Psychopathology (HiTOP) to hashtags and subreddits. The data are used for both binary classification (148k, 21k, and 42k for the training,

validation, and test set, respectively) and multi-class classification (84k, 10k, and 10k for the training, validation, and test set, respectively). PsychBERT is based on a two-stage framework: in the first stage, PsychBERT determines if a conversation describes a mental-health disease or not (binary classification), while in the second stage PsychBERT along with a multi-class classifier assign one of the 6 labels to the posts with a mental disease. In both stages, the classification performance of PsychBERT achieves the best F1 score (0.98 in the first stage, and 0.63 in the second stage). In addition, the authors investigate methods based on backpropagation to enhance the interpretability of the model.

Naseem et al. (2022) propose *PHS-BERT* for public health surveillance (PHS) tasks. The model is obtained by further pre-training BERT on tweets related to mental-health topics, and then fine-tuning it to binary and multi-class classification tasks related to various mental diseases such as depression, suicide, stress, anorexia as well as other health-related topics. To this purpose, 25 datasets are used that contain posts from Reddit and Twitter, such as R-SSD (500 samples, 5 classes) for suicide risk detection, SAD (6850 samples, 2 classes) for identification of stressful situations, Depression.Reddit.1 (3553 samples, 4 classes) and Depression.Twitter.1 (1793 samples, 3 classes) for early detection of depression states. The performance of PHS-BERT is compared to general-purpose TLMs as well as to BioBERT and MentalBERT. Experimental results show that PHS-BERT achieves F1-score higher than the baselines in all tasks and datasets.

Rawat et al. (2022) focus on suicide risk detection from EHR notes. They construct the *Suicide Attempt and Ideation Events* (ScAN) dataset, which is a subset of the MIMIC III data, containing about 12k EHR notes of patients in hospital-stay with 19k expert-annotated events related to suicidal ideations (SI) and attempts (SA), along with attributes indicating methods of suicide adopted. In addition, they release a novel TLM pre-trained on such corpus, named *Suicide Attempt and Ideation Events Retriever* (ScANER). This consists of a retrieval module followed by a predictor module. The retrieval module is obtained by further pre-training RoBERTa on MIMIC dataset and aims to detect sentences with suicidal evidences as a binary classification problem at paragraph level. The predictor module is a multi-head attention model that identifies the type of the diagnosed suicidal behavior (i.e., SA or SI). ScANER is trained on 40k paragraphs, validated on 7k paragraphs and tested on 11k paragraphs of ScAN. The retrieval module detects suicidal evidences with a macro-weighted F1-score of 0.83, while the predictor module obtains a macro F1-score of 0.78 for SA detection and a macro F1-score of 0.60 for SI detection.

Liu et al. (2021d) analyze volunteers' compliance with conversational formats and procedures to learn more about user groups' demographics and behaviors that indicate distress. Shout data are used for this study, which include 10M messages and 271k conversations of various lengths, from which 169 conversations (8844 messages) were sampled and annotated using six behavior labels, namely setting intention, enquiring, expressing distress, reflecting, corresponding, and discord. Since most conversations are longer than 512 tokens, the Longformer model is utilized and further pre-trained with MLM task on the whole pre-processed Shout dataset. Moreover, data augmentation is performed to improve the training. Three tasks are

targeted: (i) to assess if the training conversational structure is being followed by the Shout participants, (ii) to examine the volunteers' behaviors and texters' emotions of distress, and (iii) to gain insights into the user-base for Shout evidencing underrepresented groups. The proposed method selects three demographic groups for prediction, i.e., people under 13, autism diagnosis, and non-binary gender, as they are known to be at higher risk for mental-health issues. Results show good performance for the conversational stages (0.88 weighted accuracy), behavioral patterns (0.95 1-hamming loss), and texter demographics (0.96 weighted accuracy).

Lokala et al. (2022) propose a bi-encoder model to detect mental-health symptoms on patients with cardiovascular diseases, evaluating the impact of gender in the identification of mental-health illnesses. To this end, the model is evaluated on two tasks, symptom identification and gender identification, using 16k posts and comments from Reddit annotated w.r.t. four mental-health symptoms and two gender labels. The model, named *GeM* (Gender Enabled MH framework), is defined as a task adaptive framework with knowledge aware entity masking, i.e., symptom and gender masking using ontologies as external knowledge. The two encoders, the one focusing on symptoms (S-encoder) and the other on gender (G-encoder), are initialized using weights from RoBERTa and further pre-trained with a task-adaptive approach on 65k posts and 150k comments from Reddit. The pre-training task is MLM as in BERT. After that, an attention score is computed on the generated encodings to obtain a mental-health context-aware representation of gender text, where the query is the gender encoding while the key/value is the symptom encoding. The dataset is divided into 75% training set, 5% validation set and 20% test set. The model outperforms baselines based on TLMs (BERT, GPT-2, RoBERTa) and LSTM in both symptom and gender identification tasks, with both F1 scores of 0.80.

6.1.3 Methods for CLEF eRisk competition

eRisk@CLEF¹ is a competition organized since 2017, that explores the evaluation methodology, effectiveness metrics and practical applications of early risk detection on the Internet, particularly those related to health and safety. eRisk@CLEF2021 challenge considers the following tasks: *Task 1 - Early Detection of Signs of Pathological Gambling*, *Task 2 - Early Detection of Signs of Self-Harm*, and *Task 3 - Measuring the severity of the signs of depression*, through a questionnaire of 21 questions with multiple answers. The data source is a collection of writings (posts or comments) from social media users. eRisk@CLEF2020 has only two tasks, which correspond to the second and third of the 2021 edition. Task 1 and 2 are evaluated through decision-based and ranking-based criteria. In the first case, the binary decisions taken for each user are considered, and the models are evaluated using various metrics, such as precision, recall, F1 and ERDE (an error measure that introduces a penalty when late correct alerts occur). In the second case, the score that represents

¹ <https://erisk.irlab.org/>

the user's level of risk is evaluated, which is ranked using standard IR metrics (e.g., P@10, NDCG) and considering the results by the systems after seeing k writings of the same user. Task 3 is evaluated using specific metrics that estimate the quality of the answers given by the proposed model w.r.t. the answers given by a user. This section overviews the methods presented at eRisk 2021 and 2020 that use TLMs.

Achilles et al. (2020) address the 2020 Task-1 by pre-training a BERT base model, followed by four parallel stacks of convolutional neural networks with max pooling layers on top. The concatenated output is processed by a LSTM layer followed by a dense layer in order to obtain a final score. The work also presents an ensemble system based on different techniques. Martinez-Castaño et al. (2020) participate in both tasks of eRisk 2020. For Task 1, all proposed models are based on XLM-RoBERTa, which turns out to be the best performing approach. The authors use data from the Pushshift Reddit Dataset collected through its public API. Regarding Task 2, the team uses the original training dataset provided by the organizers, for the training of RoBERTa and XLM-RoBERTa models. Uban and Rosso (2020) use TLMs and an ensemble model for the 2020 Task-1. In particular, they fine-tune a pre-trained BERT base model with a sequence sampling strategy for data augmentation. Alhuzali et al. (2021) propose an approach based on pre-trained models to extract features from users' posts plus a random forest classifier on top, for the 2021 Task-3. For the training phase, they use the 2019 and 2020 datasets provided by the organizers and ELMo, BERT, and SpanEmo models for feature extraction. Barros et al. (2021) use the standard dataset to train two models for the 2021 Task-2, the first one is a BERT model with an SVM classifier and the second one is an improved version introducing two new features: emojis and sentiment analysis through VADER (Hutto and Gilbert, 2014), a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media. Basile et al. (2021) address all 2021 tasks. The test data for Task 1 is collected from Reddit, since no training data was released by the organizers, while for Task 2 the team uses Reddit data in addition to the existing one. For Task 1, they train a text classifier over BERT. Two approaches are designed for Task 2, one of which uses a TLM to classify each post individually and then to predict a label for each user, based on the probabilities assigned to their writings. Also for Task 3, one of the two proposed approaches uses two classifiers that are trained based on RoBERTa and DistilRoBERTa over external data from Reddit. Bucur et al. (2021) present automatic data crawling from mental-health subreddits and employ pre-trained BERT on all 2021 tasks. For Task 1, data are collected from 2111 Reddit users crawling their posts even from unrelated subreddits to avoid gambling posts. The crawling for Task 2 is used to increase the size of the original dataset. The same BERT architecture is fine-tuned for Task 1 and Task 2 producing models by different combinations of training data. Finally, for Task 3, both user post text embeddings and questionnaire-answer text embeddings are produced by the BERT encoder, which is fine-tuned through contrastive learning. Inkpen et al. (2021) employ multiple deep learners for the 2021 Task-3 to classify information extracted from postings after a preliminary selection phase to filter only posts that are relevant to each question. S-BERT and S-RoBERTa are used for deep contextual post representations, and models are trained for each of the

questions. Maupomé et al. (2021) train a RoBERTa model for the 2021 Task-2. The produced embeddings are averaged to get the document representations. Documents are combined with a function which gives more importance to recent posts than to past ones. Spartalis et al. (2021) address the 2021 Task-3 using both feature-based and fine-tuning strategies. For the feature-based approach, various S-BERT models are used to extract post and subjects embeddings in combination with non-neural classifiers. In the other strategy, a RoBERTa model is fine-tuned for a classification task. Nisa and Muhammad (2021) participate at the 2021 Task-2 using BERT to generate word embeddings as document representations for logistic regressors. Wu and Qiu (2021) describe a RoBERTa-based approach on the 2021 Task-3 focusing on post-processing weighting mechanisms to mitigate the tendency of the systems to give lower level of severity. A pre-trained RoBERTa model is used for each question to extract the final four levels as input for a linear classifier on top.

Overall, the result report provided by the organizers of the eRisk@CLEF competitions (erisk.irlab.org) show that, in the 2020 Task-1, TLMs achieve the best decision-based and ranking-based performances, and in particular Martínez-Castaño et al. (2020) obtain top results according to most of the assessment criteria. This trend has been consolidated in the 2021 competition, since TLMs achieve top results on Task-2 and Task-3 based on several assessment criteria.

6.2 Chapter Review

This research investigated TLMs-based approaches designed to address mental-health issues. The current landscape appears to mostly explore the identification of symptoms of depression and suicidal thoughts, although there is a bunch of studies dealing with multiple mental illnesses. While some methods are limited to simply fine-tune a TLM to mental-health related texts, other methods employ domain-adaptations by further pre-training a TLM, thus exploiting the knowledge of TLMs on the general domain and, at the same time, specializing this knowledge also for mental diseases. Also, some works at the CLEF eRisk competition use TLMs to get embeddings for similarity calculations, while others use data-augmentation and ensemble strategies. The majority of the datasets involved in the studies corresponds to social media posts, while clinical/suicidal notes and text messaging conversations are used marginally. Nonetheless, the number and type of data sources used is varying over the studies.

Major findings from the selected methods in this survey can be summarized as follows: (i) non-domain-adapted, i.e., fine-tuned, methods behave as good as conventional methods in the tasks at hand; (ii) the domain-adapted methods are able to improve significantly upon non-TLMs based methods, and in some cases further pre-training reveals to be beneficial for increasing performance w.r.t. non-domain-adaptation; (iii) in some cases, the use of external knowledge bases (i.e., ontologies, taxonomies) has a relevant impact on the model performances, since it enhances the domain knowledge of TLMs.

We also notice a number of weaknesses of the existing approaches, which represent room for improvements and further study. Firstly, few approaches offer insights into *explainability* and/or *interpretability* of the language models, highlighting how these features are relevant to increase the trust of artificial intelligence tools to support humans in decision-making processes. Secondly, most of the mental-health tasks are reduced to a classification problem (e.g., sentiment analysis), but the potential of TLMs can be exploited for many other purposes, such as entailment, question answering, entity recognition. Furthermore, a little discussion is provided to avoid or mitigate the *bias* of TLMs, in terms of different aspects of the users/patients, such as culture, age and gender. Our study has also a number of limitations. We focused our investigation on a high-resource language (English), however, since mental-healthcare is of utmost importance worldwide, it is clearly essential to explore the use of TLMs for low-resource languages as well. Also, we focused on textual data sources, however more advanced investigation should include a multimodal or cross-modal analysis, since social media and EHRs provide not only texts, but also images and multimedia data.

Chapter 7

Transformer-based Language Models for Moral-Foundation-Theory-related Tasks

Although the linguistic choices of individuals might reflect their moral values, the connection between language and morality is not always clear and may be expressed differently across linguistic patterns (Kennedy et al., 2021). Despite this complexity, TLMs—by learning lexical, semantic, and factual information from a large corpus of training data—inevitably capture a wide range of cultural and moral biases embedded within the language (Abdulhai et al., 2023; Hämmerl et al., 2023; Schramowski et al., 2022). Therefore, they can serve as effective tools for capturing and analyzing moral dimensions. In this respect, one emerging trend is to integrate TLMs with moral information to infer moral foundations from input data and/or analyze the multifaceted moral dimensions embedded in language models.

In this respect, one of the emerging trends is to integrate TLMs with moral information for inferring moral foundations from the input data and analyzing the moral dimensions embedded in TLMs. These approaches are beneficial for applications in social sciences, policy-making, and even in enhancing AI ethics. For example, TLMs can be employed to execute moral judgment tasks like moderating content on social platforms (Franco et al., 2023); or to analyze political speeches and campaigns, as well as business strategies, identifying potential manipulative tactics (Liu et al., 2022; Luceri et al., 2024). Hence the need to study AI systems for morality becomes paramount. Furthermore, recent studies have shown that TLMs can be trained to recognize and reflect moral judgments, or can be inspected to reveal certain moral tendencies present in the pre-training datasets (Guo et al., 2023a; Preniqi et al., 2024; Zhang et al., 2023). Therefore, the study of methodologies integrating TLMs and moral foundations can provide critical insights into how TLMs address task related to principles that govern human behavior and how they perceive moral values. This can also foster a form of transparency that becomes the basis for user’s confidence in AI systems

We provided a systematic overview of the recent advances in the integration of moral foundations within TLMs, from the perspective of the MFT (Zangari et al., 2024). Nevertheless, the association between language and moral foundation theory is complex and variable, since moral values often manifest indirectly or contextually. Furthermore, moral values alone may not fully capture the complexity of human

behavior as expressed by individuals Kennedy et al. (2021). The aim of our work is to analyze the current approaches to combining the MFT and TLMs to systematize and provide valuable insights into how TLMs perceive and generate relevant content within a moral framework. Contributions are summarized as follows:

- **New taxonomy:** We provide a comprehensive overview of the TLMs involved in the context of the MFT. Additionally, We formalize the essential strategies and tasks adopted within the domain of MFT-informed TLMs. Based on our theoretical framework, We organize the existing works in the literature into two main categories: (i) *Moral-driven TLMs*, which focus on training TLMs on MFT-based data; and (ii) *Moral-targeted TLMs*, which analyze the moral foundations embedded in TLM responses at the inference phase.
- **Moral Surveys, Lexicons, and Datasets:** We review and categorize the most commonly used datasets, surveys and lexicons for training and assessing TLMs on MFT-related tasks.
- **Findings, Challenges, and Future Directions:** We discuss key findings from the reviewed works, highlight ongoing challenges in the field, and suggest future research directions. We emphasize the importance of constructing reliable and inclusive datasets, along with the pivotal role of TLMs in the context of MFT. This includes considering aspects such as the model size and the strategies employed for training or fine-tuning the TLMs.

By providing a structured overview of the intersection between the fields of TLMs and MFT, this survey aims to bridge the gap between moral psychology and advanced computational linguistics, laying the groundwork for future research and development in the creation of morally aware AI systems.

To the best of our knowledge, this is the first effort to provide a comprehensive and systematic review dedicated to exploring the combination of MFT and TLMs. By bridging the gap between moral psychology and advanced computational linguistics, this work aims to lay the groundwork for future research and development in creating morally aware AI systems.

7.1 Related Works

Recent advancements in the field of moral values within TLMs include the design of algorithms capable of detecting morally relevant contents and the implementation of procedures or evaluation metrics that reflect the moral perceptions of TLMs. However, the existing literature lacks comprehensive reviews that explore the intersection of MFT and TLMs within a computer science perspective. Existing surveys mainly focus on social science, evaluating the integration of TLMs within human society from a broader outlook. For instance, Adilazuarda et al. (2024) investigate cultural values representation and inclusion in TLMs, highlighting the absence of a consistent definition of culture and employing different proxies to represent cultures. They categorized two broad approaches to studying cultural values inside TLMs: observing

the response of the TLMs to various inputs and observing the internal state of TLMs. Wang et al. (2023c) review the existing alignment tools for TLMs, with emphasis on training approaches and model assessment strategies, pointing out the limits of TLMs in humans, such as misreading human commands and producing biased or erroneous material. Guo et al. (2023b) offer a review of the evaluation methods of TLMs, including aspects related to morality, such as the typically used datasets related to the MFT. Similarly, Chang et al. (2024) provide an overview of the evaluation methods for TLMs in terms of robustness, ethics, biases, and trustworthiness. Gallegos et al. (2023) summarize bias evaluation and mitigation techniques for TLMs, expanding the notion of fairness and harm to improve the fairness capabilities of TLMs. The work that most closely aligns with ours is Reinig et al. (2024), which provides a broad review on modeling morality in text analysis, covering various moral theories and methodologies such as Transformer models, dictionary-/rule-based approaches, and machine learning techniques. The authors focus on the theoretical challenges in operationalizing moral values, highlighting issues such as low inter-annotator agreement and the scarcity of resources available for training and evaluation, particularly in non-English languages. The presence of such work reflects an emerging interest in the intersection of moral values and AI, confirming that the recent research is gradually recognizing the need to address morality within computational models.

Differences with existing surveys. We systematically review works that are framed within the context of TLMs and the MFT. Compared to other works based on the combination of TLMs and human psychological aspects, such as Adilazuarda et al. (2024), We do not focus solely on reviewing methods that observe the responses or the status of the TLMs, but We also consider fine-tuned or further pre-trained TLMs for detecting moral values. Similarly to Wang et al. (2023c), We aim to review the evaluation and training strategies, but for the specific area of MFT and TLMs. Guo et al. (2023b), Chang et al. (2024) and Gallegos et al. (2023) investigate moral values only from an evaluation perspective and within the much broader scope of ethics, safety, and AI alignment.

In contrast to (Reinig et al., 2024), our work focuses specifically on the integration of MFT and TLMs. While Reinig et al. (2024) offers a high-level overview of the research connecting morality and AI, our study provides a detailed examination of each individual contribution. Moreover, We introduce a taxonomy that categorizes works based on moral-driven and moral-targeted TLMs, while also exploring the involved architectures and task adaptation strategies. Additionally, We place greater emphasis on cross-domain generalization and the moral biases inherent in TLMs, offering a detailed exploration of their role in MFT-related tasks.

Despite the increasing interest in facing moral values within TLMs, there is a lack of comprehensive surveys exploring the integration of MFT and TLMs. Existing works often addresses MFT and PLMs separately or marginally discuss their combination. Our work fills this gap by providing the first systematic review dedicated to exploring how MFT principles can be operationalized within TLMs. We introduce a novel taxonomy categorizing approaches into moral-driven and moral-

targeted TLMs, delve into specific architectures and training strategies, and offer an in-depth analysis of relevant datasets and lexicons.

7.2 Methods

We organize existing studies into two main categories. The first one regards **Moral-driven TLMs**, which refer to works that adopt training strategies for learning model that predict or are aligned with moral values. The general approach here is to fine-tune or further pre-train TLMs for predicting the moral values of the underlying moral foundations hidden in the data. The second category, namely **Moral-targeted TLMs**, regards methods that seek to assess the moral values embedded into the TLMs without any moral task-specific adaptation, to probe their responsiveness to morally challenging issues and situations. The general approach is to analyze the TLM responses to determine their beliefs and the effects of training on a particular dataset. Fig. 7.1 illustrates how these two categories intertwine with various aspects of TLMs in the context of MFT, including MFT-related tasks, data types, model architectures, and problem-solving strategies. Solid lines represent hierarchical (parent-child) dependencies, while dashed lines indicate relational dependencies. Examining these interconnections within each category reveals correlations and patterns that help structure and describe the intersection of TLMs and MFT.

Table 7.1 shows the selected works for this survey. The table is partitioned into two subtables based on the aforementioned categories. For each method, We describe the utilized model(s) (Model column); the addressed task (MFT-related task column); the strategies for tailoring TLMs to the downstream application (Strategy column); the data used for training and/or evaluation (Data column); and the source code availability (Code column).

We also propose a meta-definition for the problem of integrating TLMs with the MFT, aiming to establish a formal foundation framework for all the works examined.

Meta-Definition (Integration of TLMs and MFT). Let M_Θ denote a TLM with parameters Θ , and T_D denote the MFT equipped with a selected set of moral dimensions D (e.g., the set of original moral dimensions proposed by Haidt and Joseph (2004a), or alternative set). An integration process between a TLM M_Θ and a MFT T_D can be described as any function belonging to a class of functions $\mathcal{F} = \{f \mid f : A \rightarrow B \times \mathbb{R} \text{ s.t. } A, B \subseteq \mathcal{T} \setminus \{\emptyset\}\}$, where $\mathcal{T} = \{M_\Theta, \Psi_{T_D}\}$ and Ψ_{T_D} is the set of possible non-empty outcome sets dependent on T_D , i.e., $2^D \setminus \{\emptyset\}$.

The meta-definition forms the foundation for the two categories in this work: Moral-driven TLMs and Moral-targeted TLMs. The key difference lies in the MFT's role: in Moral-Driven TLMs, MFT guides training, allowing the model to predict moral values from input data without MFT during inference; Moral-targeted TLMs rely on MFT at inference to prompt and interpret the model's moral responses.

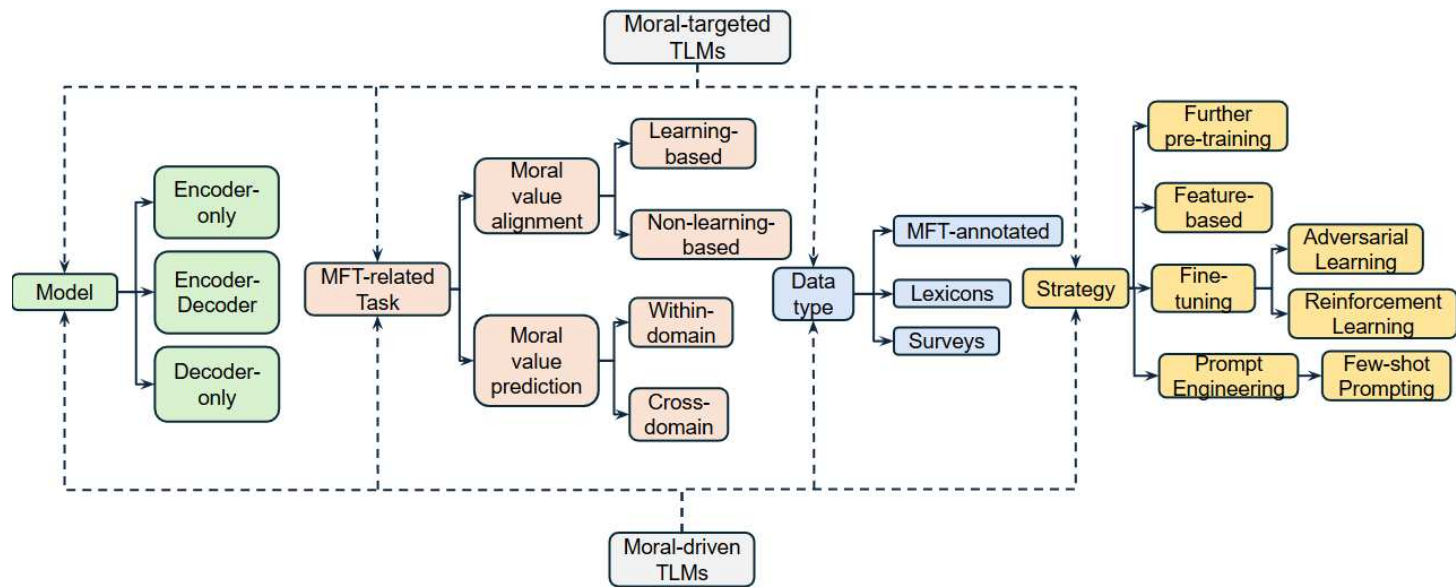


Fig. 7.1: The interplay between the two categories of methods (Moral-targeted TLMs and Moral-driven TLMs) and various aspects related to TLMs within the context of MFT, which encompass: the architecture of the models (green rectangles), the types of MFT-related tasks addressed (red rectangles), the data employed (blue rectangles) and the strategies used for problem resolution (yellow rectangles). Dashed lines indicate dependency relationships, while solid lines indicate parent-child dependencies within the tree.

Table 7.1: Summary of the selected works in this survey. For each work, We indicate the reference of the source work (Ref.); the main component(s) used in the corresponding approach (Model); the addressed task (MFT-related task); the strategies for tailoring TLMs to the downstream application (Strategy); the data used for training and/or evaluation (Data); and the source code availability (Code).

Ref.	Model	MFT-related task	Data	Strategy	Code
Moral-driven TLMs					
Roy and Goldwasser (2021)	DRaIL + BERT	Within-domain moral values prediction	Twitter (Extended Congress)	Fine-tuning	
Pacheco et al. (2022)	DRaIL + BERT	Within-domain moral values prediction	Twitter (Covid-19)	Fine-tuning	✓
Bulla et al. (2022)	SqueezeBERT	Within-domain moral values prediction	Twitter (MFTC)	Fine-tuning	✓
Rao et al. (2023)	BERT	Within-domain moral values prediction	Twitter (MFTC, Congress, Covid)	Fine-tuning	
Trager et al. (2022)	BERT	Within-domain moral values prediction	Reddit (MFRC), Twitter (MFTC)	Fine-tuning	
Liscio et al. (2022)	BERT	Cross-domain moral values prediction	Twitter (MFTC)	Further pre-training + Fine-Tuning	✓
Guo et al. (2023a)	BERT + neural architectures	Cross-domain moral values prediction	Twitter (MFTC, Covid, Congress) and news (MFNC)	Fine-tuning + Adversarial learning	✓
Preinigi et al. (2024)	BERT	Within-domain and cross-domain	Twitter (MFTC),	Fine-tuning +	✓

Continued on the next page

Ref.	Model	MFT-related task	Data	Strategy	Code
		moral values prediction	Facebook, Reddit (MFRC)	Adversarial learning	
Nguyen et al. (2024a)	RoBERTa	Within-domain and cross-domain moral values prediction	Twitter (MFTC), Reddit (MFRC), MFNC, MIC, MFVs, Extended Congress, Social Chemistry	Fine-tuning	✓
Kobbe et al. (2020)	S-BERT, BERT	Within-domain and cross-domain moral values prediction	MFD, Twitter (MFTC), Extended ArgQuality	Fine-tuning	
Alshomary et al. (2022)	BERT + Project Debater	Within-domain moral value prediction and arguments generation	Extended ArgQuality	Fine-tuning	✓
Dognin et al. (2024)	Open-Assistant 12 B	Learning-based moral values alignment	MIC	Fine-tuning + Reinforcement Learning	
Mokhberian et al. (2022)	RoBERTa	Within-domain moral values prediction	MFTC, MFRC, MFNC	Fine-tuning	

Moral-targeted TLMs					
Abdulhai et al. (2023)	GPT-3	Non-learning-based moral value alignment	MFQ	1-shot prompting	✓
Simmons (2023)	GPT-3, GPT3.5 and OPT	Non-learning-based moral value alignment	Social scenario datasets + MFDv1, MFDv2, and eMFD	0-shot prompting	
He et al. (2024)	various LLMs	Non-learning-based	Tweets	0-shot prompting	✓

Continued on the next page

Ref.	Model	MFT-related task	Data	Strategy	Code
		moral values alignment			
Nunes et al. (2024)	GPT-4, Claude 2.1	Non-learning-based moral values alignment	MFQ, MFVs	0-shot prompting	✓
Jiang et al. (2024)	DAMF + Social-PLM	Moral foundations prediction	Tweets about Covid-19	Feature-based	
Carrasco-Farre (2024)	Claude 3 Opus	Non-learning-based moral value alignment	persuasion dataset+ eMFD	0-shot prompting	
Fraser et al. (2022)	T5	Non-learning-based moral values alignment	MFQ	0-shot prompting	
Kennedy et al. (2021)	BERT	Non-learning-based moral values alignment	Facebook status updates + MFDv1, MFDv2	Feature-based,	✓
Xie et al. (2020)	S-BERT	Moral value prediction	MAAFS, MFVs	Feature-based	
Hämmerl et al. (2023)	S-BERT	Non-learning-based moral values alignment	MFQ	Feature-based	✓

Moral-driven TLMs. TLMs are here adapted during the training stage to make inference based on the MFT. The integration process involves a TLM M_Θ in which knowledge about the MFT is embedded into the parameters Θ . Moral-driven TLM can hence be expressed by a function $f \in \mathcal{F}$ such that:

$$f : M_\Theta \mapsto \Psi_{T_D} \times \mathbb{R}.$$

The process for learning the parameters Θ typically employs fine-tuning or pre-training techniques on data labeled with moral dimensions D . In the case of moral values prediction task, We have $f : M_\Theta \mapsto 2^D \times \mathbb{R}$, in which the integration of $\mathbb{R}$ takes into account the case in which We are dealing with both classification and regression tasks (e.g., ranking moral values). As an example, consider a BERT model trained on social media posts labeled with the Care and Harm moral values. When applied to new social media posts, the model can predict whether a post reflects Care or Harm, or to score how much a post contains these moral values.

Moral-targeted TLMs. MFT is here used to analyze and assess the moral values embedded within the responses generated by TLMs. The integration process involves evaluating the responses of M_Θ in order to assess its behavior w.r.t. the MFT T_D . Moral-targeted TLMs can hence be expressed by a function $f \in \mathcal{F}$ such that:

$$f : M_\Theta \times \Psi_{T_D} \mapsto \Psi_{T_D} \times \mathbb{R}.$$

In this case, knowledge of the MFT is not embedded into the parameters Θ , but it is injected through a transformation Ψ_{T_D} . The function f takes the MFT T_D as input and as a basis to analyze the output produced by M_Θ , hence producing a response belonging to the subspace of $\Psi_{T_D} \times \mathbb{R}$. For instance, considering the Non-learning based moral value alignment task, an element of Ψ_{T_D} in the domain corresponds to prompts with elements of T_D (e.g., moral survey datasets). Ψ_{T_D} in the co-domain can be the power set of D . For instance, consider prompting GPT-3 with moral dilemmas, such as “Is it acceptable to lie to protect someone’s feelings?”. The model’s responses are then analyzed using the MFT to determine which moral dimensions, such as Fairness or Loyalty, are reflected in the answers.

Task-adaptation strategies. The works discussed in this Section propose various strategies to address a morality-related task (cf. Table 7.1). In Section 3.2, we formalized the organization of methods by adaptation strategy, where the concepts of task-adaptive and domain-adaptive pre-training were defined and the distinction between task adaptation and domain adaptation was discussed. In this study, we focus exclusively on task-adaptive methods, as none of the works discussed in this Section, except for Liscio et al. (2022), involve domain adaptation. For the sake of simplicity, we also include approaches that do not involve any training on a specific task or domain, since task adaptation in this cases is handled through external tools to the TLM. To facilitate the reader, we briefly outline these strategies with concise definitions.

One of the most common approach of the discussed methods is *fine-tuning*, i.e. performing additional training on a smaller, task-specific dataset. Another common

strategy is the *feature-based* approach, where the TLM’s weights are frozen, and its outputs are used as context-sensitive word embeddings for a separate task-specific model (Min et al., 2024). A number of works opted for *further pre-training*, which involves continuing the pre-training phase of model on an intermediate corpus that is more closely aligned with the target domain or task. Other works proposed *few-shot learning*, in which task descriptions and examples (0 or k examples, respectively 0-shot or k -shot learning) are used to guide a TLM in performing the task. Finally, some works employed *adversarial learning* and *reinforcement learning* in the fine-tuning phase, for enhancing the robustness of the models by exposing them to adversarial attacks and aligning TLMs with human preferences.

7.2.1 Moral-driven methods

In this section, We describe moral-driven TLMs. Initially, We discuss methods that leverage BERT-based model for moral values prediction on social data in both within-domain and cross-domain scenarios. Then, We analyze studies that identify moral values to support argument analysis and moral values alignment.

TLMs for moral values prediction in social discourses. Bulla et al. (2022) extend the work of Hoover et al. (2020a) by replacing the LSTM model with a TLM. Specifically, SqueezeBERT is fine-tuned on the labeled tweets of the MFTC dataset for the task of multi-label classification of moral values. Each of the moral dimensions has been defined as the union of the two moral values that compose it, e.g., for the Care moral value, positive tweets are the ones labeled either with Care or Harm. Experimental results show that the BERT-based model outperforms the LSTM in all seven domains of the MFTC, achieving an F1 score ranging from 47.0 to 86.0 across the moral dimension. Similarly, Rao et al. (2023) fine-tune BERT on Covid-19, MFTC, and Congress datasets to study partisan differences in moral language. They find conservatives use more negative moral language than liberals, indicating a more morally charged tone in conservative Covid-19 tweets. Political elites use moral rhetoric more than non-elites, suggesting leaders frame Covid-19 issues in moral terms, influencing public opinion. Issues like masking and lockdowns are more morally charged in conservative discussions, while vaccines are more prominent in liberal discourse. Trager et al. (2022) fine-tune BERT on the moral dimensions of the MFRC, both in multi-label and binary classification settings. In the former case, all moral categories are predicted simultaneously. In the latter case, each label is predicted separately. Experiments have shown that multi-label BERT achieves lower performance than BERT as a binary classifier, with F1 scores ranging from 0.37 to 0.51. On the MFTC, achieved F1 scores range from 0.54 to 0.82. A cross-domain strategy is also considered, by down-sampling the MFTC dataset to have the same number of samples for each label as MFRC for consistency purposes. Then, a BERT model is trained on the MFRC and tested on the down-sampled MFTC and vice versa. Interestingly, models trained on the MFRC and tested on the MFTC have better classification performance, with F1 scores ranging from 0.28 to 0.53

vs. 0.31-0.43 across the moral dimensions. This is mainly ascribed to the Fairness Foundation’s split into Proportionality and Equality, as described in Section 2.4.3, which makes cross-domain training more difficult.

The effect of cross-domain classification on the moral prediction task have been thoroughly investigated by Liscio et al. (2022). Each topic of the MFTC is treated as a different domain, thus obtaining seven datasets each corresponding to a topic (e.g., Black Lives Matter, Sandy, and so on), and train a BERT model on different settings to evaluate its generalizability and transferability capabilities for moral values prediction as well as its catastrophic forgetting property. Each setting varies based on the source of training data, evaluation, or fine-tuned data. The MFTC dataset is divided into source and target data to test various multi-label classification scenarios. The source data consist of six out of the seven MFTC topics, treated as the available data, while the target data are derived from a novel topic of the MFTC. The settings are as follows: (i) in the *source* scenario, the source data correspond to the training set; (ii) in the *target* scenario, the training set corresponds to the target data; (iii) in the *fine-tune* scenario, the classifier is trained on the source data, and then fine-tuned on the target data; (iv) in the *all* scenario, the training set includes both source and target data. In each scenario, the classifier is evaluated on both source and target data, resulting in eight total combinations. Best results are obtained considering the scenario in which the training set includes both source and target data, obtaining 0.75 of F1 score when the target dataset is related to the topic “ME TOO”. Moreover, similar to Trager et al. (2022), pre-training a BERT classifier on the MFTC yields better performance on unseen data, even when little training data are available. However, fine-tuning on a novel topic causes catastrophic forgetting of the topic it was pre-trained with, even when fine-tuning on a small portion of data from the novel topic. Specifically, this issue arises in the *fine-tune* scenario (iii), where the classifier is first trained on source topic data and then fine-tuned on target topic data. Catastrophic forgetting results particularly evident when evaluating the model on an unbalanced source derived from the Davidson topic (cf. Sec. 2.4.3), in which the shift in label distribution with other topics such as Black Lives Matter is pronounced.

Guo et al. (2023a) propose *DAMF*, a framework for moral value prediction on heterogeneous datasets, leveraging an adversarial training strategy. Rather than training the classifier on the aggregation of data, features from different datasets are mapped into a shared common embedding space. Specifically, a BERT model is used to encode text, and a domain classifier is learned with the purpose of detecting the correct domain the data come from. The domain classifier works as an adversary, hence pushing BERT embeddings to be domain invariant. The moral foundations are the classified based on the learned domain-invariant embedding. *DAMF* is evaluated using different datasets in the cross-domain setting: Covid-19, Congress, MFTC and the MFNC. Results show that model performance improves when fusing heterogeneous datasets. When using Covid-19, Congress and MFNC as training data, the achieved F1 score performance on the MFTC is equal to 0.43.

MoralBERT (Preniqi et al., 2024) is a framework in which both a base BERT and a BERT trained with the adversarial learning strategy used by (Guo et al., 2023a)

are fine-tuned for single-label and multi-label moral values prediction. Similarly to (Trager et al., 2022) and (Liscio et al., 2022), the single-label setting outperforms the multi-label one in terms of F1 score. MoralBERT with adversarial learning outperforms MoralStrength by a large margin and achieves F1 score equal to 0.65 and 0.52, in the single-label and multi-label cases, respectively. The performance degradation in the multi-label setting is likely due to the fact that BERT model is fine-tuned on a specific moral value (e.g., Care), and therefore they fail to capture the relationship arising between different moral values. In a cross-domain scenario (training on two social media datasets, and testing on the third left-out social media dataset), the overall performance tends to decrease since the distribution of moral values differs from dataset to dataset, even if they are from the same domain.

MFormer (Nguyen et al., 2024a) is a collection of five binary classifiers based on the RoBERTa-base architecture (Liu et al., 2019), developed for predicting moral values in textual data. Each classifier is fine-tuned to predict one moral foundation (e.g., Care/Harm) and is fine-tuned on a diverse array of data. Specifically, the MFRC, MFTC, and MFNC datasets were combined into a single dataset for training the model. Notably, unlike other approaches such as (Guo et al., 2023a), the authors chose not to consider the polarity (virtue or vice) of moral dimensions, as these are loosely defined concepts and assigning them can introduce noise due to human annotator bias. MFormer outperformed simple baselines such as MFD (lexicon-based methods), GloVe (embedding similarity), and supervised classifiers (logistic regression) in within-domain moral foundation classification tasks. In cross-domain evaluations, MFormer was tested on the Moral Foundation Vignettes, Moral Arguments, Social Chemistry, and Moral Integrity Corpus datasets, achieving good performance across all moral dimensions.

Methods such as Roy and Goldwasser (2021) and Pacheco et al. (2022) integrate TLM encoders within declarative frameworks, by associating a BERT classifier to logical rules. Both works exploit DRaIL (Pacheco and Goldwasser, 2021), a framework that combines logical reasoning and deep learning for addressing NLP tasks. Roy and Goldwasser (2021) use MFT for elucidating the political stances of American politicians and analyzing their positions. This is addressed by using logical rules within DRaIL to model tweets and other meta-data, from the Extended Congress dataset. Each rule corresponds to a classifier that maps features on the left-hand side of the rule to the predicted output on the right-hand side. Three rules are defined to describe the foundation of a tweet, the ideology, and the topic. Each base rule is associated with a BERT model to encode tweets. Combining multiple rules encoded with BERT is shown to be useful in prediction, achieving the highest performance in terms of the F1 score (0.52). Concerning the MFT, a strong correlation is found between the usage of MFT and politicians' nuanced stances on particular divisive issue, based on Twitter. For example, considering the topic of 'immigration', Democrats employ the term Care to refer to dreamers and young people, while Republicans use this term when discussing 'border wall' and 'border patrol'. Pacheco et al. (2022) apply a similar architecture to identify the moral foundation in tweets about the Covid-19 vaccine. Each tweet is annotated with its moral foundation and vaccination stance. Relational properties of tweets are described using logical rules, indicating

whether a tweet includes a moral judgment, its specific moral foundation, and vaccination stance. Like (Roy and Goldwasser, 2021), tweets are encoded using BERT. Jointly modeling moral judgment and vaccination stance improves performance on the moral prediction task. Mokhberian et al. (2022) aim to improve the performance of RoBERTa in moral value prediction by identifying and removing noisy annotations in the MFTC, MFRC and MFNC datasets. Noisy annotations reflect the subjectivity of human judgments on moral issues, which often lead to annotator disagreement based on personal beliefs, but also ambiguity of the texts and human errors. Specifically, the authors quantify such disagreement through entropy and identify outliers by evaluating how well a text labeled by an annotator aligns with other texts assigned the same label. Training the model on the refined datasets show significant improvements in classification performance, including higher F1 scores and accuracy.

TLMs to support argument analysis. The approaches by Kobbe et al. (2020) and Alshomary et al. (2022) aim to identify moral values for supporting argument analysis. Kobbe et al. (2020) propose different variants of S-BERT to generate text encoding. In one of these, the MFD is used as seed data, and S-BERT is fine-tuned to encode moral foundations in DBpedia¹ data labeled utilizing the MFD through a weak-supervision strategy. In another approach, S-BERT is enhanced with sense-disambiguation features through a word sense-disambiguated version of the MFD based on WordNet synsets. For evaluating the methods, both the MFTC and the extended ArgQuality corpus are used. In one setting, MFTC is used for training different models in a within-domain scenario. When dealing with a cross-domain scenario, models are trained on the MFTC and tested on the extended ArgQuality corpus. In both cases, multi-label BERT is the best-performing method, with an average F1 score equal to 0.45 and 0.308, respectively. The lower performance for the moral class would be explained by the differences in class distribution between the two datasets. Interestingly, in the experiments related to the correlations of multi-label BERT between arguments and moral values, a weak positive correlation is observed between argument quality and moral sentiment for the two most frequent categories on the ArgQuality corpus.

Alshomary et al. (2022) propose Moral Debater to examine how morally framed arguments influence different audiences. A model is learned to identify moral foundations in arguments by fine-tuning BERT on the MoralArg dataset. By testing it on the ArgQuality corpus, the model is found to outperform the multi-label BERT model of Kobbe et al. (2020), obtaining an average F1 score of 0.40 vs. 0.28. The learned model is used to study the effect of morally framed arguments on the audience by combining it with IBM’s Project Debater (Slonim, 2022). Given as input a controversial topic, a stance on the topic, and a set of morals to be targeted, it generates arguments with the given stance. The findings indicate that the generated arguments are aligned with the audience’s moral beliefs. In the case of conservative views, arguments that emphasize loyalty, authority, and purity are compelling, whereas for liberal views, arguments respond strongly to themes of care and fairness.

¹ <https://www.dbpedia.org/>

Adapting TLMs to support specific moral values. Considering the task of learning-based moral value alignment, Dognin et al. (2024) propose Contextual Moral Value Alignment Through Context-Based Aggregation (CMVA-GS), which leverages the MFT to create a series of moral agents based on OpenAssistant 12B. Each moral agent is fine-tuned using reinforcement learning, with reward functions tailored to its respective moral values. The reward function indicates the reward for a specific action at a given prompt based on the moral value being optimized. The moral agents' responses are then aggregated contextually to generate the final output. The fine-tuning and evaluation dataset is the Moral Integrity Corpus (MIC), which is labeled on five moral dimensions. The CMVA-GS models tend to achieve the highest ROUGE scores, indicating better alignment with moral values compared to other TLMs such as Llama-7B and Llama-13B.

7.2.2 Moral-targeted methods

We further explore the existing literature of works examining the moral alignment of TLMs. We firstly review the literature that leverages TLMs, which inherit moral inclinations during their training phase or are influenced by prompt design to provide biased responses in relation to specific moral values. Then, We delve into studies which explore the inherent moral biases in the contextual embeddings of encoder-decoder and encoder-based TLMs.

Moral foundations discovery through prompt engineering. Abdulhai et al. (2023) analyze the moral foundations of GPT-3 models (*curie*, *babbage* and *davinci*) to understand the values encoded from their training data and their potential impact on unforeseen tasks. They examine whether these models exhibit specific moral stances, the consistency of these stances across contexts, and how prompts can influence their endorsement of certain moral foundations.

Questions from the MFQ are exploited to construct prompts containing a description of the task, a rating scale specifying the model response (from 0, meaning 'not relevant', to 5, which is 'extremely relevant'), and a static example to show the model how to structure its response. To evaluate potential cultural and political biases in GPT-3, its default responses are analyzed and prompted with a political affiliation. By comparing the response of GPT-3 models with human studies, it is observed that less expensive GPT-3 models (*curie*, *babbage*) exhibit greater differences from human moral foundations, while more expensive models (*davinci*) align more closely, suggesting that larger models may better capture moral political values. As regards the consistency of TLMs' moral foundations across different contexts, certain dimensions (e.g., Fairness and Purity) are weighted more strongly than others. Additionally, results reveal moral correlations in the model's responses (prompts exalting Authority also lead to higher scores on the Purity traits, similarly for Fairness and Harm). Overall, the results suggest that while GPT-3 models may exhibit certain political tendencies, these can be modified through deliberate prompting, which also affects behavior in downstream tasks.

Simmons (2023) examines the “moral mimicry” phenomena, i.e., whether TLMs can replicate the US political party moral biases. The goals are (i) to assess whether TLMs utilize moral terminology appropriately in different situations; (ii) if they can emulate human moral biases when given a political identity; and (iii) if and how the model size influences such capability. Prompts are intended to motivate language models to provide clear moral justifications. Each prompt contains a scenario, a political identity statement and a moral stance. The scenario is a description of circumstances or behaviors that need moral assessment. A political identity statement is a text referencing political beliefs, such as “liberal” or “conservative”. A moral stance labels the scenario’s acceptance or disapproval such as “moral” or “immoral”. Five distinct prompt templates are utilized to assess sensitivity to specific phrasings. Three dictionaries (MFDv1, MFDv2, and eMFD) are used to evaluate the moral foundational content of the responses from GPT-3, GPT-3.5, and OPT models. To focus on foundational content, the positive or negative connotations of words are removed, merging Virtues and Vices into a single category (for example, Care/Harm). Experimental results for the first goal show that TLMs increase their use of specific moral foundations when those foundations are salient in a given scenario. However, TLMs diverge more from human consensus than individual humans, indicating they do not perfectly replicate human use of moral foundations. When conditioned with political identity, TLMs apply more individualizing foundations (Care/Harm, Fairness/Cheating) when prompted with a liberal identity, and more binding foundations (Authority/Subversion, Sanctity/Degradation, Loyalty/Betrayal) when prompted with a conservative identity. Larger models, especially within the OPT family, demonstrate stronger moral mimicry.

Jiang et al. (2024) combine DAMF with Social-PLM (Jiang and Ferrara, 2024) to detect moral values on a Covid-19 dataset (Chen et al., 2020) and analyze social media user groups. Users valuing fairness and authority participate more in discussions, while those emphasizing Care and Purity show distinct engagement patterns. Tweets combining multiple moral foundations are rare but attract more engagement.

Carrasco-Farre (2024) examines the ability of TLMs to adjust their arguments based on the prompt, employing persuasive strategies that require different levels of cognitive effort and utilizing different moral language. To this purpose, the study utilizes the responses of the experiments performed by Anthropic,² in which human participants were presented with various claims on diverse topics (e.g., police body cameras and lab-grown meat) and corresponding persuasive arguments generated by Claude 3 Opus or human-generated. Persuasiveness is measured by shifts in agreement with claims before and after exposure to arguments. The TLM is prompted to produce arguments using different persuasive techniques (for example, arguments aimed at persuading undecided or opposed individuals). The analysis focuses on both cognitive effort, through readability and perplexity, and the moral-emotional language, evaluated using sentiment analysis and the frequency of words expressing MFT moral traits. The TLM use more moral language than humans overall, with a higher frequency of both positive and negative moral foundations, contributing to

² <https://www.anthropic.com/news/measuring-model-persuasiveness>

its persuasive power. The study suggests that the strategic use of moral language by TLMs, including negative bias such as harm and cheating, plays a significant role in persuasion.

He et al. (2024) focus on evaluating the affective alignment of 36 TLMs (such as LLaMA-2, GPT-3, OPT and etc.). The goal is to measure how closely the models' emotional and moral tone aligns with humans ones from liberal and conservative perspectives. It uses two prompting approaches: default prompting, where TLMs generate responses without ideological context, and steered prompting, where TLMs are guided to adopt a liberal or a conservative perspective. The responses generated by the models are compared to existing tweets on contentious topics like COVID-19 and abortion. Using DAMF, the moral values from these tweets are extracted, while the political ideologies of the users are identified based on the political bias of the news sources shared in their tweets. The results reveal that TLMs tend to express higher levels of Care and Loyalty than human tweets, skewing towards liberal perspectives, even when prompted with conservative viewpoints.

Nunes et al. (2024) explore whether GPT-4 and Claude 2.1 exhibit moral hypocrisy by testing the consistency of the models' abstract principles and whether such abstract principles align with their judgments in real-life scenarios. The models were assessed with MFQ for abstract values and the MFVs for concrete moral judgments. Specifically, they are prompted to respond to MFQ and MFVs with numerical answers (e.g., on a Likert scale) without providing any explanatory text or reasoning. While both models showed consistency in handling abstract values, comparable to human respondents, they struggled to apply these values across different concrete contexts, revealing a form of moral hypocrisy.

Moral bias analysis in TLMs embeddings. Fraser et al. (2022) explore the Delphi framework (Jiang et al., 2021) for the MFT. It is a TLM built on the UNICORN (Lourie et al., 2021) model, which is a specialized version of T5 for commonsense question answering. Delphi is fine-tuned on the COMMONSENSE NORM BANK dataset (Jiang et al., 2021), consisting of various moral judgments on everyday situations. Delphi responds to ethical queries in various formats, predicting moral judgments that align with general societal views. The system's final task is to provide moral judgments in response to user queries. According to Fraser et al. (2022), when using the MFQ as a basis for their Delphi queries, the two most crucial foundations are Care and Fairness.

Kennedy et al. (2021) aim to assess a direct connection between individual moral concerns and language, examining both moral and general language (i.e., not necessarily expressing moral concerns). In this study, data are gathered from 2.6k participants, including their responses to the MFQ and 107k Facebook posts. Two approaches are employed to relate moral concerns to language. The first is MFD-based, testing the connection between moral language (via MFD and MFDv2) and individual concerns. The second approach uses general language, applying bag-of-words with LDA and text embeddings via GloVe and BERT. A regression analysis is carried out to predict each MFQ score from the text representations. Experimental results reveal that MFD-based measures are generally less effective at predicting

moral concerns compared to NLP-techniques. In particular, BERT is able to capture differences in language w.r.t. moral concerns that were not captured by individuals' lexicons.

Xie et al. (2020) introduce a text-based method for predicting human moral judgments of moral scenarios. The aim is to ascertain whether human-judged moral categories can be deduced solely from text, leveraging contextualized language models. The study assesses different representations of moral scenarios and their efficacy in categorizing nuanced moral differences, even with limited training data. The scenarios are gathered from three independent datasets, including MAAFS and MFVs. The experimental settings evaluate several classifiers with different representations of moral vignettes, including contextual embeddings with S-BERT. S-BERT embeddings achieve the highest accuracy across all datasets and classification methods.

Hämmerl et al. (2023) explore whether moral standards learned from dominant languages, such as English, are imposed on other languages during cross-lingual transfer learning. They employ the MORALDIRECTION framework for multilingual TLMs to encode action verbs with positive and negative connotations. These verbs are used in template questions to generate sentence embeddings, and PCA is applied to derive a “moral direction” subspace. The scores are then normalized to estimate moral values. The framework is tested with multilingual models (e.g., mBERT, XLM-R), monolingual models, and S-BERT models for each language, showing high correlation between S-BERT and human moral judgments across languages, though slightly lower for Arabic and Czech with XLM-R.

Training with parallel data eliminates cultural differences, aligning models more closely with English. Analysis of parallel sentence pairs highlights discrepancies caused by literal translations and lexical confusion, especially in Czech-English pairs.

TLMs' moral dimensions are evaluated using the MFQ by comparing model scores against human responses across cultures. While models can generally distinguish right from wrong, they struggle with complex sentences and negations. The models' scores do not fully align with human responses and vary by language, with greater discrepancies observed in models trained on smaller language-specific datasets, underscoring the impact of training data on moral judgments.

7.3 Chapter Review

Our discussion so far has provided insights into the potential of TLMs in modeling the framework of the MFT. These insights might contribute to a deeper understanding of the complexities involved in integrating MFT and TLMs and the ongoing efforts to improve their performance.

Main trends in the existing literature. Observing Table 7.1, recurring patterns can be noticed regarding models selection, tasks addressed, data employed, and strategies used (the interplay among the categories and these aspects are shown in Fig.

7.1). From the perspective of model choice, BERT and BERT-based models represent the most commonly used technology when creating moral-driven TLMs, while TLMs are primarily employed as moral-targeted TLMs. Clearly, the choice of the model has a significant impact on the adaptation strategy. Therefore, a clear separation of strategies can be observed between moral-driven TLMs, which exclusively use fine-tuning strategies to adapt to the MFT task, and moral-targeted TLMs, which adopt few-shot prompting and feature extraction. Moral-driven TLMs are frequently employed for tasks involving moral values prediction, while moral-targeted TLMs are predominantly utilized for moral value alignment tasks. Specifically, MFT-annotated datasets are generally used in the context of within-domain and cross-domain moral values prediction tasks. Conversely, lexicons and moral surveys are primarily used for addressing moral value alignment challenges. As regards the employed data, We can notice that moral surveys and moral lexicons are mostly used to evaluate moral value alignment, while MFT-annotated datasets are primarily involved in tasks of moral values prediction.

Correlation between moral foundations and political stances. A prominent theme is the correlation between moral foundations and political stances. Roy and Goldwasser (2021) demonstrate a strong link between the use of moral foundation language in tweets and the nuanced political positions of U.S. politicians on issues such as gun control and immigration. This correlation is further explored by the work of Simmons (2023), who finds that changing the political identity in prompts affects the likelihood of expressing particular moral foundations, with TLMs using more individualizing foundations when prompted with a liberal identity, and more binding foundations when prompted with a conservative identity. Additionally, Rao et al. (2023) highlights partisan differences in moral language, noting that conservatives tend to use more negative moral language compared to liberals, suggesting a more critical and morally charged tone in conservative tweets.

TLMs vs Lexicon-based methods. The superiority of TLMs over traditional and lexicon-based methods is evident in several studies. (Bulla et al., 2022) notice enhanced performance with TLMs upon earlier works, while (Xie et al., 2020) highlight the superior performance of contextual embeddings in predicting human moral judgments. Additionally, Kennedy et al. (2021) point out the limitations of lexicon-based measures, suggesting that predefined moral dictionaries may not fully capture the complexity and variability of moral language as expressed naturally by individuals. The moral nuances in natural language can be better captured by TLMs. Furthermore, Roy and Goldwasser (2021) and Pacheco et al. (2022) investigated the integration of TLMs with logical reasoning frameworks. They show that combining TLMs such as BERT with declarative frameworks like DRaiL enhances the performance of NLP tasks by effectively modeling complex relational properties and dependencies in text data. This approach allows for the definition of logical rules that capture these dependencies, leading to improved performance.

Effect of training strategies. Training strategies, including fine-tuning and adversarial learning, have shown several enhancements in moral value prediction tasks. The works proposed by Kobbe et al. (2020), Alshomary et al. (2022), Dognin et al.

(2024), and Guo et al. (2023a) highlight the effectiveness of these approaches in aligning model outputs with desired moral values.

Fine-tuning emerges as the predominant strategy for adapting models to specific tasks and datasets, involving additional training on task-specific datasets to capture nuances and improve performance. Adversarial learning, as employed by DAMF (Guo et al., 2023a) and MoralBERT (Preniqi et al., 2024), is used to create domain-invariant embeddings, improving cross-domain performance by reducing the impact of domain-specific biases. Prompt engineering and reinforcement learning are notable strategies for aligning TLMs with specific moral values without extensive fine-tuning.

Effect of prompting. Several works have examined how the construction of prompts can elicit specific moral traits and influence TLM responses. This aspect is crucial for understanding how TLMs reflect and generate moral judgments, providing insights into the moral tendencies embedded in the models' pre-training data. Abdulhai et al. (2023) and Simmons (2023) show that deliberate prompting can significantly influence the moral foundations exhibited by models, affecting their responses and behaviors in various tasks. This highlights the malleability of TLMs and their potential to be guided through carefully designed prompts to exhibit desired moral and ethical considerations. Additionally, modifying the models' moral stances through prompting affects their behavior in downstream tasks, such as donation decisions, demonstrating the practical implications of moral prompting.

Effect of cross-domain strategies. Cross-domain strategies and their impact on model performance have been extensively explored. Trager et al. (2022), Liscio et al. (2022), Preniqi et al. (2024), Guo et al. (2023a) and Alshomary et al. (2022) demonstrate that training models on heterogeneous datasets generally improves performance. However, challenges remain due to class distribution differences and the phenomenon of catastrophic forgetting when models are fine-tuned on new domains, as noted by Liscio et al. (2022). This issue highlights the difficulty of maintaining model accuracy across varied datasets.

Cross-lingual approach and lack of diverse datasets. Cross-lingual transfer of moral standards is another critical area of investigation. Hämmerl et al. (2023) explores whether moral standards learned from dominant language sources (e.g., English) are imposed on other languages during cross-lingual transfer learning, which could lead to negative consequences. Training with parallel data tends to eliminate cultural differences, aligning models more closely with English, but models trained on smaller, language-specific datasets show more variability in moral judgments.

However, the scarcity of labeled datasets in non-English languages limits our understanding of how TLMs perform in predicting moral values across diverse linguistic and cultural settings. An exception is Simonsen and Widmann (2023), which collected more than 20k sentences from political communication across six languages (Danish, Dutch, English, German, Swedish, and Spanish) to study moral appeals in party manifestos from ten Western democracies. It is worth mentioning also the *Moral Machine* (Awad et al., 2018), an online platform which exposes moral dilemmas to face as car drivers and collected 40M decisions in 10 languages from

millions of people in 233 countries. While these efforts represent important steps forward, they remain limited in scope and coverage.

Recent studies on cultural bias in NLP models reveal that TLMs often mirror the cultural norms of high-resource languages. For instance, Tao et al. (2024) shows that widely used TLMs tend to align more closely with the cultural values of English-speaking and Protestant European countries compared African-Islamic countries. Navigli et al. (2023) highlights that most NLP models are heavily influenced by high-resource languages, resulting in skewed representations of cultural norms and values. The dominance of these languages in training corpora amplifies their influence on downstream tasks, often neglecting the unique cultural dimensions of low-resource languages. Johnson et al. (2022) demonstrate the challenges associated with low-resource languages, where limited data availability hinders the training and performance of TLMs. Creating balanced and representative datasets from various cultural and linguistic backgrounds will ensure that TLMs can better generalize moral judgments across different contexts. Overall, the ability of TLMs to effectively represent and predict moral values in linguistically and culturally distinct low-resource languages remain uncertain. A comprehensive training on diverse datasets is crucial for improving the alignment of TLMs with moral values across different languages and cultures. Without such efforts, there is a risk of perpetuating moral biases that favor high-resource languages and aligning poorly with the moral standards of underrepresented communities.

Moreover, the majority of existing datasets are sourced from social media platforms, as indicated in Table 7.1, which limits the diversity of language and context represented. The reliance on social media data, primarily in English, does not capture the full spectrum of moral values across different cultures and languages.

The role of model size. Model size also plays a crucial role in moral alignment. The works proposed by Abdulhai et al. (2023) and Simmons (2023) indicate that larger models, such as GPT-3, align more closely with human moral foundations compared to smaller models. This suggests that model complexity contributes to better capturing and representing moral values.

Bias on moral judgment and reliability of manually annotated data. The potential bias in moral judgments is discussed by Fraser et al. (2022), who hypothesize that Delphi’s moral principles are likely influenced by the moral views of its annotators, specifically young, English-speaking North Americans. This highlights a broader concern: the responses of systems trained on morally annotated datasets may not be universally applicable or unbiased. Several works have highlighted the inherent subjectivity and potential biases in annotating moral judgments, emphasizing the influence of annotators’ worldviews, beliefs, and demographic characteristics. For instance, Ziems et al. (2022) point out that any collection of moral judgments will inherently reflect the worldviews of its annotators. In their work, they employed Amazon Mechanical Turk workers, which tend to be more educated, less religious, and more likely to be unemployed w.r.t. the general population, according to Goodman et al. (2013). Similarly, Kobbe et al. (2020) and Mokhberian et al. (2022) emphasize that the personal beliefs of human annotators can influence the

annotation process, raising questions about the validity of annotations employed in training and evaluating automated systems. This issue is further elaborated by Johnson and Goldwasser (2018), who observed the political bias of annotators when labeling tweets. To mitigate this, annotators were instructed to label from the perspective of the politician who wrote the tweet. Also Trager et al. (2022) noted that moral judgments are highly subjective and likely to be influenced by the annotators' sociopolitical and demographic contexts. They argue that while such biases may have minimal impact on tasks like sentiment analysis, they can profoundly affect domains involving moral values. They provided detailed psychological and demographic metadata about their annotators and recognize that their dataset reflects biases specific to its English-speaking, affluent, and predominantly liberal annotator pool, which limits its representativeness of global moral concerns. Similarly, Hoover et al. (2020a) discuss disagreements among annotators on moral foundations, highlighting the difficulty in achieving consensus due to the absence of objective criteria for moral coding. They warn that resolving such disagreements excessively can artificially inflate inter-annotator reliability, masking the inherent subjectivity of moral judgments. To this regard, Weber et al. (2021) provide further evidence on the subjectivity and variability inherent in human moral annotations. Their study shows that even highly trained annotators (i.e., individuals who have undergone extensive preparation specifically tailored to the annotation task) fail to achieve high inter-annotator reliability, emphasizing that moral intuition extraction is strongly influenced by individuals' demographic and ideological characteristics. They argue that traditional methods relying on small groups of expert annotators may inadvertently inflate inter-annotator reliability through implicit consensus-building, potentially at the expense of validity. Instead, they propose leveraging crowd-sourcing platforms to aggregate annotations from a diverse population, acknowledging and embracing the inherent subjectivity of moral judgments as a more valid approach.

These considerations highlights the need for a careful selection of annotators, ensuring a sufficient demographic and moral diversity to promote balanced and fair moral judgments. A related issue is the reliability of manually annotated data. Before moral values can be reliably utilized for argument analysis, the quality of the annotations must be ensured. A low inter-annotator agreement doubt the validity of the findings, raising concerns about the consistency and accuracy of these annotations (Kobbe et al., 2020; Fraser et al., 2022). Therefore, there is an emergency for robust and standardized annotation processes.

Moral labeling discrepancy. Another limitation regards the data sources utilized in the field (Table 2.5). Among the various datasets, a notable discrepancy is observed in the choice of labels, with some datasets employing 10 or more labels and others opting for 5. The underlying motivation for this decision is that both labeling schemes are closely related, and it is often ambiguous which end of the dimension is being addressed, particularly in the context of negated sentences; for instance, the sentence “We could never hurt you” could either be classified under Harm due to its use of vocabulary related to this dimension or be annotated as the opposite, Care, since it expresses an inability to harm someone, thereby aligning more strongly with

the Virtue class (Kobbe et al., 2020). This duality underscores the complexity and subjectivity inherent in labeling and categorizing moral dataset, since also TLMs exhibit inconsistencies, especially with complex phrases and negation (Hämmerl et al., 2023).

Ontologies to guide TLMs. Given the flexibility of TLMs and their significant responsiveness to prompting, there is a clear need for models that can be guided by predefined ontologies to provide consistent and coherent responses aligned with a specific worldview. The core issue here is that morality is deeply influenced by cultural factors, which vary widely across different societies. By integrating ontologies, which encapsulate structured knowledge and worldviews, models can be better directed to align their outputs with culturally relevant moral standards. This approach would involve embedding cultural and moral frameworks directly into the model's training process, ensuring that the generated responses adhere to these predefined moral principles. Such a method would mitigate the influence of arbitrary prompts and steer the model towards more stable and culturally coherent moral judgments.

Explainability and interpretability. A critical area lacking in current research is the explainability and interpretability of TLM outputs concerning moral judgments. Most existing works do not explore the reasoning behind the model's outputs, leaving a gap in understanding how decisions are made. This lack of transparency can undermine trust in these models, especially in sensitive applications involving moral and ethical considerations. Incorporating methods similar to the Chain-of-Thought approach, as demonstrated in the work of Jin et al. (2022), can significantly enhance the interpretability of TLMs. By explicitly detailing the steps and reasoning paths taken to arrive at a given solution, users can gain insights into the underlying processes of the model. This not only improves trust and reliability but also allows for the identification and correction of biases and errors in the model's reasoning. Developing frameworks for explainability and interpretability in moral prediction tasks is essential for advancing the field and ensuring the ethical deployment of TLMs.

Chapter 8

E2MoCase: A Dataset for Emotional, Event and Moral Observations in News Articles on High-impact Legal Cases

The media plays a pivotal role in shaping public perception and discourse, especially concerning legal issues (Sadaf, 2011). Media narratives often depict events laden with emotions and moral undertones, influencing individuals' interpretations of controversial topics. These narratives use emotionally charged language to describe events with social implications, hence triggering specific moral evaluations. Analyzing human behavior requires a multifaceted approach that integrates different dimensions of values, which are "beliefs" that guide how individuals evaluate behaviors and events (Schwartz, 1992). Among these, morality stands as a crucial aspect for understanding behavioral responses, especially in controversial and biased topics (Haidt and Joseph, 2004b). However, relying only on moral information can be limiting. In fact, morality is often the driving force behind value-based judgments and emotional reactions (Raney, 2011). *Emotions* are immediate, intuitive responses that can precede rational deliberation, influencing how people interpret events. Moreover, *events* are actions that carry social or moral implications, described by words where the involved entities cause moral and emotional reactions; for example, witnessing injustice can evoke emotions like anger or empathy, shaping the moral judgment and interpretation of the event. Note that we are aware that the scientific literature does not offer a universally agreed-upon interpretation for the notions of emotions and events. The definitions we provided above are those adopted in this work.

The aspects of emotions, moralities and events, are deeply interconnected. Studying controversial topics, such as media narratives on legal cases, can be valuable for analyzing human-driven values, as these narratives often frame events emotionally, influencing moral evaluations and reinforcing or challenging existing value systems.

Despite advancements in AI-based NLP tools, there is a lack of labeled data that jointly captures events, emotions, and moralities

To face these challenges, we introduce a new dataset for **Emotional, Event and Moral Observations in high-impact legal Cases**, dubbed **E2MoCase**,¹ specifically designed to identify events, emotions and moralities in media narratives reporting on legal cases subject to bias (Greco et al., 2024b). Our dataset can be used for several

¹ We make our data available upon request.

NLP tasks based on its key-constituting dimensions (i.e., moralities, emotions, and events), but also to support related applications, such as bias detection, by training models on E2MoCase to recognize the moral, emotional, and factual characteristics of biased texts. This would be a step toward creating a more equitable landscape for legal and media activities, where the advancements in AI can positively impact society.

We summarize our contributions as follows:

- We present E2MoCase, a novel dataset for analyzing emotions, moral values, and events in media coverage of legal cases.
- We show how emotions, moral values, and events arise from legal biased narratives.
- We provide an in-depth validation of the dataset through comparative experiments on morality and emotion detection.

8.1 Related Works

The current landscape of datasets tends to focus on either emotional analysis, moral reasoning, or event detection separately, with limited exploration of the intersection of these elements.

Some exploration into integrating these elements can be found in Cherichi and Faiz (2016), who investigated the enhancement of social event detection using big data values, suggesting the potential for integrating various data types, including emotional and moral signals. Heron et al. (2018) provided a dataset specifically designed to capture moral emotions in dyadic interactions, offering a unique perspective on the interplay between moral considerations and emotional states. Gloor (2022) explored methodologies for measuring moral values from facial expressions, indicating a pathway towards the incorporation of these measurements into broader datasets that encompass event detection and moral evaluations. Zlatintsi et al. (2017) proposed COGNIMUSE, a dataset annotated with sensory and semantic saliency, events, cross-media semantics, and emotions, designed for training algorithms for event detection, summarization, classification, and emotion tracking in videos. Moreover, the study by Hromic and Hayes (2014) provided insights into handling large-scale data where emotions and moralities are inherently intertwined.

In recent years, there has been a wide recognition that comprehensive data sets integrating moral, emotional and event dimensions are needed for effectively analyzing media narratives. As suggested in (Kepplinger et al., 2012), media depictions shape recipients' cognitions, emotions, and opinions, which extend beyond the original frames and are influenced by individual processing. In this regard, Zhang et al. (2024) introduced MoralEvents, a dataset consisting of structured annotations of moral events across news articles from U.S. media reports, to study how moral values are implicitly conveyed through events rather than explicit moral language. Lei et al. (2024) proposed the EMONA dataset, containing news articles annotated

with moral judgments of events, to analyze how these opinions reflect ideological bias and media framing. Li et al. (2016b) constructed a corpus of news articles for event-based emotion analysis where events are used for emotion classification.

The need for such frameworks is even more exacerbated when it comes to media narratives that report legal cases. Such narratives are inherently multidimensional, encompassing factual events, emotional responses, and moral judgments. Considering these aspects is essential for training AI models that capture the complexity of how legal issues are framed and perceived in public discourse. To the best of our knowledge, there is no dataset in the literature that meets the combined need for capturing factual, emotional, and moral dimensions within legal media narratives.

To address this gap in the literature, We introduce E2MoCase, a holistic dataset combining emotions, moral values, and events in media narratives reporting on high-impact legal cases. In capturing the complex interplay of emotions, moral values, and events in legal media narratives, E2MoCase provides an unprecedented, comprehensive resource for developing AI models that are sensitive to the multidimensional nature of public discourse on legal matters.

8.2 Data Collection

Fig. 8.1 shows the construction process of E2MoCase, which consists of the following phases: **Case selection**, which was carried out semi-automatically to identify legal narratives that have attracted extensive media coverage; **News retrieval** from reputable newspapers concerning the cases selected in the case selection phase; **Dataset design and annotation**, whereby We automatically annotated news items at paragraph level with values of morality, emotions, and events, employing well-recognized domain-specific TLMs, namely *MoralBERT* (Preniqi et al., 2024), *EmoLLaMA* (Liu et al., 2024) and *GoLLIE* (Sainz et al., 2023a).

Eventually, each data instance of E2MoCase includes (i) the subject of the case; (ii) the news article’s identifier; (iii) the numerical identifier of the paragraph within the article; (iv) the paragraph translated into English (if the original language was different) and (v) its original language, (vi) the numerical scores associated with each of the emotion labels, (vii) the numerical scores associated with each of the morality labels, and (viii) a JSON string for each of the detected events.

In the following, We elaborate on each step performed for the construction of E2MoCase.

8.2.1 Case selection

We first selected more than 100 candidate cases related to legal matters that had significant media impact due to evidences of cultural biases, such as religious, political, gender, racial and media biases (Spinde et al., 2023). Specifically, We found

such cases through both an extensive manual searching and the use of generative TLM tools with online search capabilities, particularly ChatGPT Plus,² Perplexity,³ and Claude-3-Opus-200K.⁴

For each case, we manually verified its accuracy in terms of reported news, We ensured it had significant media impact and it was covered by reputable newspaper agencies. Example of prompts used to find the cases are shown in the Appendix C.1.

8.2.2 News retrieval

Given the candidate cases, We gathered the corresponding news data through the use of the Swissdox API,⁵ which facilitates the retrieval of extensive volumes of Swiss media data for scholarly research. The Swissdox API provides news articles from the last 25 years, in 5 different languages: English, French, German, Italian and Romansh. It requires as main parameters: (i) the temporal range of the news to be gathered, (ii) the source news outlets, (iii) the language of the news, and (iv) a list of keywords that can be combined using the logical operators AND and OR. Considering parameter (i), We chose the year in which the case broke out, up to the year of this research study, i.e., 2024; for parameters (ii) and (iii), We selected any news outlets and language, while for (iv) We used the names of the main characters involved in the case combined with AND operator. Fig. C.1 in Appendix C.2 shows an example of a query submitted to SwissDox.

Then, We analyzed the candidate cases and selected 39 cases that We manually identified as crucial for encompassing cultural biases, such as gender and racial (more details in Appendix C.3).

We finally built a corpus of 97,351 news paragraphs, which were translated into English to accommodate the predominance of English-based TLMs in recent research on moral, emotion, and event detection. The translations were performed using Google Translate API calls. While we acknowledge potential inaccuracies, especially for low-resource languages like Romansh, it is worth emphasizing that Romansh articles are minimal in comparison to those in Italian, French, and German—languages well supported by automated translation tools.

8.3 Dataset Design and Annotation

In our framework, We utilized a paragraph-level annotation to retain the contextual integrity of the original text. We believe that paragraphs represent a good balance

² <https://chatgpt.com/>

³ <https://www.perplexity.ai/>

⁴ <https://www.anthropic.com/claude>

⁵ <https://liri.linguistik.uzh.ch/wiki/langtech/swissdox/start>

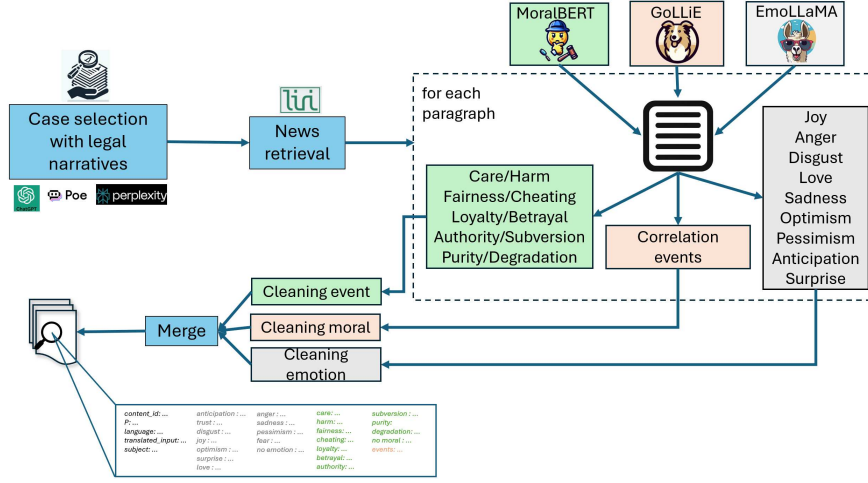


Fig. 8.1: Workflow diagram illustrating the construction of E2MoCase.

between input length and semantic coherence preservation, retaining contextual integrity and supporting a fine-grained analysis. Specifically, each paragraph within E2MoCase was annotated with three silver labels, where each label consists of one or more values from each of the following types:

- **Emotional tone**, identifying the presence and intensity of specific emotions in the text: *joy, love, optimism, trust, sadness, disgust, pessimism, fear, anticipation, surprise*.
- **Moral trait**, which pertains to the underlying morality aspect according to the MFT.
- **Events occurrence**, which focuses on the events detected within the paragraph and the subjects involved. Events are defined as specific occurrences involving one or more participants (Xiang and Wang, 2019), providing precise contextual information. In our setting, an event always includes a *mention*, i.e., a phrase/sentence describing the event. It may also include *entities*, i.e., the participants involved in the event, and their respective *roles*.

The above annotations were independently performed by using automated tools, as shown in Fig. 8.1. We argue that the combined use of three types of labels to associate with news articles is crucial for a comprehensive understanding of the phenomena in the legal narratives conveyed by modern media.

Emotion detection. For detecting emotions in news paragraphs, We leveraged EmoLLMs (Liu et al., 2024), a series of open-source TLMs for comprehensive affective analysis. EmoLLMs are obtained by fine-tuning various TLMs on the *Affective Analysis Instruction Dataset*, an instruction dataset based on the data from *SemEval 2018 Task 1: Affect in Tweets* (Mohammad et al., 2018). A comprehensive analysis of the performance of EmoLLMs on in-domain and out-domain data indicate

that EmoLLMs perform exceptionally well in both regression tasks and classification tasks, surpassing the best model of SemEval 2018 leaderboard and achieving SOTA compared to the other open-source TLMs (Falcon, Vicuna, LLaMA-7B-chat, LLaMA-13B-chat). Encouraged by these remarkable results, We opted to use the best-performing EmoLLM from the test set evaluations, i.e., EmoLLaMA-chat-13B, for the emotion classification (E-c) task and the emotion intensity regression (EI-reg) task. This allowed us to identify a list of detected emotions from the input text (E-c), and to quantify the intensity of each detected emotion (EI-reg). In our experiments, We performed 8-bit quantization on the model, since it significantly reduces the computational requirements, allowing for faster inference times without substantial loss of accuracy.

Moral classification. We detected the moral values present in the news articles’ paragraphs according to the MFT. For this purpose, we utilized MoralBERT (Preniqi et al., 2024), a suite of BERT-based sequence classifiers, each fine-tuned for detecting and quantify moral values in social media data. In Preniqi et al. (2024), best results in out-of-domain contexts were obtained in the single-label scenario (i.e., one model designed for a specific moral value) and performing domain-adaptation with adversarial training. This is the configuration we have adopted for our analysis. To the best of our knowledge, MoralBERT represents the most valuable current approach for detecting morality.

Event extraction. We extracted events from each paragraph using GoLLIE (Sainz et al., 2023a), an TLM derived from Code-LLaMa (Rozière et al., 2024) and fine-tuned to comply with annotation guidelines expressed as code-based instructions. By following the specific prompt described in GoLLIE (Sainz et al., 2023a) for event extraction, We defined the target event encapsulating the guidelines and instructions as docstrings. Specifically, We focused on what We called as *Correlation Event*, i.e., any dynamics, state change or relationship involving one or more entities. The definition of the Correlation Event and some examples of the extracted events are shown in Appendix C.4. In our experiments We used Code-LLama with 13B parameters and 8 bit quantization. Note that We also tested InstructUIE (Wang et al., 2023b) for the task of event extraction, which is a fine-tuned version of Flan-T5 with 11B parameters. However, We empirically found out that the results provided by GoLLIE were more accurate and consistent across different paragraphs, i.e., same output format. Note that, in the following, we will use the term “event” to intend correlation events.

8.4 Dataset Analysis

News articles collected from SwissDox come from web pages, which may include various events of the day or insertions between paragraphs that are unrelated to the case of interest. To address this issue, We have identified these outliers by leveraging emotional information, and manually filtered out them. Table 8.1 shows the statistics

Table 8.1: Statistics of E2MoCase and its variants: E2MoCase.noEvents and E2MoCase.full

	E2MoCase	E2MoCase.noEvents	E2MoCase.full
# paragraphs	50, 975	46, 276	97, 251
avg # tokens	275.106 $\pm$ 245.303	139.402 $\pm$ 220.950	210.532 $\pm$ 243.647
avg # emotions	1.164 $\pm$ 0.757	1.634 $\pm$ 0.680	1.678 $\pm$ 0.657
avg # morals	3.517 $\pm$ 3.870	1.773 $\pm$ 1.644	2.795 $\pm$ 2.424
avg # events	3.597 $\pm$ 2.940	0.0 $\pm$ 0.0	1.885 $\pm$ 2.785

of E2MoCase with its variants, i.e., E2MoCase.noEvents, obtained by removing paragraphs that do not contain events, and E2MoCase.full, which also includes paragraphs that do not contain events. We can notice that the event extraction process identified events in only 50,975 paragraphs, resulting in the exclusion of nearly half of the initial paragraphs.

We can also observe high variability in the paragraph length in terms of number of tokens (counted using WordPiece tokenizer) in all the variants; in addition, the average number of detected emotions per paragraph is 1.16 in all cases, whereas the average number of moral values is significantly lower in paragraphs without events compared to those with events. This also indicates that the presence of events, involving actions, decisions or situations, causes a stronger moral response.

We further analyze E2MoCase from the perspective of each type of label to gain insights about the underlying media narratives. Figs. 8.2a and 8.2b show the distribution of emotions and morality traits, respectively. On the one hand, it can be observed that the most frequently detected emotions are *anger* and *disgust*, which are also strongly correlated (Fig. 8.3b). On the other hand, the distribution of moral values is much more balanced, with a predominance of *cheating*, *subversion*, and *betrayal* values, which are also strongly correlated (Fig. 8.3a) along with *harm*. This difference between emotion and morality labels may be due to the different distribution (Figs. 8.2b and 8.2a). Furthermore, emotions tend to have a low score distribution, mostly below 0.5 (Fig. 8.6).

From Fig. 8.3c, We observe that *cheating*, *betrayal*, *subversion* and *degradation* moral values show a significant positive correlation with emotions such as *anger* and *disgust*. This suggests that narratives involving negative moral values are likely to evoke strong negative emotions and vice versa. Nonetheless, as depicted in Fig. 8.4, the events behind the news selected for E2MoCase, are quite controversial such as *murder*, *shoot*, *rape*. This is expected given the topics covered by the collected news articles. Additionally, We observed that in the paragraphs lacking any emotions or moral dimensions, which are approximately 3% of the total, the associated events include innocuous facts like *buy*, *sign*, *submit*, *announce*, *move*, and *win* (Fig. 8.5). It is likely that these events do not evoke strong moral or emotional responses since they are perceived as routine or neutral actions, devoid of inherent effective implications.

Finally, Fig. 8.7 shows the mean similarity heatmaps for pairs of texts labeled with moral categories (m_1, m_2) from the E2MoCase, MFTC, and MFRC datasets. In

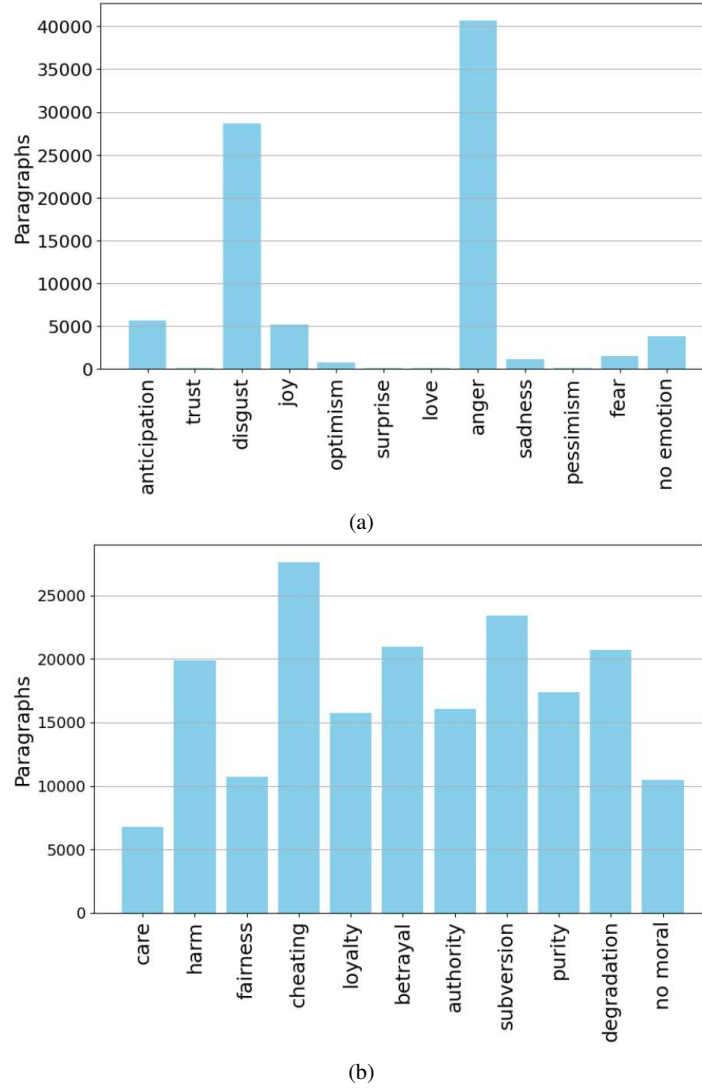
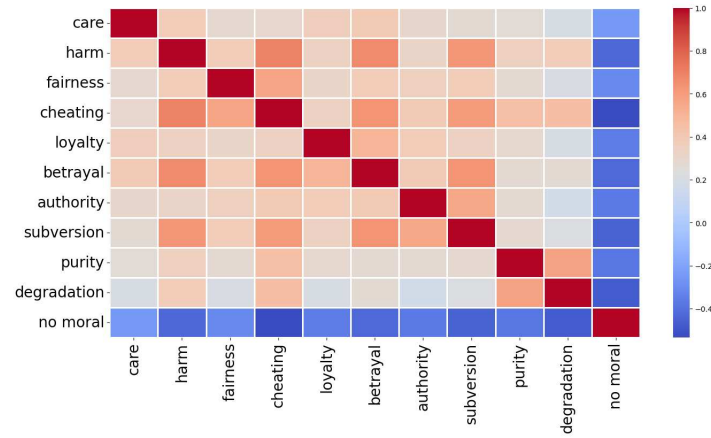
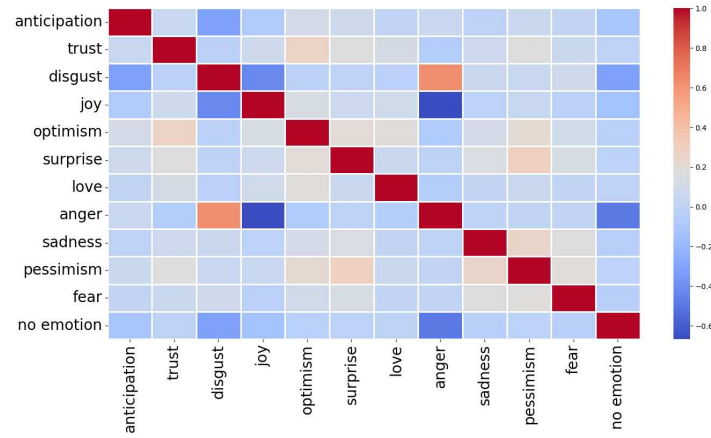


Fig. 8.2: (a) Distribution of emotions and (b) moralities

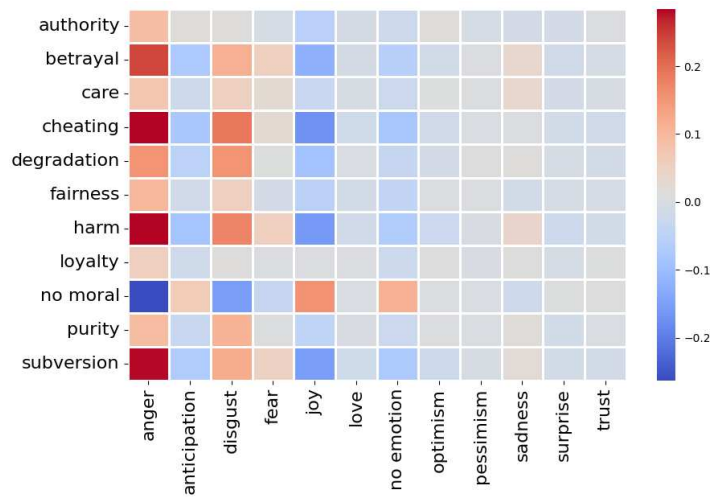
all three datasets, the labels are generated by MoralBERT. Note that MoralBERT is originally trained on MFTC and MFRC, hence the assigned labels for these datasets can be considered reasonably robust and, consequently, the corresponding heatmaps serve as a benchmark for comparison with the heatmap from E2MoCase. The three heatmaps show a high degree of similarity, suggesting firstly that MoralBERT consistently labels texts across the datasets, i.e. it assigns the same label ($m_1 = m_2$)



(a)



(b)



(c)

Fig. 8.3: Pearson's correlation of morality labels (a), emotion labels (b), and between emotions and morality labels (c) in E2MoCase.

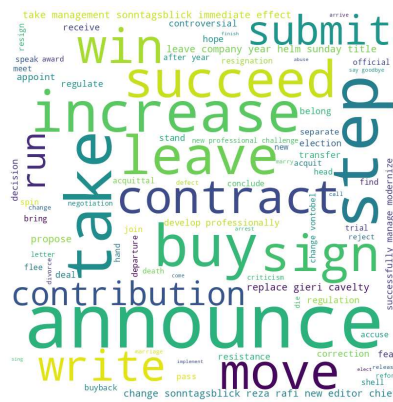
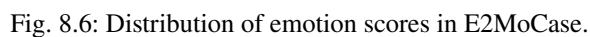
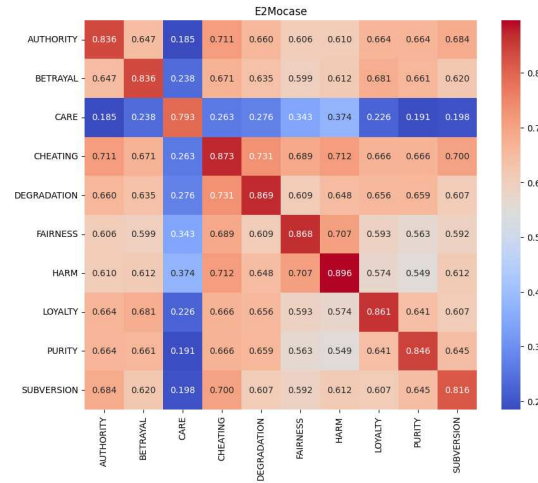


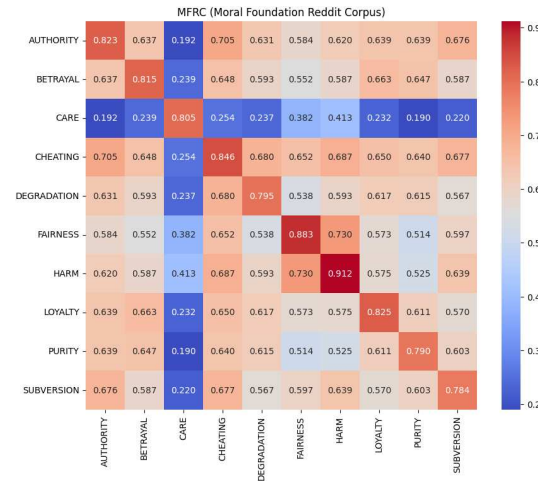
Fig. 8.5: Word cloud of events for texts with no emotion and no morality.



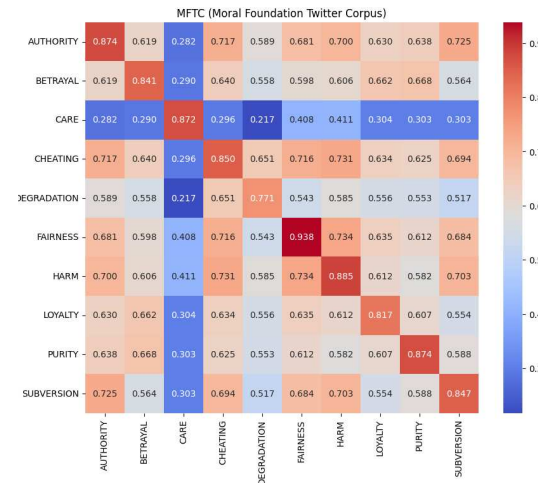
to texts with high similarity (likely discussing similar topics) and different labels ($m_1 \neq m_2$) to texts with lower similarity.



(a)



(b)



(c)

Fig. 8.7: Similarity heatmaps of moral values on E2MoCase (a), MRFC (b) and MTFC (c).

8.5 Dataset Validation

Since the emotions, moralities, and events in E2MoCase have been obtained automatically (i.e., through MoralBERT, EmoLLaMA, GoLLiE), it is essential to provide some guarantee of the reliability of these labels. To this purpose, our approach was to train models on E2MoCase for tasks of moral value detection and emotion detection in order to validate the test performances on each of the three dimensions separately.

Table 8.2: F1-score performance results on the morality classification task. Columns correspond to test datasets.

MFTC						
	care	fairness	loyalty	authority	purity	overall F1
BERT _{MFTC}	0.407	0.358	0.189	0.259	0.199	0.282
BERT _{MFRC}	0.663	0.533	<u>0.312</u>	0.396	<u>0.272</u>	0.491
BERT _{E2Mo}	<u>0.481</u>	<u>0.495</u>	0.347	0.396	0.314	<u>0.406</u>

MFRC						
	care	fairness	loyalty	authority	purity	overall F1
BERT _{MFTC}	0.770	0.736	0.589	0.679	0.604	0.691
BERT _{MFRC}	<u>0.619</u>	0.475	0.296	0.321	0.250	0.392
BERT _{E2Mo}	0.544	<u>0.581</u>	<u>0.432</u>	<u>0.380</u>	<u>0.314</u>	<u>0.450</u>

Validation of moral labels. Considering the moral dimensions, We examined the validity of the labels in E2MoCase for detecting moral values in human-annotated texts. Specifically, We fine-tuned a BERT base model (bert-base-uncased) on E2MoCase for morality classification and assessed the model’s performance on two state-of-the-art datasets for moral values classification, based on the Moral Foundation Theory: MFRC and MFTC (cf. Section 2.4.3). We considered the task of moral value classification in a *multi-label setting*. We binarized the MoralBERT-based numerical scores associated with morality values in E2MoCase (i.e., scores equal to or greater than 0.5 were mapped to 1, otherwise to 0). We performed a K-Fold cross-validation strategy, with $K = 5$. Following other works (Trager et al., 2022; Araque et al., 2020b), We adopted the common approach of reducing the moral labels to 5, without distinguishing between virtues and vices (i.e., *care* refers to either *care* and *harm*). The fine-tuning was conducted for 5 epochs with a batch size of 8, using the Adam optimizer, and a learning rate of $5e-5$.

In Table 8.2, We compare the results on MFRC and MFTC test sets obtained by BERT trained on E2MoCase (dubbed BERT_{E2Mo}), BERT trained on MFTC (BERT_{MFTC}) and BERT trained MFRC (BERT_{MFRC}), under the same settings described above. We report the mean F1 score for each moral label, along with an overall evaluation of morality detection using the averaged micro F1 score (overall F1). Bold values, resp. underlined values, correspond to the best, resp. the second-best, results for each moral value (analogous remarks holds for the subsequent

Table 8.3: F1-score performance results on the emotion classification task. Columns correspond to test datasets.

SemEval							
	anger	disgust	fear	joy	optimism	sadness	overall F1
BERT _{SemEval}	0.579	0.557	0.399	0.607	0.461	0.418	0.510
BERT _{E2Mo}	0.599	0.577	0.158	0.538	0.011	0.149	0.443

DailyDialogue						
	anger	disgust	fear	joy	sadness	overall F1
BERT _{SemEval}	0.448	0.153	0.147	0.689	0.466	0.559
BERT _{E2Mo}	0.379	0.112	0.061	0.790	0.376	0.515

GoEmotion						
	anger	disgust	joy	optimism	sadness	overall F1
BERT _{SemEval}	0.196	0.138	0.176	0.169	0.128	0.161
BERT _{E2Mo}	0.346	0.222	0.204	0.022	0.053	0.224

AffectiveText						
	anger	disgust	fear	joy	sadness	overall F1
BERT _{SemEval}	0.205	0.114	0.331	0.477	0.316	0.268
BERT _{E2Mo}	0.179	0.099	0.112	0.548	0.110	0.199

tables). As expected, the *in-domain* models, i.e. models trained and tested on the same dataset, generally achieve the best performance. More interestingly, in most cases BERT_{E2Mo} outperforms its *out-of-domain* counterparts (i.e., BERT_{MTFC} for MFRC and BERT_{MFRC} for MFTC), which are trained on human-annotated labels. In few cases, BERT_{E2Mo} even outperforms the in-domain models (e.g., for the Loyalty label in the MFRC dataset).

Based on the above results, there is a basis for asserting that the morality labels of E2MoCase are consistent with human judgment, given that the model trained on these labels performs sufficiently well in detecting human-provided labels. In addition, Fig. 8.7 provides further evidence on the validity of the moral labels of E2MoCase, showing that: (i) texts with the same label are sufficiently similar to each other and distinct from texts with different labels, and (ii) the labeling patterns across E2MoCase, MFRC, and MFTC datasets are consistent to each other.

Validation of emotion labels. Similarly, We tested the validity of the emotion labels of E2MoCase. In this case, We used the data from (Mohammad et al., 2018) as a comparison, for two main reasons: the emotion labels in this dataset are human-annotated and correspond to those of E2MoCase. As described in Section 8.3, the aim of Mohammad et al. (2018) is to evaluate various computational methods for detecting affective states, such as emotions and sentiments, expressed in tweets. We employ the data designed for emotion classification (E-c) task, hereinafter referred to as *SemEval*, containing about 11k instances. However, looking at Fig. 8.2a, We observe that the labels *trust*, *surprise*, *love* and *pessimism* are almost absent in

Table 8.4: F1-score results on moral and emotion classification tasks, with E2MoCase as test set and trigger words.

Moral Classification							
	care	fairness	loyalty	authority	purity	overall F1	
tf.idf+mlp _{par}	0.567	0.691	0.623	0.632	0.649	0.638	
tf.idf+mlp _{trig}	0.335	0.369	0.382	0.356	0.379	0.366	
tf.idf+mlp _{par-trig}	0.587	0.715	0.656	0.662	0.668	0.668	

Emotion Classification								
	anger	anticipation	disgust	fear	joy	optimism	sadness	overall F1
tf.idf+mlp _{par}	0.888	0.168	0.691	0.234	0.466	0.155	0.210	0.729
tf.idf+mlp _{trig}	0.884	0.087	0.703	0.145	0.359	0.055	0.080	0.736
tf.idf+mlp _{par-trig}	0.895	0.195	0.674	0.240	0.493	0.145	0.244	0.728

E2MoCase. Since it would not be possible to validate E2MoCase on these labels, We excluded them from the validation process and fine-tuned a BERT-base-uncased model on both E2MoCase and *SemEval* for the remaining labels. We employed the same training settings as previously described for morality classification. However, here the emotion labels of E2MoCase were binarized by selecting any emotion score greater than 0. This is due to the fact that, as shown in Figs. 8.2a and 8.2b, while the morality scores in E2MoCase are evenly distributed, the emotion scores in E2MoCase are highly imbalanced and the values are mostly below 0.5 (Fig. 8.6). Firstly, We tested the performance of BERT trained on E2MoCase, referred to as BERT_{E2Mo}, and BERT trained on SemEval, referred to as BERT_{SemEval}, on the SemEval test set to compare in-domain and out-of-domain performance. Additionally, We evaluated the performance of both models on other three emotion classification datasets, selecting only the instances with labels matching with those used to train BERT_{E2Mo} and BERT_{SemEval}, namely:

- *DailyDialogue* (Li et al., 2017): a manually-labeled dataset of conversations about daily life. We considered about 1k instances from the a test set with labels *anger*, *disgust*, *fear*, *joy* and *sadness*.
- *GoEmotion* (Demszky et al., 2020): a dataset consisting of Reddit comments with 27 human-annotated emotions. From the test set, We selected 880 instances with labels *anger*, *disgust*, *joy*, *optimism* and *sadness*.
- *AffectiveText* (Strapparava and Mihalcea, 2007): a dataset containing human-annotated headlines from sources like CNN and newspapers, with a development set of 250 instances. We considered the human-annotated labels *anger*, *disgust*, *fear*, *joy* and *sadness*.

Table 8.3 presents the results on the aforementioned datasets. Similar to Table 8.2, the average F1 scores for each label and an overall averaged micro F1 score are shown. On SemEval, BERT_{E2Mo} proves to be competitive with its in-domain counterpart for the labels *anger*, *disgust*, and *joy*, while it reports very low scores for the labels *fear*, *optimism*, and *sadness*. This is not surprising, since *fear*, *optimism*, and *sadness* are underrepresented in E2MoCase, as shown in Fig. 8.2a. This may have impacted

on the training process of $BERT_{E2Mo}$, making it less effective to recognize examples with these emotions. Similarly, low performance on these labels are observed on the DailyDialogue, GoEmotions, and AffectiveText datasets. For the remaining labels, $BERT_{E2Mo}$ proves to be competitive, and in some cases superior, to $BERT_{SemEval}$ on these datasets. On DailyDialogue, $BERT_{E2Mo}$ recognizes the *joy* label, with an average F1 score of 0.79, and achieves an overall micro F1 score close to that of $BERT_{SemEval}$. On GoEmotion, $BERT_{E2Mo}$ outperforms $BERT_{SemEval}$ on the *anger*, *disgust*, and *joy* labels, and achieves the highest overall micro F1 score. On AffectiveText, $BERT_{SemEval}$ generally proves to be the better model. This could be due to the fact that this dataset is part of the SemEval 2007 competition, while $BERT_{SemEval}$ is trained on a dataset from the SemEval 2018 competition. Since they belong to the same competition, they may share common features. Nevertheless, $BERT_{E2Mo}$ achieves better performance than $BERT_{SemEval}$ on the *joy* label.

However, we emphasize that the emotion-classification task was conducted to validate the generated labels, but E2MoCase is not designed for such task. Our intent was to demonstrate that a model trained on our dataset can satisfactorily label human-annotated data, thereby confirming the reliability of its training labels. This validation focuses on the most represented emotions, where sufficient examples enable the effective training of the model. Unlike balanced emotion-datasets, E2MoCase reflects the natural imbalance of emotion distribution in real-world controversial issues. Therefore, by applying undersampling or data augmentation to balance the data, the nature of the dataset would be compromised. Consequently, we selected human-annotated examples from these balanced datasets that correspond to well-represented labels in E2MoCase. For these labels, the model trained on E2MoCase performs comparably to—and in some cases better than—the same model trained on human-annotated datasets.

Validation of events. Finally, We provide preliminary tests on the validity of the events provided by E2MoCase. Unlike morality and emotions, events are not numerical scores but rather strings in JSON format containing actions and subjects involved. In this case, rather than evaluating a model trained on E2MoCase for event extraction—a complex task that would require a considerable and unnecessary implementation effort at this stage—We assess whether the actions, which serve as the *trigger* elements of an event, are useful for identifying emotions and moralities. More specifically, We encode paragraphs and triggers of E2MoCase using TF-IDF vectorial representation. Hence, We train an MLP model for moral and emotion classification using paragraphs, triggers, or the concatenation of both as input. We choose TF-IDF as the encoding strategy instead of language models like BERT because the latter benefit from contextual information, which is absent in our representation of events. Table 8.4 shows the performance, in terms of F1 scores, obtained by using TF-IDF representation of paragraphs only ($tf.idf+mlp_{trig}$), of triggers only ($tf.idf+mlp_{par}$), or both ($tf.idf+mlp_{par+trig}$) as input. Notably, combining triggers and paragraphs improves performance in most cases as concerns the emotion classification task, and in all cases as concerns the moral value classification task.

8.6 Chapter Review

We proposed E2MoCase, the first dataset designed to link moralities, emotions and events to press narratives of legal cases. These three aspects are pivotal for understanding how media biases shape public perception of legal issues. While the dataset was labeled using automatic procedures based on well-recognized domain-specific TLMs, our validation methodology showed that E2MoCase enables NLP tools to achieve competitive performance on human annotated datasets.

Future work will focus on expanding the dataset to include a wider array of sources and languages, increasing the robustness of the analysis. In particular, it would be interesting to link the retrieved news with official legal documents of the associated legal trials. We also aim to refine the tools used to detect events and classify emotional and moral content, improving the precision with which biases can be identified.

Besides recognizing the need for a more robust validation process support on human annotators, as well as the opportunity for maintaining E2MoCase with fresh news from SwissDox or other similar reliable sources, one further limitation is the linguistic scope. Currently, the collection only contains articles translated into English, which may result in the loss of complex meanings found in the original language versions. This may affect annotations of emotions and moral qualities, particularly in languages where cultural factors heavily influence emotional expression and moral interpretation. We emphasize that the original paragraphs of E2MoCase cover different languages, such as French, German and Italian.

Also, cross-cultural analysis of moral and emotional representations is another area that requires further exploration. The same legal case may evoke different moral and emotional responses in different countries due to cultural and social differences. Cross-cultural studies, such as retrieving news of a case from different countries' media, could provide insights into these differences, increasing the dataset's applicability and impact in global research contexts and providing a broader spectrum of media representations of legal narratives.

Chapter 9

ME²-BERT: Are Events and Emotions what you need for Moral Value Prediction?

Moral foundations, emotions, and events are complex aspects of human cognition, which are often treated separately since capturing their combined effects is challenging, especially due to the lack of annotated data. Leveraging their interrelations hence becomes crucial for advancing the understanding of human moral behaviors. Moral foundations embedded in language, reflecting what people think is "right" and "wrong", serve as a mirror for understanding human behaviors (Van de Poel and Royakkers, 2023). Emotions, such as anger and joy, are complex states of feeling that result in physical and psychological reactions influencing both thought and behavior (Cambria et al., 2012). Emotions are closely linked with morality (Horne and Powell, 2016), indicating that leveraging emotional information can simplify morality analysis tasks. Events, acting as occurrences involving one or more entities, can encapsulate a wide range of contextual information (e.g., actions, participants), which can be mapped to moral foundations. Thus, events might be helpful as a new layer of representation for textual documents, enabling their analysis from different perspectives, particularly for morality analysis tasks, where contextual information are crucial (Haidt and Graham, 2007).

We propose a TLM-based framework, namely ME²-BERT, for Moral Value Prediction through **E**vents and **E**motions, achieved by fine-tuning **BERT** (Zangari et al., 2025). ME²-BERT is specifically designed for leveraging emotions, events and moralities during the learning process, and can generalize across different types of data never seen during the training process. We devise a domain-adaptation strategy for learning robust domain invariant representations based on events, i.e., by aligning to a common embedding space both texts that contain events and texts that do not contain events. This is accomplished through a denoising auto-encoder, which is also constrained to learn emotion-aware text encodings through a contrastive learning strategy. The learned embeddings are then fed to an adversarial classifier, for learning domain-invariant representations, and to a moral foundation classifier, which performs the main task. ME²-BERT is fine-tuned on the E2MoCase dataset—specifically on its **E2MoCase.full** version—which, to the best of our knowledge, is the only available dataset linking moralities, emotions and events together within textual data (news articles). We argue that by forcing the model to learn domain-

invariant representations from event-based domains, which both contain news from diverse media outlets, the model can effectively adapt to new datasets with different linguistic styles and biases, thus becoming a general moral foundation classifier.

We summarize our contributions as follows:

1. We design a novel holistic framework for fine-tuning TLMs by integrating moralities, emotions, and events. To the best of our knowledge, ME²-BERT is the first framework which jointly leverage emotions and events for moral value prediction.
2. We propose a domain identification strategy based on events, and an emotion-aware denoising auto-encoder module, which acts as adversary of an event-based domain classifier for learning domain-invariant representations.
3. Our experimental evaluation on existing datasets with moral foundation annotations has shown the significance of our framework against several methods for moral foundation prediction, including lexicon-based, BERT-based and TLM approaches, with average percentage increase in F1-score ranging from 15% to 33% in average. Also, the experimental results indicate the usefulness of all components of ME²-BERT, including those exploiting emotions and events.

Why including events and emotions for moral value prediction? When addressing moral foundation analysis tasks, relying solely on textual information may be insufficient, as morality involves a wide array of human behaviors, including emotions and situational contexts (Haidt and Joseph, 2004a). Moral principles are deeply embedded in the events people experience. Events, defined as specific occurrences at a particular time and place involving one or more participants (Xiang and Wang, 2019), provide contextual information, often marked by trigger verbs or nouns. Events do not merely provide background; they actively contribute to the moral interpretation of situations. They also serve as powerful tools for data augmentation, which is crucial in the field of morality, due to the scarcity of reliable, annotated data (Kobbe et al., 2020). By incorporating events, we can diversify the input without needing multiple data sources, thus allowing models to better generalize and learn domain-invariant representations. At the same time, the link between emotions and morality is well-established, though the exact nature of this relationship is still debated (Cameron et al., 2015).

By jointly integrating events and emotions for moral value prediction, we gain a twofold advantage. First, the use of events within texts creates a domain shift in data distribution, allowing the framework to learn domain-invariant representations that enhance its ability to generalize across different domains. Second, by leveraging emotions, the model can better capture the subjective nature of morality due to the intrinsic link between these two aspects (Tekin and Ekici, 2023). Fig. 9.1 shows an example of text associated with emotions, moralities and events.

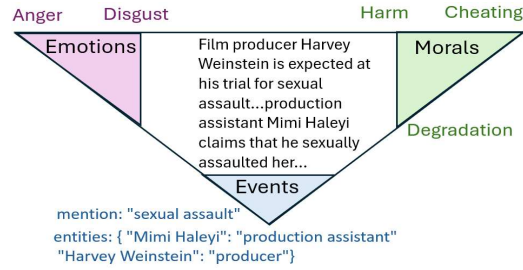


Fig. 9.1: Text segment with emotions, moralities and events extracted from E2MoCase.

9.1 Related Works

Early works for moral value prediction are lexicon-based, utilizing lists of words linked to moral foundations. The Moral Foundations Dictionary (MFD) (Graham et al., 2009) is one of the first lexicons developed for this purpose. Its extensions, MFDv2 (Frimer, 2019) and the extended Moral Foundation Dictionary (eMFD) (Hopp et al., 2021), expand the MFD by including more words and a larger set of annotated text samples. Distributed Dictionary Representations (DDR) (Garten et al., 2018) combines the MFD with word embeddings, representing a concept in the semantic space through the vector representation of the words in the MFD. MoralStrength (Araque et al., 2020a) and LibertyMFD (Araque et al., 2022) further enhance the MFD by quantifying the relevance and strength of words related to the five moral foundations. These methods rely on predefined word lists and lack adaptability to diverse linguistic contexts.

Recent works have collected datasets annotated with moral foundations and used them for training deep-learning models. Hoover et al. (2020b) introduced the Moral Foundation Twitter Corpus (MFTC), which comprises seven distinct Twitter datasets focused on morally relevant issues, whose validity was tested using a LSTM model. In a similar vein, Trager et al. (2022) developed the Moral Foundation Reddit Corpus (MFRC) and established a series of baseline models employing BERT. Liscio et al. (2022) approached each dataset within the MFTC as a separate domain and trained BERT under different configurations.

DAMF (Guo et al., 2023a) and MoralBERT (Preniqi et al., 2024) are approaches relatively close to our work. In fact, both employ domain adaptation to train BERT on heterogeneous data sources; however, DAMF was trained in a semi-supervised fashion, having access to some data from the target domain, while MoralBERT showed a significant drop in performance in out-of-domain scenarios, indicating that it cannot effectively handle the domain shift problem between source and target data. Zhang et al. (2024) released the MoralEvent dataset, comprising of news article annotated with events and moralities, while E2MoCase connected moralities, events and emotions.

Like DAMF and MoralBERT, we employ a domain adaptation strategy to build a model for morality inference on unseen data. However, we leverage events and emotions to build a more robust domain-invariant representation. While events, often treated as graph-structured data, are mainly used for text augmentation (Shorten et al., 2021), and emotions are closely linked to morality (Tekin and Ekici, 2023; Ugazio et al., 2012), their combined use in deep learning for moral foundation prediction has remained unexplored so far. Their joint use enables training a domain-adaptive model driven by emotional information while relying on a single data source rather than depending on several heterogeneous datasets.

9.2 Preliminaries

In this section, we introduce preliminary definitions. A table of the notations used is provided in Table C.2. We define the problem statement and provide an overview of the unsupervised domain adaptation paradigm. It is worth noting that the problem statement pertains to the moral foundations prediction task detailed in Section 2.3.3. However, we restate the problem here, using slightly different notation to avoid unnecessary symbolic representations for the purposes of this work.

Problem definition: Moral Value Prediction. We define moral foundation prediction as the task of predicting the moral foundations contained in textual data according to the principles outlined by the MFT. We are given a dataset $\mathcal{D} = \{(d_i, \mathbf{y}_i)\}$, where d_i represents a text and $\mathbf{y}_i \in \mathbb{R}_*^{|\mathcal{M}|}$ is a vector of non-negative real-value scores associated with the dimensions of the morality theory, i.e., in our setting MFT with $|\mathcal{M}| = 5$. Our goal is to learn a function $f : \mathcal{T} \rightarrow \{0, 1\}^{|\mathcal{M}|}$ that predicts the moral values for any $d_i \in \mathcal{T}$, where $\mathcal{T}$ is the space of all possible input texts. Whenever none of the MFT dimensions are present in a text d_i , its content is treated as *non-moral* (*Nm.*) (Hoover et al., 2020b). Following other works (Trager et al., 2022), we frame the problem as a multi-label classification task. This is in line with the fact that most of the existing moral data are provided with binary labels.

Domain adaptation. *Unsupervised Domain adaptation (UDA)* is a transfer learning technique designed to address domain shifts in data distribution, enhancing out-of-domain prediction performance (Ben-David et al., 2010). UDA involves unsupervised learning as it relies on labeled data from a source domain while assuming no labels for a target domain. Formally, let $\mathcal{D}_s = \{(d_i^{(s)}, \mathbf{y}_i^{(s)})\}_{i=1}^{n_s}$ and $\mathcal{D}_t = \{d_j^{(t)}\}_{j=1}^{n_t}$ denote the labeled source domain and unlabeled target domains, respectively, where n_s and n_t are the number of source and target samples and $n = n_s + n_t$. Due to the domain shift, both the marginal distribution ($\mathbb{P}$) and the conditional distributions ($\mathbb{Q}$) of the two domains differ, i.e., $\mathbb{P}_s(d^{(s)}) \neq \mathbb{P}_t(d^{(t)})$, $\mathbb{Q}_s(\mathbf{y}^{(s)}|d^{(s)}) \neq \mathbb{Q}_t(\mathbf{y}^{(t)}|d^{(t)})$, and are not known a priori. The goal of UDA for moral value prediction is to learn a function $f : \mathcal{T} \rightarrow \mathbb{R}^{|\mathcal{M}|}$ that predicts the moralities of samples from the target domain by minimizing domain discrepancy between source and target (Singhal et al., 2023).

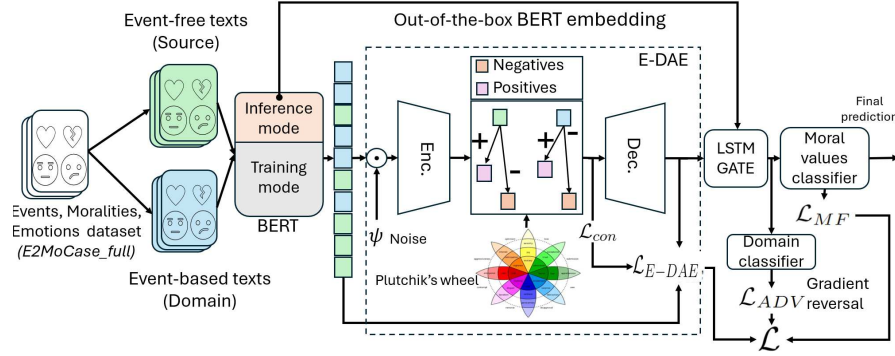


Fig. 9.2: The ME²-BERT framework for moral value prediction. ME²-BERT leverages emotion and event information for learning domain-invariant text representations.

Common domain adaptation methods include *domain adversarial networks*, which introduce a domain discriminator to enforce the learning of domain-invariant embeddings (Ganin et al., 2016). This approach has been effective in predicting moral foundations from text (Guo et al., 2023a). Another strategy, based on the idea that a domain-adaptive framework should effectively reconstruct target domain data, involves learning domain-invariant embeddings using a *denoising auto-encoder* (DAE) (Vincent et al., 2008). The DAE is trained to reconstruct the BERT embeddings while being informed with emotional information using a contrastive learning strategy. The reconstructed embeddings are fed into both an adversarial-learning-based domain classifier (Ganin et al., 2016) to map the embeddings from different domains into a common space, and a moral foundation classifier. Note that the use of MFT may carry inherent cultural biases and potentially conflict with specific cultural contexts. Nonetheless, we want to emphasize that the conceptual

9.3 Methodology

The overall architecture of ME²-BERT is shown in Fig. 9.2. ME²-BERT is a framework for fine-tuning TLMs incorporating (i) events to learn domain-invariant text embeddings, and (ii) emotions to drive the learning of morally-relevant text embeddings for the moral value prediction task. Firstly, source and target domains are selected based on whether they do not contain events or contain events, respectively. In our training dataset, i.e., E2MoCase, this events-based selection ensures a strong domain shift between the two types of corpora. Then, a BERT-based model generates the embeddings for source and target texts, which are input into a *denoising auto-encoder* (DAE) (Vincent et al., 2008). The DAE is trained to reconstruct the BERT embeddings while being informed with emotional information using a contrastive learning strategy. The reconstructed embeddings are fed into both an adversarial-learning-based domain classifier (Ganin et al., 2016) to map the embeddings from different domains into a common space, and a moral foundation classifier. Note that the use of MFT may carry inherent cultural biases and potentially conflict with specific cultural contexts. Nonetheless, we want to emphasize that the conceptual

design of our proposed framework is general and not inherently dependent on the MFT. That is, our framework can equally be adapted to other theories of morality.

Event-based domain identification. By exploiting the fine-tuning of ME²-BERT on the E2MoCase dataset, we use events to create a domain shift in features and labels distribution between text segments without events (referred to as source domain) and text segments with events (referred to as target domain), thus forcing the model to learn robust, domain-invariant representations. When events are available, we model a text segment as a JSON object representing each event as a tuple of its corresponding trigger words and involved entities — as shown in Fig. 9.3 — to provide BERT with contextual information and to enable it to extract meaningful text embeddings. Indeed, it is common for a narrative to be introduced through events, e.g., to mold public opinion (Zhang et al., 2021b). Also, our distinction of the source and target texts is useful to make the model align event-free and event-based text representations, enhancing its generalization capabilities.

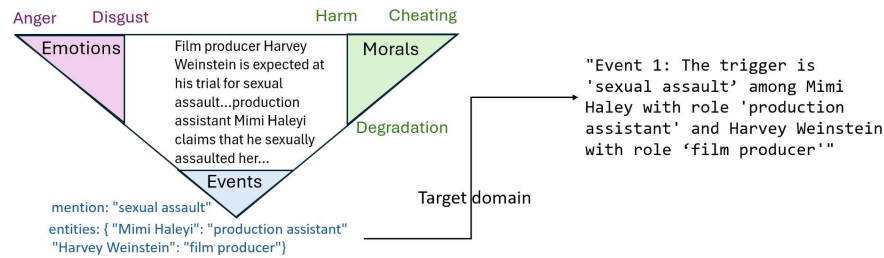


Fig. 9.3: Event encoding in descriptive text string.

Our domain identification strategy could not always result in a domain shift, and in general, it is required to analyze the fine-tuning data. Fig. 9.4a shows the UMAP plot (McInnes et al., 2018a) of the BERT embeddings of the source and target domains on E2MoCase, which differ in the embedding space, while Fig. 9.5 shows the labels distribution, revealing that they have also different moral focus. The predominance of non-moral data in the source domain suggests that event-free paragraphs often carry “neutral” content from the morality viewpoint. This can lead to data imbalance issues (Liscio et al., 2022).

Integration of emotional values. We provide emotional information through the Plutchik’s wheel of emotions (Plutchik, 2001), which defines eight basic emotions arranged in pairs of opposites (cf. Appendix C.5). We group the primary emotions of the Plutchik’s wheel into opposing pairs, resulting in four broad categories: *anger/fear*, *trust/disgust*, *joy/sadness*, and *surprise/anticipation*. These categories are used within our contrastive learning framework to guide the model’s training to incorporate emotional state information in the prediction of moral foundations.

Emotion-aware Denoising Auto-Encoder. The Emotion-aware Denoising Auto-Encoder (E-DAE) module in our ME²-BERT is designed to learn robust embeddings

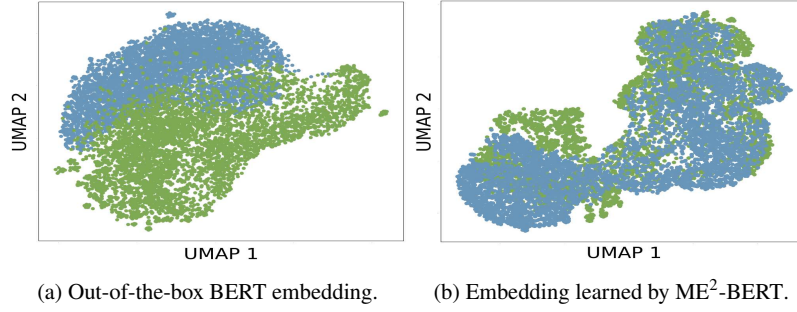


Fig. 9.4: UMAP embedding. Green and blue points show the [CLS] token encoding of the source and target domains, respectively.

leveraging emotion information for both event-based and event-free texts (i.e., domain). By utilizing a denoising auto-encoder as a transformation function, we filter out domain-specific noise and features contained in the BERT-based encodings (Guo et al., 2023a). This approach improves robustness and generalizability so as to better support the alignment of embeddings across different domains (Wang et al., 2021a; Lopez-Avila and Suárez-Paniagua, 2023). Following Clinchant et al. (2016), we apply the E-DAE to the union of the source and target samples. Given d_i , let $\mathbf{x}_i = \text{BERT}(d_i)$ be its BERT embedding generated during training. We corrupt $\mathbf{x}_i$ with noise ψ , i.e., $\mathbf{x}'_i = \mathbf{x}_i \odot \psi$, where ψ is a random variable sampled from a Bernoulli distribution with probability p , and $\odot$ indicates element-wise multiplication. The auto-encoder aims to reconstruct the BERT embeddings without noise ($\mathbf{x}_i$) given the corrupted embeddings $\mathbf{x}'_i$:

$$\mathbf{h}_i = f_{enc}(\mathbf{x}'_i), \quad \hat{\mathbf{x}}_i = f_{dec}(\mathbf{h}_i), \quad (9.1)$$

where $\hat{\mathbf{x}}_i$ is the reconstructed encoding from the auto-encoder; f_{enc} and f_{dec} are the encoder and decoder functions, resp., both corresponding to two-layer MLPs; $\mathbf{h}_i$ is the bottleneck representation learned by the encoder. The reconstruction loss is given by the Mean Squared Error (MSE):

$$\mathcal{L}_{\text{MSE}} = \frac{1}{n} \sum_{i=1}^n \|\mathbf{x}_i - \hat{\mathbf{x}}_i\|^2, \quad (9.2)$$

where n is the number of text segments. To favor the discovery of latent representations that are emotion-aware, we add a separate contrastive term for each domain that encourages samples with the same emotions to be closer in the embedding space, while those with different emotions are pushed far apart. This contrastive objective ensures that the learned representations capture emotion information, which is crucial for morality analysis tasks (Tekin and Ekici, 2023). Eq. 9.3 defines the source loss $\mathcal{L}_{\text{tr}}^{(s)}$, which is a triplet margin objective:

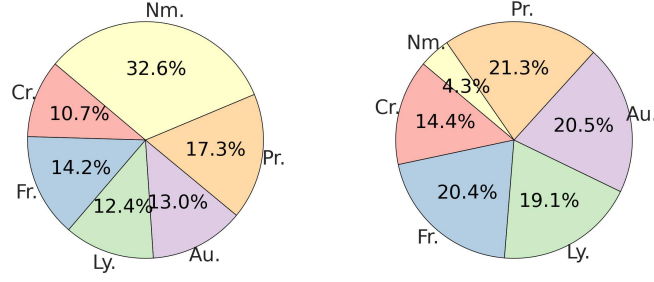


Fig. 9.5: Distribution of the moral labels for source domain (on the left) and target domain (on the right).

$$\mathcal{L}_{tr}^{(s)}(i, p, r) = \max\left(0, c(\mathbf{h}_i, \mathbf{h}_p) - c(\mathbf{h}_i, \mathbf{h}_r) + m\right), \quad (9.3)$$

where $\mathbf{h}_p$ and $\mathbf{h}_r$ are the latent representations of the positive and negative samples associated with the text segment d_i from the source domain; m is the margin enforced between positive and negative pairs and $c(\cdot, \cdot)$ is the cosine similarity distance. Analogous contrastive term is used for the samples belonging to the target domain ($\mathcal{L}_{tr}^{(t)}$). By applying a separate contrastive term to the source and target domains we ensure that important emotional information are preserved, thus preventing the loss of domain-specific information which might be blurred during the denoising process (Kang et al., 2019; Lopez-Avila and Suárez-Paniagua, 2023). The pairs of positive and negative samples are selected based on the Plutchik's categories described in Section 9.3, where each text instance is associated with the emotions having the highest score in E2MoCase. The overall loss function for training the E-DAE corresponds to the sum of reconstruction loss and the contrastive loss:

$$\mathcal{L}_{E-DAE} = \mathcal{L}_{MSE} + \mathcal{L}_{con}, \quad (9.4)$$

where $\mathcal{L}_{con} = \mathcal{L}_{tr}^{(s)} + \mathcal{L}_{tr}^{(t)}$, and $\mathcal{L}_{tr}^{(s)}$, $\mathcal{L}_{tr}^{(t)}$ are the source and target contrastive terms, respectively.

LSTM gate mechanism. The reconstruction provided by the E-DAE module may lead to the loss of semantic information inherent in the original text. To overcome this issue, we propose a LSTM-style gated fusion mechanism that integrates general pre-trained knowledge with the fine-tuned representations. Let $\mathbf{x}_i^{(oob)}$ denote the embedding produced by the out-of-the-box BERT model, and $\hat{\mathbf{x}}_i$ the embeddings produced by E-DAE. The gating mechanism is defined as:

$$\begin{aligned} \mathbf{f}_i &= \sigma(\mathbf{W}_f \mathbf{x}_i^{(oob)} + \mathbf{b}_f), & \mathbf{q}_i &= \sigma(\mathbf{W}_q \hat{\mathbf{x}}_i + \mathbf{b}_q), \\ \mathbf{c}_i &= \mathbf{f}_i \odot \mathbf{x}_i^{(oob)} + \mathbf{q}_i \odot \hat{\mathbf{x}}_i, & \mathbf{o}_i &= \sigma(\mathbf{W}_o \mathbf{c}_i + \mathbf{b}_o), \\ \tilde{\mathbf{x}}_i &= \mathbf{o}_i \odot \tanh(\mathbf{c}_i), \end{aligned} \quad (9.5)$$

where σ is the sigmoid function; $\mathbf{W}_f, \mathbf{W}_q, \mathbf{W}_o$ are weight matrices; $\mathbf{b}_f, \mathbf{b}_q, \mathbf{b}_o$ are bias vectors; $\odot$ denotes element-wise multiplication and $\mathbf{f}_i, \mathbf{q}_i$ and $\mathbf{o}_i$ are the forget, input and output gates, respectively. $\mathbf{f}_i$ and $\mathbf{q}_i$ determine the amount of information to retain from $\mathbf{x}_i^{(oob)}$, and $\tilde{\mathbf{x}}_i$, resp., while $\mathbf{c}_i$ is the vector state combining the out-of-the-box and the reconstructed embeddings; $\mathbf{o}_i$ controls the contribution of $\mathbf{c}_i$ to the final gated representation $\tilde{\mathbf{x}}_i$.

Event classifier. To strengthen the domain invariance effect, we integrate an adversarial learning module applied to the gated representations, i.e., $\tilde{\mathbf{x}}_i$. By introducing an adversarial discriminator that challenges the model to produce embeddings indistinguishable across domains—thus acting as adversary of the E-DAE module—we further enforce the learning of domain-invariant embeddings (Ganin et al., 2016; Guo et al., 2023a). The loss function for the domain classifier is defined in Eq. 9.6, which is to be maximized during the training through a gradient reversal layer (Ganin et al., 2016):

$$\mathcal{L}_{ADV} = -\frac{1}{n} \sum_{i=1}^n \left[y_i^{(D)} \log(\sigma(g(\tilde{\mathbf{x}}_i))) + (1 - y_i^{(D)}) \log(1 - \sigma(g(\tilde{\mathbf{x}}_i))) \right], \quad (9.6)$$

where g is the domain classifier (MLP), which aims to distinguish between the two domains; σ is the sigmoid function, and $y_i^{(D)}$ is the ground truth label indicating which domain the input text belongs to.

Optimization. To predict the moral foundations, we use a MLP shared for both domains, which is fed with the denoised and domain-invariant BERT embedding $\tilde{\mathbf{x}}_i$. Since we address the problem as a multi-label classification task, the loss function, i.e., $\mathcal{L}_{MF}$, is a Cross-Entropy with sigmoid activation. ME²-BERT is trained to optimize the following loss:

$$\mathcal{L} = \mathcal{L}_{E-DAE} + \mathcal{L}_{MF} - \lambda_{dom} \mathcal{L}_{ADV}, \quad (9.7)$$

where λ_{dom} is an hyperparameter indicating the importance of the adversarial learning module.

9.4 Computational Complexity aspects

We discuss the time complexity of ME²-BERT, under the following assumptions: the sequence length of each sample in the batch is set to the maximum value T , and the hidden dimension of each neural model out of the BERT encoder is d_b .

Generating the BERT embeddings requires $O(n_b L T^2 d_B)$ (Vaswani et al., 2017), where L is the number of BERT layers, n_b is the number of samples in the batch,

and d_B is the hidden dimension. With a bottleneck layer of size d_b , and using two-layer MLPs for both the encoder and decoder, the autoencoder requires $O(n_b d_B d_b + n_b d_b^2) \approx O(n_b d_B d_b)$, since $d_b \ll d_B$. The LSTM gate requires $O(n_b d_B^2)$ operations, stemming from the linear transformations and the additional element-wise operations described in Eq. 9.5. The moral classifier requires $O(d_B d_b)$ operations. Therefore, during inference, the cost is $O(n_b L T^2 d_B + n_b d_B^2)$.

At training time, the domain classifier adds a cost of $O(d_B d_b)$ and we need to compute the loss functions. The triplet margin loss is the most computationally expensive, requiring cosine distance calculations for up to $O(n_b^2)$ triplets.

9.5 Experimental Evaluation

We design our experimental evaluation to pursue the following objectives: (i) to measure the effectiveness of ME²-BERT w.r.t. SOTA NLP methods and TLMs for moral value prediction; (ii) to carry out an ablation analysis which shows the impact of each component of ME²-BERT; (iii) to validate our event-based domain identification strategy; (iv) to evaluate the performance of ME²-BERT in the single-label setting, where each moral foundation is treated independently, and (v) to analyze the abilities of ME²-BERT to detect the polarity of each moral foundation.

Datasets. As previously said, we fine-tuned ME²-BERT on E2MoCase, specifically the E2MoCase_full dataset. We evaluate all methods on three datasets that are manually annotated with moralities: Moral Foundation Twitter Corpus (MFTC) (Hoover et al., 2020b), Moral Foundation Reddit Corpus (MFRC) (Trager et al., 2022), and the Extended MFD (eMFD) (Hopp et al., 2021). Note that we refer to the dataset as eMFD; other studies (Mokhberian et al., 2022; Nguyen et al., 2024b) refer to this dataset as MNFC (Moral Foundations News Corpus). Here, we follow Guo et al. (2023a) in referring to it as eMFD, but we clarify that we are referring to the dataset and not the dictionary constructed by Hopp et al. (2021). All these datasets are thoroughly described in Section 2.4.3. Furthermore, we also used the MoralEvents dataset (cf. Section 8.1) for the evaluation our domain-identification strategy.

Competing methods. We compare ME²-BERT with different classes of competitors: (i) *BERT-based methods* specifically trained for moral foundation prediction, including BERT-base trained on E2MoCase dataset (hereinafter BERT-E2MoCase), the baselines proposed by Trager et al. (2022), namely BERT trained on MFTC (*BERT-MFTC*) and BERT trained on MFRC (*BERT-MFRC*). Additionally, we include BERT trained on both MFTC and MFRC (*BERT-MFTRC*), *DAMF* and *Moral-BERT*. (ii) *TLMs* with zero-shot inference strategy, which include three of the most recent open models: *Llama-3.1*, *Gemma-2* and *Mistral-Nemo*; (iii) *lexicon-based methods* designed for moral classification, including *MoralStrength* and *DDR*.

9.5.1 Experimental setting

We evaluate the performance of each model in the out-of-domain setting, where training and test data come from different domains. Although ME²-BERT is independent of the TLM used, we employ BERT base uncased to be fair with our competitors—*although, we also experimented with a selection of Sentence-Transformers*. ME²-BERT and DAMF are fine-tuned on the E2MoCase dataset, while all the other BERT-base models are fine-tuned on the MFRC and/or the MFTC datasets (e.g., BERT-MFRC). All BERT-based models are evaluated utilizing a 5-fold cross validation strategy, except for MoralBERT for which we use its single-label version in inference mode, as provided by the authors. Since it uses the the same strategy as DAMF, and it is fine-tuned on MFRC and MFTC datasets, we report its performance solely on eMFD data to maintain consistency with the out-of-domain setting. For TLMs and lexicons, we report the average scores over five runs. We implemented ME²-BERT, BERT-E2MoCase, BERT-MFRC, BERT-MFTC and BERT-MFRC using PyTorch library.¹ For DAMF,², MoralStrength³ and DDR,⁴ we used their publicly available source code. For MoralStrength, we employ the best performing model (on average) according to Araque et al. (2020a). For MoralBERT,⁵ LLaMA-3.1,⁶ Gemma-2⁷ and Mistral-Nemo,⁸ we use the models available on HuggingFace.⁹

Hyper-parameter setting. Fine-tuning for all models was performed over 10 epochs with a batch size of 8. For all BERT-based models, we set a maximum sequence length of 256 due to the length of paragraphs in E2MoCase (about 210 tokens per paragraph on average, with high variability), with padding to the maximum length. We adopted the token [CLS] as text embeddings. For ME²-BERT, we applied noise to the auto-encoder by using the dropout technique (i.e., by randomly setting to zero some values with probability p). We performed a grid search to select the best learning rate and dropout values, i.e., 5.0E-5 and 0.3, respectively. The margin m of the contrastive function dynamically varied from 1 to 0 using an exponential function. The term λ_{dom} was also adjusted exponentially from 0 to 1 (Ganin et al., 2016; Guo et al., 2023a). BERT-E2MoCase, BERT-MFRC, BERT-MFTC, and BERT-MFTRC were fine-tuned using the same hyper-parameter setting as ME²-BERT. All other competing methods were carried out using the default hyperparameters specified in their corresponding source code. Considering TLMs, each model is asked to respond to the prompt described in the following paragraph,

¹ <https://pytorch.org/>

² <https://github.com/fionasguo/DAMF/tree/master>

³ <https://pypi.org/project/moralstrength/>

⁴ <https://github.com/USC-CSSL/DDR>

⁵ <https://github.com/vjosapreniqi/MoralBERT>

⁶ <https://huggingface.co/meta-llama/Meta-Llama-3.1-8B-Instruct>

⁷ <https://huggingface.co/google/gemma-2-9b-it>

⁸ <https://huggingface.co/mistralai/Mistral-Nemo-Instruct-2407>

⁹ <https://huggingface.co/>

with the role of *expert in Moral Foundation Theory*. We set the models' *temperature* to 1.0E-2, so as to reduce randomness and ensure the selection of the most probable tokens for more precise and deterministic responses. We ran each model five times and reported its average performances.

The Uniform Manifold Approximation and Projection (UMAP) (McInnes et al., 2018a) graphs were plotted using the following hyper-parameters: number of neighbors and minimum distance equals to 80 and 0.5, resp., and using cosine similarity as a distance measure. The plots in Figs. 9.4 and 9.10 were generated using 10,000 random samples.

Prompt construction for TLMs. The same prompt is constructed for all models in a zero-shot setting, providing the paragraph and the list of possible moral traits as options. Specifically, the following prompt is provided:

Prompt

Given the following news article paragraph:

{paragraph}

Categorize the text's moral traits as 'neutral' or according to the moral foundation theory with the following: *{list of all possible morals traits}*.

Your response must follow the following pattern: [*i*list of detected moral traits_{*i*}]. Your response must contain just the list of the detected moral traits, do not add any additional word or introductions.

9.5.2 Results

Comparative evaluation. Table 9.1 reports the F1-score values achieved by ME²-BERT and the competing methods on each dataset. ME²-BERT stands out as the best approach yielding an average F1-score of 0.450 across all datasets, while the second best method is DAMF (0.384), which also utilizes a domain adaptation strategy, but it does not leverage events and emotions. Despite the TLMs benefit from extensive pre-training on diverse text sources, the best TLM (Mistral-Nemo) yields low performance than ME²-BERT, especially on the MFTC and MFRC datasets (with percentage decrease of about 20% in average). At the moral foundation level, ME²-BERT consistently performs well across all moral foundations, particularly for the MFRC and MFTC datasets. On the eMFD dataset, ME²-BERT is outperformed by TLMs in 3 out of 5 moral categories. Nevertheless, ME²-BERT remains competitive, achieving performance comparable to TLMs on average.

Ablation analysis. Table 9.2 summarizes the F1-scores by ME²-BERT and several simplified variants. We observe that the best results are achieved by the full

Table 9.1: Comparative evaluation: F1-scores. Best results are in bold, second-best results are underlined.

	Care	Fairness	Loyalty	Authority	Purity	Non-moral	AVG
MFRC							
ME ² -BERT	0.636	0.585	0.345	0.490	0.363	0.669	0.515
DAMF	0.457	<u>0.535</u>	0.260	<u>0.419</u>	0.311	0.520	0.417
BERT-MFRC	<u>0.619</u>	0.475	0.296	0.321	0.250	0.507	0.411
BERT-E2MoCase	0.481	0.495	<u>0.347</u>	0.396	<u>0.314</u>	0.610	<u>0.440</u>
LLaMA-3.1	0.481	0.481	0.298	0.392	0.273	0.010	0.323
Mistral-Nemo	0.496	0.505	0.353	0.405	0.233	0.245	0.373
Gemma-2	0.519	0.484	0.312	0.394	0.252	0.103	0.344
MoralStrength	0.434	0.463	0.209	0.335	0.183	0.120	0.291
DDR	0.465	0.276	0.201	0.377	0.280	<u>0.623</u>	0.371
MFTC							
ME ² -BERT	0.688	0.673	0.551	0.521	0.453	0.546	0.572
DAMF	0.540	0.583	0.486	0.485	0.397	0.459	<u>0.492</u>
BERT-MFRC	0.568	0.499	0.325	0.435	0.342	<u>0.527</u>	0.449
BERT-E2MoCase	0.544	0.581	0.432	0.380	0.314	0.510	0.460
LLaMA-3.1	0.627	<u>0.609</u>	0.469	0.541	<u>0.452</u>	0.144	0.474
Mistral-Nemo	0.631	0.518	0.459	0.483	0.416	0.407	0.486
Gemma-2	<u>0.642</u>	0.603	0.466	<u>0.524</u>	0.419	0.271	0.487
MoralStrength	0.589	0.531	<u>0.501</u>	0.494	0.379	0.442	0.489
DDR	0.503	0.392	0.314	0.375	0.412	0.432	0.405
eMFD							
ME ² -BERT	0.309	0.225	<u>0.217</u>	0.227	0.220	0.373	0.262
DAMF	0.283	0.231	0.192	0.256	0.185	0.317	0.244
MoralBERT	0.279	0.226	0.220	0.235	0.111	0.007	0.196
BERT-MFTRC	0.296	0.229	0.210	0.236	0.064	0.422	0.243
BERT-E2MoCase	0.297	0.161	0.101	0.155	0.037	<u>0.480</u>	0.210
LLaMA-3.1	<u>0.330</u>	<u>0.287</u>	0.145	<u>0.273</u>	<u>0.201</u>	0.207	0.240
Mistral-Nemo	0.314	0.267	0.170	0.270	0.167	0.371	<u>0.260</u>
Gemma-2	0.333	0.293	0.209	0.279	0.178	0.211	0.250
MoralStrength	0.244	0.220	0.208	0.200	0.182	0.436	0.248
DDR	0.210	0.118	0.153	0.191	0.080	0.484	0.206

ME²-BERT, while BERT-E2MoCase is the worst model, demonstrating the effectiveness of the domain-adaptive approach. As emotions and/or events are integrated, the performance tends to increase. The variant with only the DAE module, excluding the adversarial learning and contrastive modules (dubbed ME²-BERT *w/o* A-C.), shows an average performance increase of nearly 10% over BERT-E2MoCase. Emotions also improve performance, as indicated by the scores achieved by the variant not using the contrastive learning module (dubbed ME²-BERT *w/o* C.). The gate component (dubbed ME²-BERT *w/o* G.) also yields benefits.

Table 9.2: Ablation study: F1-scores. Best results are in bold, second-best results are underlined.

	Care	Fairness	Loyalty	Authority	Purity	Non-moral	AVG
MFRC							
BERT-E2MoCase	0.481	0.495	0.347	0.396	0.314	0.610	0.440
ME ² -BERT w/o A.	0.566	0.555	0.294	0.466	0.327	0.672	0.480
ME ² -BERT w/o C.	0.587	0.541	0.278	0.451	0.316	0.636	0.468
ME ² -BERT w/o A-C.	0.583	0.525	0.304	0.421	0.328	0.623	0.464
ME ² -BERT w/o G.	<u>0.629</u>	<u>0.577</u>	0.332	<u>0.481</u>	<u>0.360</u>	0.672	<u>0.510</u>
ME ² -BERT	0.636	0.585	<u>0.345</u>	0.490	0.363	<u>0.669</u>	0.515
MFTC							
BERT-E2MoCase	0.544	0.581	0.432	0.380	0.314	0.510	0.460
ME ² -BERT w/o A.	0.624	0.623	0.525	0.457	0.419	0.509	0.526
ME ² -BERT w/o C.	0.650	0.622	0.528	0.478	0.422	0.517	0.536
ME ² -BERT w/o A-C.	0.615	0.610	0.501	<u>0.518</u>	0.392	0.503	0.523
ME ² -BERT w/o G.	<u>0.687</u>	0.682	<u>0.529</u>	0.514	<u>0.431</u>	<u>0.519</u>	<u>0.560</u>
ME ² -BERT	0.688	<u>0.673</u>	0.551	0.521	0.453	0.546	0.572
eMFD							
BERT-E2MoCase	0.298	0.162	0.101	0.155	0.037	0.480	0.205
ME ² -BERT w/o A.	0.284	<u>0.246</u>	<u>0.233</u>	0.239	0.206	0.297	<u>0.251</u>
ME ² -BERT w/o C.	<u>0.300</u>	0.193	0.179	0.157	0.134	<u>0.427</u>	0.232
ME ² -BERT w/o A-C.	0.284	0.250	0.242	0.218	0.228	0.130	0.225
ME ² -BERT w/o G.	0.298	0.218	0.221	0.239	0.199	0.285	0.243
ME ² -BERT	0.309	0.225	0.217	<u>0.227</u>	<u>0.220</u>	0.373	0.262

Table 9.3: F1-scores of ME²-BERT and BERT-base (BERT-ME) fine-tuned on MoralEvent. Best results are in bold.

	Care	Fairness	Loyalty	Authority	Purity	Non-moral	AVG
MFRC							
ME ² -BERT	0.387	0.226	0.139	0.162	0.144	0.283	0.224
BERT-ME	0.348	0.006	0.000	0.003	0.000	0.636	0.165
MFTC							
ME ² -BERT	0.502	0.162	0.132	0.262	0.206	0.161	0.238
BERT-ME	0.360	0.011	0.000	0.008	0.000	0.376	0.126
eMFD							
ME ² -BERT	0.210	0.219	0.085	0.170	0.089	0.240	0.169
BERT-ME	0.135	0.006	0.000	0.014	0.000	0.459	0.102

Fine-tuning on MoralEvents. To further assess the validity of our event-based domain identification strategy, we fine-tune ME²-BERT and BERT-base on the *MoralEvent* dataset (Zhang et al., 2024), which contains about 400 news article labeled with morals and events. Since emotion labels are not provided, we use the version of ME²-BERT w/o contrastive learning (ME²-BERT w/o C.). Table 9.3 reports the F1-scores achieved by the two methods. The overall performances are

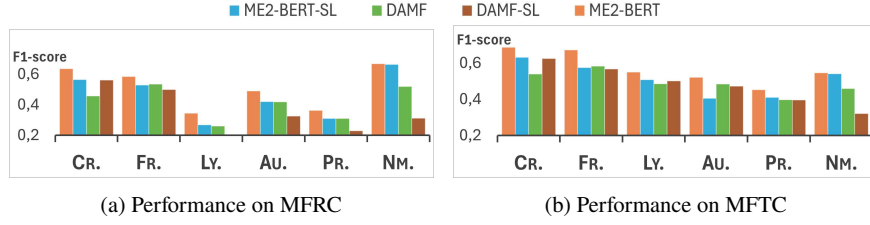


Fig. 9.6: ME²-BERT and DAMF performance in the single-label (-SL) vs multi-label classification settings.

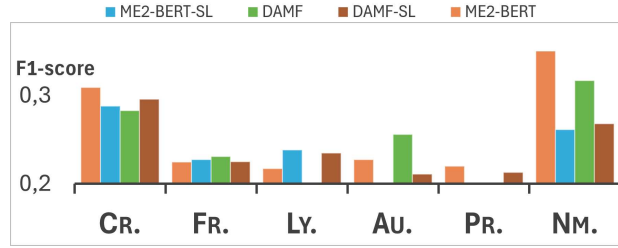


Fig. 9.7: Single-label classification on the eMFD dataset.

lower than the version trained on E2MoCase, likely due to the smaller size of MoralEvent. However, ME²-BERT yields significantly higher F1-scores on all moral foundations, while BERT fails to identify them (both precision and recall are zero for several dimensions), resulting in a high score for the non-moral label.

Single-label classification. We explore the moral value prediction problem in the single-label setting, where a single model is fine-tuned for each moral foundation. Figs. 9.6 and 9.7 show the performance of ME²-BERT and the best competing method, i.e., DAMF, on the MFRC and MFTC datasets. Interestingly, both ME²-BERT and DAMF tend to perform better in the multi-label setting than their single-label counterparts (ME²-BERT-SL and DAMF-SL, respectively). While previous works (Liscio et al., 2022; Trager et al., 2022) found that the multi-label setting is more challenging, we argue that this behavior may stem from their domain-adaptive nature, enabling them to better handle cross-dimension dependencies.

Fig. 9.7 shows the F1 scores yielded by ME²-BERT when trained in multi-label and single-label setting (ME²-BERT and ME²-BERT-SL, respectively), against the best competing method (DAMF and DAMF-SL). Similar to the performance observed in Fig. 9.6 on the MFRC and MFTC datasets, the multi-label models tend to outperform their single-label counterparts also in this case.

Moral polarity detection. Fig. 9.8 shows ME²-BERT’s performance in predicting the polarity (virtue/vice) of each moral foundation, such as care/harm, on the MFTC dataset, compared to the best competing methods for each category (e.g., LLaMA-3.1 for TLMs). Overall, ME²-BERT outperforms the competitors on most moral

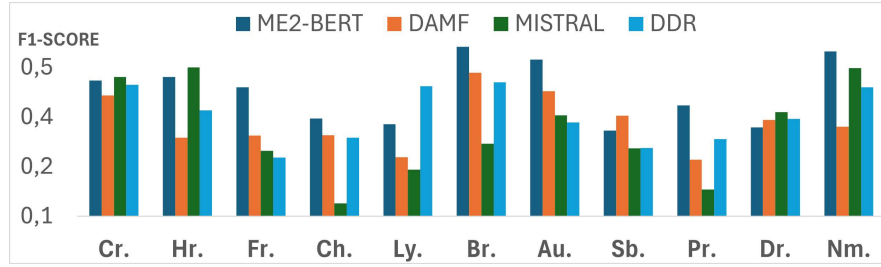


Fig. 9.8: Prediction of virtue/vice moral foundations on the MFTC dataset.

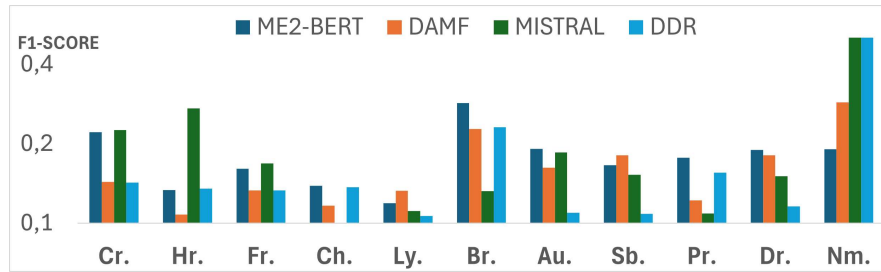


Fig. 9.9: Detection of polarity of moral dimensions on the eMFD dataset.

foundations, yielding the highest average F1-score (0.427), followed by DDR (0.353) and DAMF (0.334).

Fig. 9.9 shows the results of ME²-BERT when compared with the best competing methods of each category (TLMs, lexicon and BERT-based) on the eMFD dataset. In this case, ME²-BERT is the second-best method with an average F1 score of 0.170, outperformed only by Mistral-Nemo, achieving 0.172 F1-score. Overall, Mistral-Nemo appears to be highly effective at predicting the Harm and non-moral polarities. ME²-BERT, on the other hand, is highly effective in predicting the moralities Betrayal (Br.), Purity (Pr.), and Degradation (Dr.). Note that the MFRC dataset does not contain information on the polarity of moral foundations.

Fine-tuning with Sentence-Transformers. In our main experiments, we used *bert-base-uncased* to ensure fairness with TLM-based competitors, which use the aforementioned version of *BERT*. However, our framework is TLM-agnostic. Table 9.4 shows the results of our framework when fine-tuned with different Sentence-Transformers as backbones,¹⁰ such as *bert-base-nli-mean-tokens* (ME²-SBERT) *sentence-t5-large* (ME²-T5), *nli-roberta-base-v2* (ME²-RoBERTa), and *all-mpnet-base-v2* (ME²-MPNET). Sentence-Transformers can capture semantically meaningful embedding, making them suitable architectures for sentence similarity tasks. Note that Sentence-Transformer models were run using 128 as maximum sequence length for efficiency purpose, since some models, like T5, have a huge number of

¹⁰ https://www.sbert.net/docs/sentence_transformer/pretrained_models.html

Table 9.4: F1-scores of ME²-T5, ME²-RoBERTa, and ME²-MPNET models fine-tuned on MFRC, MFTC, and eMFD datasets. Best results for each moral foundation in bold, second-best underlined.

	Care	Fairness	Loyalty	Authority	Purity	Non-moral	AVG
MFRC							
ME ² -BERT	<u>0.636</u>	<u>0.585</u>	<u>0.345</u>	<u>0.490</u>	<u>0.363</u>	0.669	<u>0.515</u>
ME ² -SBERT	0.629	0.583	0.359	0.483	0.352	0.660	0.511
ME ² -T5	0.661	0.604	<u>0.345</u>	0.503	0.384	0.715	0.535
ME ² -RoBERTa	0.630	0.575	0.289	0.472	0.346	<u>0.685</u>	0.500
ME ² -MPNET	0.596	0.558	0.270	0.427	0.295	0.563	0.452
MFTC							
ME ² -BERT	0.688	<u>0.673</u>	0.551	<u>0.521</u>	0.453	0.546	<u>0.572</u>
ME ² -SBERT	<u>0.692</u>	0.669	0.548	0.515	<u>0.451</u>	0.516	0.565
ME ² -T5	0.710	0.678	<u>0.549</u>	0.564	0.433	0.585	0.586
ME ² -RoBERTa	0.681	0.642	0.535	0.506	0.438	<u>0.573</u>	0.562
ME ² -MPNET	0.673	0.616	0.531	0.467	0.429	0.528	0.541
eMFD							
ME ² -BERT	0.309	0.225	0.217	0.227	0.220	0.373	0.262
ME ² -SBERT	0.299	0.238	0.218	0.242	0.243	0.288	0.255
ME ² -T5	0.292	0.272	0.247	<u>0.251</u>	0.243	0.212	0.253
ME ² -RoBERTa	0.324	0.265	0.254	<u>0.265</u>	<u>0.241</u>	0.423	0.295
ME ² -MPNET	<u>0.323</u>	<u>0.269</u>	0.254	0.271	0.237	<u>0.397</u>	<u>0.292</u>

parameters (355M). ME²-BERT was fine-tuned using 256 as maximum sequence length.

Overall, the best performing method is ME²-T5 with average F1-score of 0.458 across all datasets, followed by ME²-RoBERTa (0.450) and ME²-BERT (0.450). ME²-T5 shows outstanding performance on social datasets like MFRC and MFTC, but its performance decreases on eMFD. By contrast, ME²-RoBERTa and ME²-MPNET perform poorly on MFRC and MFTC but achieve the highest scores on eMFD. This highlights the strong differences between these datasets, as evidenced by their feature distributions in Fig. 9.4b. Other models, such as ME²-BERT and ME²-SBERT, exhibit more balanced performance across all datasets. Notably, ME²-BERT achieves comparable, and in some cases even better, performance than ME²-SBERT. We believe that the good performance achieved by ME²-BERT w.r.t. S-BERT are due to the task-specific fine-tuning process, as the relative performance of a model can also depend on the similarity of the pretraining and target tasks (Peters et al., 2019).

Visual interpretations. Fig. 9.4b shows the UMAP of the embeddings learned by ME²-BERT on E2MoCase, illustrating effective domain alignment despite the initial feature distribution differences. Considering the evaluation data, Fig. 9.10 shows their UMAP visualization using BERT (a) and ME²-BERT embeddings (b),

respectively. The out-of-the-box BERT embeddings are clearly separated clusters, reflecting distinct domain-specific features. By contrast, ME²-BERT aligns these datasets into a more overlapping space, suggesting it has learned to capture high-level, shared textual features rather than domain-specific ones. This indicates that ME²-BERT effectively abstracts moral concepts that generalize across different domains, improving its ability in moral value prediction.

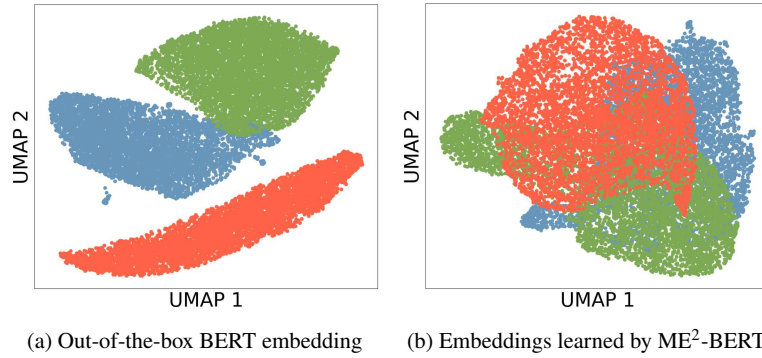


Fig. 9.10: UMAP of the dataset embeddings. Green, blue and orange points indicate samples from the MFRC, MFTC and eMFD datasets, respectively.

Qualitative analysis of the results. We analyze the quality of ME²-BERT’s responses by visually assessing instances from E2MoCase in which ME²-BERT correctly predicts all relevant moral foundations, and instances in which it makes wrong inference. 9.11 shows how ME²-BERT predicts moral dimensions of the training dataset. In 9.11a and 9.11b, it successfully predicts all relevant moral dimensions. The predicted values correspond to the nature of the events and the societal challenges they evoke. Specifically, in 9.11a the model captures the moral dimensions in the context of a murder case, while in 9.11b, it identifies Authority/Subversion and Purity/Degradation in a narrative about social protests.

In 9.11c, the model is unable to predict Loyalty/Betrayal, likely due to the sparsity of the event representation, which hinders to capture this moral dimension. In 9.11d, the instance focuses on interpersonal conflict and reconciliation between two individuals. In this case the model wrongly predicts Care/Harm.

9.6 Chapter Review

We presented ME²-BERT, a novel holistic framework leveraging emotions and events for moral value prediction. The promising performance of ME²-BERT supports our hypothesis of inter-dependence among emotions, events and moralities. Neverthe-

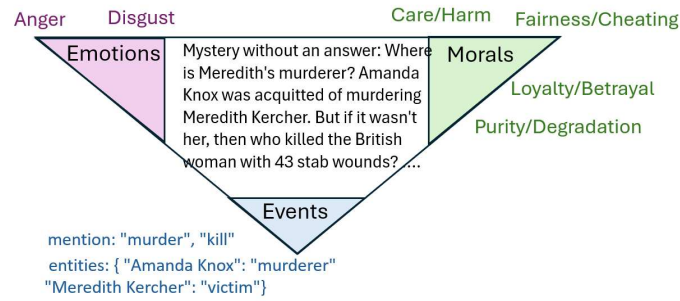
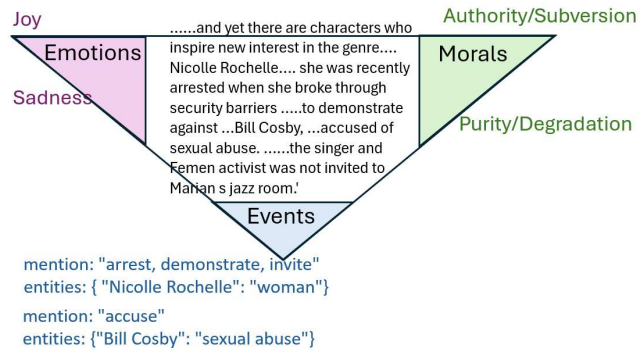
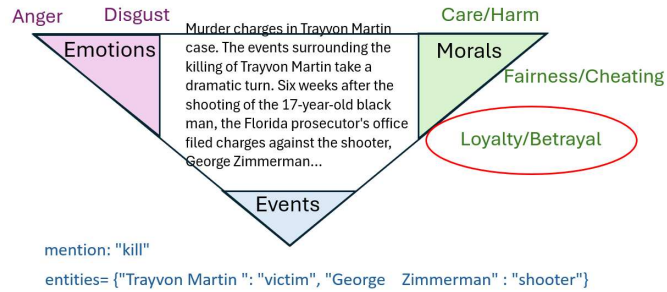
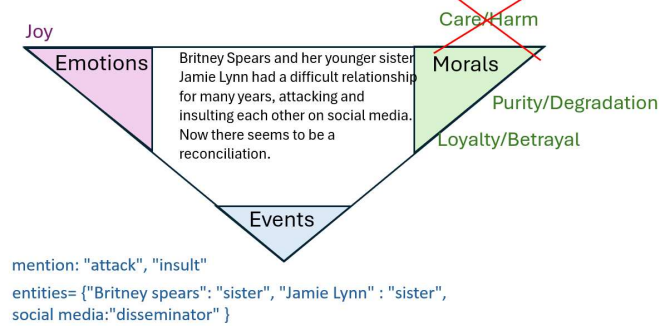
(a) ME²-BERT correctly predicts all moral foundations.(b) ME²-BERT correctly predicts all moral foundations.(c) ME²-BERT fails to predict Loyalty/Betrayal.(d) ME²-BERT wrongly predicts Loyalty/Betrayal.

Fig. 9.11: Instances of E2MoCase in which ME²-BERT successfully predicts moral foundations and instances where it does not. The red circle represents a false negative, while the cross represents a false positive.

less, further investigation is needed to deeply understand the multifaceted nature of the interactions arising among these fundamental aspects, for developing more effective and interpretable TLM-based framework for moral foundation prediction and related tasks.

ME²-BERT is designed to be in principle independent from the particular TLM employed in the encoding phase as well as from the dataset used to fine-tune the TLM. Therefore, other or more advanced BERT-like models could easily be used in alternative to BERT—indeed, we also experimented with Sentence-Transformer models. Less straightforward would be the replacement of E2MoCase since, to date, there are no alternatives to it that share the same features incorporating annotations at level of emotions, events and moral foundations at the same time. Nonetheless, some choices in ME²-BERT were influenced by the characteristics of E2MoCase. First, while our strategy for categorizing emotions has a strong theoretical foundation (Plutchik, 2001), the label’s selection process was constrained by the emotion values available in E2MoCase. For other datasets, different choices might prove more effective, especially given that the exact nature of interactions between morality and emotions in psychology remains unclear and warrants further discussion. Second, E2MoCase exhibits a domain shift between event-free and event-based paragraphs, making it well-suited for our domain adaptation strategy. However, this choice may not always be optimal for the model’s effectiveness and requires further investigation on other datasets. Though, as shown in our experimental section, our event-based domain identification strategy also proved effective for the MoralEvent dataset, where ME²-BERT achieved better performance than BERT-base. Overall, the integration of events and emotions depends on the characteristics of the fine-tuning dataset and is subject to the expertise of domain specialists.

We employed a straightforward strategy for selecting emotion categories for our contrastive learning strategy, where each text sample is linked to a single emotion—the one with the highest score in E2MoCase. However, since emotions are fluid and texts often convey a mix of emotions (Cambria et al., 2012), a more effective strategy would involve utilizing a wider range of emotions present in a text. We aim to investigate this approach in future works.

Finally, we cannot definitively answer whether emotions and events are what we need for moral value prediction. We believe this matter should first be addressed using tools from moral psychology and then integrated into the development of AI tools for moral tasks. By contrast, our study offers a complementary perspective by starting with AI tools to explore whether emotions, events, and moralities are interdependent and can be effectively leveraged for moral tasks. Another factor preventing a conclusive answer is the lack of reliable datasets annotated simultaneously with morality, emotions, and events. Nevertheless, we believe our work serves as a starting point that paves the way for future interdisciplinary research in this area.

Conclusions

The first part of the thesis explored the application of TLMs to the legal domain, addressing critical challenges inherent in processing legal texts. By examining a diverse range of methods, TLMs proved to be instrumental in tasks such as legal understanding and summarization.

We firstly presented a comprehensive framework for categorizing and analyzing TLM-based methodologies in the legal domain, with specific attention to task-adaptive and domain-adaptive methods. At its time of developing, the study was the first systematic one on problems, tasks and benchmarks in the legal AI area, for which we covered about thirty different TLMs used to build more than three hundred AI approaches and methods addressing retrieval, classification, prediction, entailment, summarization, generation, information extraction, and many other tasks relevant in the legal domain. Key findings unveiled the impact of TLMs in advancing legal AI, particularly in multilingual contexts, long document processing, and low-resource language tasks. Despite these advancements, several aspects can still be addressed in a more effective and efficient way: *(i)* most models require extensive fine-tuning to adapt to the formal and specialized syntax of legal texts, which may limit their generalizability across jurisdictions; *(ii)* there is a need for high-quality annotated datasets remains a bottleneck, especially for tasks involving multilingual or low-resource legal systems; *(iii)* challenges such as model explainability and fairness highlight potential biases in datasets and predictions, requiring further investigation to ensure ethical deployment in real-world applications.

Through the lens of European constitutional texts, we demonstrated the capabilities of TLMs in analyzing citizens' rights and duties. Notably, we proved that GPT-3 has a substantial knowledge of the European constitutions, accurately describing relevant topics without being provided with any portion extracted from the constitutions. Through the use of Sentence-Transformers, we also observed a significant similarity among constitutional texts addressing common topics, a crucial prerequisite for justifying a summarization phase. In this regard, we proposed a two-stage pipeline: first, prompting GPT-3 to generate topic-specific summaries for individual countries, then synthesizing unified cross-country summaries for each topic. These summaries, evaluated against multiple criteria, effectively capture the

essence of constitutions across countries while preserving the unique characteristics of each. Our research limitations include challenges in distinguishing highly similar topics, comparison with more recent and open models, cost constraints that restricted fine-tuning and prompt engineering.

The second part of the thesis presented an investigation of the abilities of TLMs in tackling the task of lexical entailment recognition. The key novelty of this study is to shed light on the actual skills of TLMs, with emphasis on Open models, to recognize entailment relations between verbs, gauging their accuracy against well-grounded, manually curated lexical resources such as WordNet and HyperLex. To accomplish our research goal, we defined three prompt types, providing different levels of contextual information, in both zero-shot and few-shot learning scenarios. We showed evidence of both abilities and limitations that arise in the examined models. Our results demonstrate that general-purpose TLMs can achieve moderate success in lexical entailment recognition, though none fully solve the task. We believe our study can advance our understanding of how TLMs grasp nuanced meanings and logical relationships among verbs, providing valuable insights into their interpretability and decision-making processes.

The third part of the thesis explored the application of TLMs in analyzing psychosocial dimensions, focusing on mental-health and moral values. By addressing these critical aspects of human behavior and societal interaction, the research contributes to understanding how TLMs can decode complex psychosocial phenomena.

Our investigation on mental-health-related tasks highlighted the strengths and limitations of TLM-based methods. These methods primarily detect depression, suicidal ideation, and other psychological states from textual data, predominantly sourced from social media. While domain adaptation and fine-tuning strategies enhance model performance, challenges such as explainability, cultural bias, and low-resource language support remain critical areas for future work. Additionally, the integration of multimodal data from diverse sources like clinical notes and multimedia platforms could further advance this domain.

In the context of moral dimensions, we provided a comprehensive analysis of TLMs concerning the Moral Foundation Theory (MFT). The study formalized MFT-related tasks and introduced a taxonomy categorizing TLM approaches into moral-driven and moral-targeted models. Various models and strategies, including fine-tuning and prompt engineering, have shown ability in accurately predicting moral values and aligning TLMs within specific moral scenarios. The study highlighted notable correlation between moral values and political stances, and it demonstrated how TLMs can be influenced by and reflect societal biases. However, the need for comprehensive and standardized annotation processes is critical to enhance the reliability of moral datasets. Additionally, constructing large-scale, multi-cultural, and multi-domain datasets is crucial for improving the generalizability of TLMs in MFT-related tasks.

Through the development of the E2MoCase dataset, this research bridged emotions, events, and moralities in media narratives about legal cases. E2MoCase provides a unique resource for exploring how emotions and events influence moral

judgment. Building on the E2MoCase dataset, we introduced ME²-BERT, a novel model designed to predict moral values by integrating emotions, events, and moralities. ME²-BERT demonstrated significant improvements in moral value prediction, particularly in out-of-domain scenarios. The promising performance of ME²-BERT supports our hypothesis of inter-dependence among emotions, events and moralities. Nevertheless, further investigation is needed to deeply understand the multifaceted nature of the interactions arising among these fundamental aspects.

Remarks on further research developments. Besides the aforementioned discussed aspects and limitations related to our proposed approaches, the deployment of TLMs in sensitive domains such as legal, mental-health, and morality has generally highlighted critical challenges that require careful consideration. Among these, *interpretability* remains paramount to fostering trust and enabling human oversight in decision-making processes. While significant strides have been made in improving *explainability*—ranging from post-hoc methods to augmented models enriched with domain-specific knowledge—further efforts are needed to balance interpretability with the complexity of TLMs. This is especially relevant in domains where decisions bear significant societal implications, such as law and ethics, underscoring the necessity of transparent reasoning pathways.

Equally important are the ethical dimensions surrounding the use of TLMs. Ensuring compliance with *privacy*, *data protection*, and *security standards* is fundamental, particularly given the challenges of filtering sensitive or toxic information without compromising the integrity of the underlying data. The rise of generative AI tools has amplified concerns over the misuse of generated content, demanding robust mechanisms for mitigating risks and adhering to evolving regulatory frameworks. Furthermore, the *fairness* and inclusivity of these models require ongoing scrutiny, particularly in addressing disparities linked to demographic, cultural, and regional factors.

Finally, the pervasive issue of *bias* within TLMs warrants urgent attention. Biases in training data—whether stemming from societal inequities, cultural oversights, or annotator subjectivity—pose significant risks of perpetuating stereotypes and unfair outcomes. This is particularly critical in tasks such as moral reasoning, where the diversity of cultural and ethical norms complicates the development of universally fair systems. Proactively identifying and mitigating such biases through interdisciplinary approaches, diverse datasets, and rigorous evaluation frameworks is essential for ensuring the equitable and ethical application of these models.

In conclusion, advancing TLMs in these sensitive domains demands a multifaceted approach that integrates interpretability, ethical safeguards, and bias mitigation. By addressing these challenges, future research can pave the way for responsible, transparent, and inclusive AI systems capable of supporting critical societal applications.

References

- Abdulhai M, Serapio-García G, Crepy C, Valter D, Canny J, Jaques N (2023) Moral foundations of large language models. CoRR abs/2310.15337, DOI 10.48550/ARXIV.2310.15337, URL <https://doi.org/10.48550/arXiv.2310.15337>
- Achilles L, Kisselew M, Schäfer J, Kölle R (2020) Using surface and semantic features for detecting early signs of self-harm in social media postings. In: Working Notes of CLEF, CEUR Workshop Proceedings, vol 2696
- Adams G, Fabbri AR, Ladhak F, Lehman E, Elhadad N (2023) From sparse to dense: GPT-4 summarization with chain of density prompting. CoRR abs/2309.04269
- Adilazuarda MF, Mukherjee S, Lavania P, Singh S, Dwivedi A, Aji AF, O’Neill J, Modi A, Choudhury M (2024) Towards measuring and modeling ”culture” in llms: A survey. CoRR abs/2403.15412, DOI 10.48550/ARXIV.2403.15412, URL <https://doi.org/10.48550/arXiv.2403.15412>
- Agirre E, Cer DM, Diab MT, Gonzalez-Agirre A (2012) Semeval-2012 task 6: A pilot on semantic textual similarity. In: Agirre E, Bos J, Diab MT (eds) Proceedings of the 6th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2012, Montréal, Canada, June 7-8, 2012, The Association for Computer Linguistics, pp 385–393, URL <https://aclanthology.org/S12-1051/>
- Agirre E, Banea C, Cer DM, Diab MT, Gonzalez-Agirre A, Mihalcea R, Rigau G, Wiebe J (2016) Semeval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In: Proc. SemEval@NAACL-HLT, pp 497–511
- Aguiar A, Silveira R, Pinheiro V, Furtado V, Neto JA (2021) Text classification in legal documents extracted from lawsuits in brazilian courts. In: Proc. of the Brazilian Conf. on Intelligent Systems (BRACIS), pp 586—600
- Ahmad WU, Chi J, Tian Y, Chang K (2020) PolicyQA: A Reading Comprehension Dataset for Privacy Policies. In: Findings of the Association for Computational Linguistics: EMNLP, Association for Computational Linguistics, Findings of ACL, pp 743–749

- Ahmad WU, Chi J, Le T, Norton T, Tian Y, Chang K (2021) Intent classification and slot filling for privacy policies. In: Proc. of the Annual Meeting of the Association for Computational Linguistics and the Int. Joint Conf. on Natural Language Processing (ACL/IJCNLP), Association for Computational Linguistics, pp 4402–4417
- Ahmed U, Lin JCW, Srivastava G (2022) Hyper-graph-based attention curriculum learning using a lexical algorithm for mental health. *Pattern Recognition Letters* 157:135–143
- Alberts H, Ipek A, Lucas R, Wozny P (2020) COLIEE 2020: Legal information retrieval and entailment with legal embeddings and boosting. In: *New Frontiers in Artificial Intelligence - JSAI-isAI 2020 Workshops, JURISIN, LENLS 2020 Workshops*, Springer, Lecture Notes in Computer Science, vol 12758, pp 211–225
- Aletras N, Tsarapatsanis D, Preotiuc-Pietro D, Lampos V (2016) Predicting judicial decisions of the European Court of Human Rights: a Natural Language Processing perspective. *PeerJ Comput Sci* 2:e93, DOI 10.7717/peerj-cs.93
- Alhuzali H, Zhang T, Ananiadou S (2021) Predicting Sign of Depression via Using Frozen Pre-trained Models and Random Forest Classifier. In: *Working Notes of CLEF, CEUR Workshop Proceedings*, vol 2936, pp 888–896
- Almazrouei E, Alobeidli H, Alshamsi A, Cappelli A, Cojocaru R, Debbah M, Goffinet É, Hesslow D, Launay J, Malartic Q, Mazzotta D, Noune B, Pannier B, Penedo G (2023) The falcon series of open language models. *CoRR abs/2311.16867*, DOI 10.48550/ARXIV.2311.16867, URL <https://doi.org/10.48550/arXiv.2311.16867>, 2311.16867
- Alshomary M, El Baff R, Gurcke T, Wachsmuth H (2022) The moral debater: A study on the computational generation of morally framed arguments. In: Muresan S, Nakov P, Villavicencio A (eds) *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, Dublin, Ireland, pp 8782–8797, DOI 10.18653/v1/2022.acl-long.601, URL <https://aclanthology.org/2022.acl-long.601>
- Althammer S, Askari A, Verberne S, Hanbury A (2021) DoSSIER@COLIEE 2021: Leveraging dense retrieval and summarization-based re-ranking for case law retrieval. *CoRR abs/2108.03937*
- Amos R, Acar G, Lucherini E, Kshirsagar M, Narayanan A, Mayer JR (2021) Privacy policies over time: Curation and analysis of a million-document dataset. In: *Proc. of the ACM Web Conf. (WWW)*, ACM, pp 2165–2176
- Anthropic A (2024) The claude 3 model family: Opus, sonnet, haiku. *Claude-3 Model Card*
- Antoun W, Baly F, Hajj HM (2020) AraBERT: Transformer-based Model for Arabic Language Understanding. *CoRR abs/2003.00104*
- Araque O, Gatti L, Kalimeri K (2020a) Moralstrength: Exploiting a moral lexicon and embedding similarity for moral foundations prediction. *Knowledge-based systems*

191:105184

- Araque O, Gatti L, Kalimeri K (2020b) Moralstrength: Exploiting a moral lexicon and embedding similarity for moral foundations prediction. *Knowl Based Syst* 191:105184, DOI 10.1016/J.KNOSYS.2019.105184, URL <https://doi.org/10.1016/j.knosys.2019.105184>
- Araque O, Gatti L, Kalimeri K (2022) Libertymfd: A lexicon to assess the moral foundation of liberty. In: *Proceedings of the 2022 ACM Conference on Information Technology for Social Good*, pp 154–160
- Artetxe M, Ruder S, Yogatama D (2020) On the cross-lingual transferability of monolingual representations. In: *Proc. of the Annual Meeting of the Association for Computational Linguistics (ACL)*, Association for Computational Linguistics, pp 4623–4637
- Artetxe M, Bhosale S, Goyal N, Mihaylov T, Ott M, Shleifer S, Lin XV, Du J, Iyer S, Pasunuru R, Anantharaman G, Li X, Chen S, Akin H, Baines M, Martin L, Zhou X, Koura PS, O'Horo B, Wang J, Zettlemoyer L, Diab MT, Kozareva Z, Stoyanov V (2022) Efficient large scale language modeling with mixtures of experts. In: *Proc. of the Conf. on Empirical Methods in Natural Language Processing (EMNLP)*, Association for Computational Linguistics, pp 11699–11732
- Askari A, Verberne S (2021) Combining Lexical and Neural Retrieval with Longformer-based Summarization for Effective Case Law Retrieval. In: *Proc. of the Int. Conf. on Design of Experimental Search & Information REtrieval Systems*, CEUR-WS.org, CEUR Workshop Proceedings, vol 2950, pp 162–170
- Atari M, Haidt J (2023) Ownership is (likely to be) a moral foundation. *Behavioral & Brain Sciences* 46
- Atari M, Graham J, Dehghani M (2020) Foundations of morality in iran. *Evolution and Human Behavior* 41(5):367–384
- Atari M, Haidt J, Graham J, Koleva S, Stevens ST, Dehghani M (2023) Morality beyond the weird: How the nomological network of morality varies across cultures. *Journal of Personality and Social Psychology*
- Athinaios K (2020) Named Entity Recognition using a Novel Linguistic Model for Greek Legal Corpora based on BERT model. BS Thesis, School of Science, Department of Informatics and Telecommunications.
- Aumiller D, Almasian S, Lackner S, Gertz M (2021) Structural text segmentation of legal documents. In: *Proc. of the Int. Conf. on Artificial Intelligence and Law (ICAIL)*, ACM, pp 2–11, DOI 10.1145/3462757.3466085
- Aumiller D, Chouhan A, Gertz M (2022) Eur-lex-sum: A multi- and cross-lingual dataset for long-form summarization in the legal domain. In: Goldberg Y, Kozareva Z, Zhang Y (eds) *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022*, Abu Dhabi, United Arab Emirates, December 7–11, 2022, Association for Computational Linguistics, pp 7626–7639,

- DOI 10.18653/V1/2022.EMNLP-MAIN.519, URL <https://doi.org/10.18653/v1/2022.emnlp-main.519>
- Avram A, Pais VF, Tufis DI (2021) PyEuroVoc: A Tool for Multilingual Legal Document Classification with EuroVoc Descriptors. In: Proc. of the Int. Conf. on Recent Advances in Natural Language Processing (RANLP), pp 92–101
- Awad E, Dsouza S, Kim R, Schulz J, Henrich J, Shariff A, Bonnefon JF, Rahwan I (2018) The moral machine experiment. *Nature* 563(7729):59–64
- Aydemir A, de Castro Souza P, Gelfman A (2020) Using BERT and TF-IDF to predict entailment in law-based queries. In: New Frontiers in Artificial Intelligence - JSAI-isAI 2020 Workshops, JURISIN, LENLS 2020 Workshops, Springer, Lecture Notes in Computer Science, vol 12758, pp 286–293
- Bahdanau D, Cho K, Bengio Y (2015) Neural machine translation by jointly learning to align and translate. In: Bengio Y, LeCun Y (eds) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings
- Bai H, Wang T, Sordoni A, Shi P (2022) Better language model with hypernym class prediction. In: Muresan S, Nakov P, Villavicencio A (eds) Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022, Association for Computational Linguistics, pp 1352–1362, DOI 10.18653/V1/2022.ACL-LONG.96, URL <https://doi.org/10.18653/v1/2022.acl-long.96>
- Bao H, Dong L, Wei F, Wang W, Yang N, Liu X, Wang Y, Gao J, Piao S, Zhou M, Hon H (2020) UniLMv2: Pseudo-Masked Language Models for Unified Language Model Pre-Training. In: Proc. of the Int. Conf. on Machine Learning (ICML), pp 642–652
- Barrios F, López F, Argerich L, Wachenchauzer R (2016) Variations of the Similarity Function of TextRank for Automated Summarization. CoRR abs/1602.03606, 1602.03606
- Barros L, Trifan A, Oliveira JL (2021) VADER meets BERT: sentiment analysis for early detection of signs of self-harm through social mining. In: Working Notes of CLEF, CEUR Workshop Proceedings, vol 2936, pp 897–907
- Basile A, Chinea-Rios M, Uban AS, Müller T, Rössler L, Yenikent S, Chulvi B, Rosso P, Franco-Salvador M (2021) UPV-Symanto at eRisk 2021: Mental Health Author Profiling for Early Risk Prediction on the Internet. In: Working Notes of CLEF, CEUR Workshop Proceedings, vol 2936, pp 908–927
- Bayrak T (2022) A comparative analysis of the world’s constitutions: a text mining approach. *Soc Netw Anal Min* 12(1):26
- Beiró MG, D’Ignazi J, Bustos VP, Prado MF, Kalimeri K (2023) Moral narratives around the vaccination debate on facebook. In: Ding Y, Tang J, Sequeda JF, Aroyo L, Castillo C, Houben G (eds) Proceedings of the ACM Web Conference 2023,

- WWW 2023, Austin, TX, USA, 30 April 2023 - 4 May 2023, ACM, pp 4134–4141, DOI 10.1145/3543507.3583865, URL <https://doi.org/10.1145/3543507.3583865>
- Beltagy I, Peters ME, Cohan A (2020) Longformer: The long-document transformer. CoRR abs/2004.05150, URL <https://arxiv.org/abs/2004.05150>, 2004.05150
- Ben-David S, Blitzer J, Crammer K, Kulesza A, Pereira F, Vaughan JW (2010) A theory of learning from different domains. *Machine learning* 79:151–175
- Bhattacharya P, Hiware K, Rajgaria S, Pochhi N, Ghosh K, Ghosh S (2019a) A comparative study of summarization algorithms applied to legal case judgments. In: Proc. of the European Conf. on IR Research (ECIR), Springer, Lecture Notes in Computer Science, vol 11437, pp 413–428
- Bhattacharya P, Paul S, Ghosh K, Ghosh S, Wyner A (2019b) Identification of Rhetorical Roles of Sentences in Indian Legal Judgments. In: Proc. of the Int. Conf. on Legal Knowledge and Information Systems (JURIX)
- Bhattacharya P, Ghosh K, Pal A, Ghosh S (2020) Hier-spcnet: A legal statute hierarchy-based heterogeneous network for computing legal case document similarity. In: Proc. of the ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR), ACM, pp 1657–1660
- Bhattacharya P, Paul S, Ghosh K, Ghosh S, Wyner A (2021) DeepRhole: deep learning for rhetorical role labeling of sentences in legal case documents. *Artif Intell Law* DOI 10.1007/s10506-021-09304-5
- Bibal A, Lognoul M, de Streel A, Frénay B (2021) Legal requirements on explainability in machine learning. *Artif Intell Law* 29(2):149–169
- Biderman S, Schoelkopf H, Anthony QG, Bradley H, O’Brien K, Hallahan E, Khan MA, Purohit S, Prashanth US, Raff E, Skowron A, Sutawika L, van der Wal O (2023) Pythia: A suite for analyzing large language models across training and scaling. In: Krause A, Brunskill E, Cho K, Engelhardt B, Sabato S, Scarlett J (eds) International Conference on Machine Learning, ICML 2023, 23–29 July 2023, Honolulu, Hawaii, USA, PMLR, Proceedings of Machine Learning Research, vol 202, pp 2397–2430
- Black S, Biderman S, Hallahan E, Anthony Q, Gao L, Golding L, He H, Leahy C, McDonnell K, Phang J, Pieler M, Prashanth US, Purohit S, Reynolds L, Tow J, Wang B, Weinbach S (2022) Gpt-neox-20b: An open-source autoregressive language model. CoRR abs/2204.06745
- Blair-Stanek A, Holzenberger N, Durme BV (2023) Can GPT-3 perform statutory reasoning? CoRR abs/2302.06100
- Bojanowski P, Grave E, Joulin A, Mikolov T (2017) Enriching word vectors with subword information. *Trans Assoc Comput Linguistics* 5:135–146
- Bowman SR, Angeli G, Potts C, Manning CD (2015) A large annotated corpus for learning natural language inference. In: Màrquez L, Callison-Burch C, Su J, Pighin

- D, Marton Y (eds) Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015, The Association for Computational Linguistics, pp 632–642, DOI 10.18653/V1/D15-1075, URL <https://doi.org/10.18653/v1/d15-1075>
- Brack A, Hoppe A, Buschermöhle P, Ewerth R (2021) Sequential sentence classification in research papers using cross-domain multi-task learning. CoRR abs/2102.06008
- Brady WJ, Crockett MJ, Van Bavel JJ (2020) The mad model of moral contagion: The role of motivation, attention, and design in the spread of moralized content online. *Perspectives on Psychological Science* 15(4):978–1010
- Branting LK, Pfeifer C, Brown B, Ferro L, Aberdeen JS, Weiss B, Pfaff M, Liao B (2021) Scalable and explainable legal prediction. *Artif Intell Law* 29(2):213–238
- Breunig MM, Kriegel H, Ng RT, Sander J (2000) LOF: identifying density-based local outliers. In: Chen W, Naughton JF, Bernstein PA (eds) Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, May 16-18, 2000, Dallas, Texas, USA, ACM, pp 93–104, DOI 10.1145/342009.335388, URL <https://doi.org/10.1145/342009.335388>
- Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, Agarwal S, Herbert-Voss A, Krueger G, Henighan T, Child R, Ramesh A, Ziegler DM, Wu J, Winter C, Hesse C, Chen M, Sigler E, Litwin M, Gray S, Chess B, Clark J, Berner C, McCandlish S, Radford A, Sutskever I, Amodei D (2020) Language models are few-shot learners. In: Larochelle H, Ranzato M, Hadsell R, Balcan M, Lin H (eds) Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual, URL <https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfc4967418bfb8ac142f64a-Abstract.html>
- Bucur A, Cosma A, Dinu LP (2021) Early Risk Detection of Pathological Gambling, Self-Harm and Depression Using BERT. In: Working Notes of CLEF, CEUR Workshop Proceedings, vol 2936, pp 938–949
- Bui D, Shin KG, Choi J, Shin J (2021) Automated extraction and presentation of data practices in privacy policies. *Proceedings on Privacy Enhancing Technologies* 2021(2):88–110
- Bulla L, Giorgis SD, Gangemi A, Marinucci L, Mongiovì M (2022) Detection of morality in tweets based on the moral foundation theory. In: Nicosia G, Ojha V, Malfa EL, Malfa GL, Pardalos PM, Fatta GD, Giuffrida G, Umeton R (eds) Machine Learning, Optimization, and Data Science - 8th International Workshop, LOD 2022, Certosa di Pontignano, Italy, September 19-22, 2022, Revised Selected Papers, Part I, Springer, Lecture Notes in Computer Science, vol 13810, pp 1–13, DOI 10.1007/978-3-031-25599-1_1, URL https://doi.org/10.1007/978-3-031-25599-1_1
- Burton K (2017) “Think Like a Lawyer” Using a Legal Reasoning Grid and Criterion-Referenced Assessment Rubric on IRAC (Issue, Rule, Application, Conclusion).

- Journal of Learning Design 10(2):57–68
- Callister PD (2020) Law, artificial intelligence, and natural language processing: A funny thing happened on the way to my search results. *Law Libr J* 112(161)
- Cambria E, Livingstone A, Hussain A (2012) The hourglass of emotions. In: *Cognitive behavioural systems: COST 2102 international training school, dresden, Germany, February 21-26, 2011, revised selected papers*, Springer, pp 144–157
- Cameron CD, Lindquist KA, Gray K (2015) A constructionist review of morality and emotions: No evidence for specific links between moral content and discrete emotions. *Personality and social psychology review* 19(4):371–394
- Carrasco-Farre C (2024) Large language models are as persuasive as humans, but why? about the cognitive effort and moral-emotional language of llm arguments. *arXiv preprint arXiv:240409329*
- Cer DM, Diab MT, Agirre E, Lopez-Gazpio I, Specia L (2017) Semeval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In: Bethard S, Carpuat M, Apidianaki M, Mohammad SM, Cer DM, Jurgens D (eds) *Proceedings of the 11th International Workshop on Semantic Evaluation, SemEval@ACL 2017, Vancouver, Canada, August 3-4, 2017*, Association for Computational Linguistics, pp 1–14, DOI 10.18653/V1/S17-2001, URL <https://doi.org/10.18653/v1/S17-2001>
- Chalkidis I, Nikolaou C, Soursos P, Koubarakis M (2017) Modeling and Querying Greek Legislation Using Semantic Web Technologies. In: *Proc of the Int. Conf. on the Semantic Web (ESWC), Lecture Notes in Computer Science*, vol 10249, pp 591–606
- Chalkidis I, Androutsopoulos I, Aletras N (2019a) Neural legal judgment prediction in english. In: Korhonen A, Traum DR, Màrquez L (eds) *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, Association for Computational Linguistics, pp 4317–4323, DOI 10.18653/V1/P19-1424, URL <https://doi.org/10.18653/v1/p19-1424>
- Chalkidis I, Fergadiotis M, Malakasiotis P, Androutsopoulos I (2019b) Large-scale multi-label text classification on EU legislation. In: Korhonen A, Traum DR, Màrquez L (eds) *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, Association for Computational Linguistics, pp 6314–6322, DOI 10.18653/V1/P19-1636, URL <https://doi.org/10.18653/v1/p19-1636>
- Chalkidis I, Fergadiotis M, Kotitsas S, Malakasiotis P, Aletras N, Androutsopoulos I (2020a) An empirical study on large-scale multi-label text classification including few and zero-shot labels. In: *Proc. of the Conf. on Empirical Methods in Natural Language Processing (EMNLP)*, Association for Computational Linguistics, pp 7503–7515, DOI 10.18653/v1/2020.emnlp-main.607
- Chalkidis I, Fergadiotis M, Malakasiotis P, Aletras N, Androutsopoulos I (2020b)

- LEGAL-BERT: The muppets straight out of law school. In: Findings of the Association for Computational Linguistics: EMNLP 2020, Association for Computational Linguistics, pp 2898–2904, DOI 10.18653/v1/2020.findings-emnlp.261
- Chalkidis I, Fergadiotis M, Androutsopoulos I (2021a) MultiEURLEX - A multilingual and multi-label legal document classification dataset for zero-shot cross-lingual transfer. In: Proc. of the Conf. on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, pp 6974–6996, DOI 10.18653/v1/2021.emnlp-main.559
- Chalkidis I, Fergadiotis M, Manginas N, Katakalo E, Malakasiotis P (2021b) Regulatory Compliance through Doc2Doc Information Retrieval: A case study in EU/UK legislation where text similarity has limitations. In: Proc. of the Conf. of the European Chapter of the Association for Computational Linguistics (EACL), Association for Computational Linguistics, pp 3498–3511, DOI 10.18653/v1/2021.eacl-main.305
- Chalkidis I, Fergadiotis M, Tsarapatsanis D, Aletras N, Androutsopoulos I, Malakasiotis P (2021c) Paragraph-level rationale extraction through regularization: A case study on european court of human rights cases. In: Toutanova K, Rumshisky A, Zettlemoyer L, Hakkani-Tür D, Beltagy I, Bethard S, Cotterell R, Chakraborty T, Zhou Y (eds) Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021, Association for Computational Linguistics, pp 226–241, DOI 10.18653/v1/2021.NAACL-MAIN.22, URL <https://doi.org/10.18653/v1/2021.naacl-main.22>
- Chalkidis I, Dai X, Fergadiotis M, Malakasiotis P, Elliott D (2022a) An Exploration of Hierarchical Attention Transformers for Efficient Long Document Classification. CoRR abs/2210.05529
- Chalkidis I, Jana A, Hartung D, II MJB, Androutsopoulos I, Katz DM, Aletras N (2022b) Lexglue: A benchmark dataset for legal language understanding in english. In: Muresan S, Nakov P, Villavicencio A (eds) Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022, Association for Computational Linguistics, pp 4310–4330, DOI 10.18653/v1/2022.ACL-LONG.297, URL <https://doi.org/10.18653/v1/2022.acl-long.297>
- Chalkidis I, Pasini T, Zhang S, Tomada L, Schwemer SF, Søgaaard A (2022c) Fairlex: A multilingual benchmark for evaluating fairness in legal text processing. In: Proc. of the Annual Meeting of the Association for Computational Linguistics (ACL), Association for Computational Linguistics, pp 4389–4406
- Chan B, Schweter S, Möller T (2020) German’s next language model. In: Proc. of the Int. Conf. on Computational Linguistics (COLING), International Committee on Computational Linguistics, pp 6788–6796, DOI 10.18653/v1/2020.coling-main.598
- Chang Y, Wang X, Wang J, Wu Y, Yang L, Zhu K, Chen H, Yi X, Wang C, Wang Y, Ye W, Zhang Y, Chang Y, Yu PS, Yang Q, Xie X (2024) A survey on evaluation of large language models. ACM Trans Intell Syst Technol 15(3):39:1–39:45, DOI

10.1145/3641289, URL <https://doi.org/10.1145/3641289>

- Chaurasia A, Prajapati SV, Tiru PA, Kumar S, Gupta R, Chauhan A (2021) Predicting mental health of scholars using contextual word embedding. In: Proc. INDIACom, pp 923–930
- Chen E, Lerman K, Ferrara E, et al. (2020) Tracking social media discourse about the covid-19 pandemic: Development of a public coronavirus twitter data set. *JMIR public health and surveillance* 6(2):e19273
- Chen J, Shi W, Fu Z, Cheng S, Li L, Xiao Y (2023) Say what you mean! large language models speak too positively about negative commonsense knowledge. In: Rogers A, Boyd-Graber JL, Okazaki N (eds) Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9–14, 2023, Association for Computational Linguistics, pp 9890–9908, URL <https://doi.org/10.18653/v1/2023.acl-long.550>
- Chen M, Tworek J, Jun H, Yuan Q, de Oliveira Pinto HP, Kaplan J, Edwards H, Burda Y, Joseph N, Brockman G, Ray A, Puri R, Krueger G, Petrov M, Khlaaf H, Sastry G, Mishkin P, Chan B, Gray S, Ryder N, Pavlov M, Power A, Kaiser L, Bavarian M, Winter C, Tillet P, Such FP, Cummings D, Plappert M, Chantzis F, Barnes E, Herbert-Voss A, Guss WH, Nichol A, Paino A, Tezak N, Tang J, Babuschkin I, Balaji S, Jain S, Saunders W, Hesse C, Carr AN, Leike J, Achiam J, Misra V, Morikawa E, Radford A, Knight M, Brundage M, Murati M, Mayer K, Welinder P, McGrew B, Amodei D, McCandlish S, Sutskever I, Zaremba W (2021) Evaluating large language models trained on code. *CoRR* abs/2107.03374
- Chen T, He T, Benesty M, Khotilovich V, Tang Y, Cho H, Chen K, et al. (2015) Xgboost: extreme gradient boosting. *R package version 04-2* 1(4):1–4
- Cherichi S, Faiz R (2016) Using big data values to enhance social event detection pattern. In: Proceedings of the ACS/IEEE International Conference on Computer Systems and Applications
- Chi J, Ahmad WU, Tian Y, Chang K (2023) PLUE: Language Understanding Evaluation Benchmark for Privacy Policies in English. In: Proc. of the Annual Meeting of the Association for Computational Linguistics (ACL), Association for Computational Linguistics, pp 352–365
- Choi JH, Hickman KE, Monahan A, Schwarcz D (2023) ChatGPT goes to law school. *Journal of Legal Education*
- Chung HW, Hou L, Longpre S, Zoph B, Tay Y, Fedus W, Li Y, Wang X, Dehghani M, Brahma S, et al. (2024) Scaling instruction-finetuned language models. *Journal of Machine Learning Research* 25(70):1–53
- Clark K, Luong M, Le QV, Manning CD (2020) ELECTRA: pre-training text encoders as discriminators rather than generators. In: 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26–30, 2020, OpenReview.net, URL <https://openreview.net/forum?id=r1xMH1BtvB>

- Clifford S, Iyengar V, Cabeza R, Sinnott-Armstrong W (2015) Moral foundations vignettes: A standardized stimulus database of scenarios based on moral foundations theory. *Behavior research methods* 47(4):1178–1198
- Clinchant S, Csúrká G, Chidlovskii B (2016) A domain adaptation regularization for denoising autoencoders. In: Erk K, Smith NA (eds) *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Association for Computational Linguistics, Berlin, Germany, pp 26–31, DOI 10.18653/v1/P16-2005, URL <https://aclanthology.org/P16-2005>
- Cobbe K, Kosaraju V, Bavarian M, Chen M, Jun H, Kaiser L, Plappert M, Tworek J, Hilton J, Nakano R, Hesse C, Schulman J (2021) Training verifiers to solve math word problems. CoRR abs/2110.14168, URL <https://arxiv.org/abs/2110.14168>, 2110.14168
- Cohan A, Derroncourt F, Kim DS, Bui T, Kim S, Chang W, Goharian N (2018) A discourse-aware attention model for abstractive summarization of long documents. In: *Proc. NAACL-HTL*, pp 615–621
- Cohan A, Beltagy I, King D, Dalvi B, Weld DS (2019) Pretrained language models for sequential sentence classification. In: *Proc. of the Conf. on Empirical Methods in Natural Language Processing and the Int. Joint Conf. on Natural Language Processing (EMNLP-IJCNLP)*, Association for Computational Linguistics, pp 3691–3697
- Conneau A, Lample G (2019) Cross-lingual language model pretraining. In: Wallach HM, Larochelle H, Beygelzimer A, d'Alché-Buc F, Fox EB, Garnett R (eds) *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp 7057–7067, URL <https://proceedings.neurips.cc/paper/2019/hash/c04c19c2c2474dbf5f7ac4372c5b9af1-Abstract.html>
- Conneau A, Rinott R, Lample G, Williams A, Bowman SR, Schwenk H, Stoyanov V (2018) XNLI: evaluating cross-lingual sentence representations. In: *Proc. of the Conf. on Empirical Methods in Natural Language Processing (EMNLP)*, Association for Computational Linguistics, pp 2475–2485
- Conneau A, Khandelwal K, Goyal N, Chaudhary V, Wenzek G, Guzmán F, Grave E, Ott M, Zettlemoyer L, Stoyanov V (2020) Unsupervised cross-lingual representation learning at scale. In: Jurafsky D, Chai J, Schluter N, Tetreault JR (eds) *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, Association for Computational Linguistics, pp 8440–8451, DOI 10.18653/v1/2020.ACL-MAIN.747, URL <https://doi.org/10.18653/v1/2020.acl-main.747>
- Cui Y, Che W, Liu T, Qin B, Wang S, Hu G (2020) Revisiting Pre-Trained Models for Chinese Natural Language Processing. In: *Findings of the Association for Computational Linguistics: EMNLP*, Association for Computational Linguistics, pp 657–668, DOI 10.18653/v1/2020.findings-emnlp.58
- Cui Y, Yang Z, Liu T (2022) PERT: pre-training BERT with permuted language model.

CoRR abs/2203.06906

- Dadgostari F, Guim M, Beling P, Livermore MA, Rockmore D (2021) Modeling law search as prediction. *Artif Intell Law* 29(1):3–34
- Dadu T, Pant K (2020) Team Rouges at SemEval-2020 Task 12: Cross-lingual Inductive Transfer to Detect Offensive Language. In: *Proc. of the Fourteenth Workshop on Semantic Evaluation, SemEval@COLING 2020*, Int. Committee for Computational Linguistics, pp 2183–2189
- Dai Z, Yang Z, Yang Y, Carbonell JG, Le QV, Salakhutdinov R (2019) Transformer-XL: Attentive Language Models beyond a Fixed-Length Context. In: *Proc. of the Annual Meeting of the Association for Computational Linguistics (ACL)*, Association for Computational Linguistics, pp 2978–2988, DOI 10.18653/v1/p19-1285
- Danescu-Niculescu-Mizil C, Lee L (2011) Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. In: Keller F, Reitter D (eds) *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics, CMCL@ACL 2011*, Portland, Oregon, USA, June 23, 2011, Association for Computational Linguistics, pp 76–87, URL <https://aclanthology.org/W11-0609/>
- Davidson T, Warmesley D, Macy MW, Weber I (2017) Automated hate speech detection and the problem of offensive language. In: *Proceedings of the Eleventh International Conference on Web and Social Media, ICWSM 2017*, Montréal, Québec, Canada, May 15–18, 2017, AAAI Press, pp 512–515, URL <https://aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/view/15665>
- Delobelle P, Winters T, Berendt B (2020) RobBERT: a Dutch RoBERTa-based Language Model. In: *Findings of the Association for Computational Linguistics: EMNLP*, Association for Computational Linguistics, pp 3255–3265, DOI 10.18653/v1/2020.findings-emnlp.292
- Demszky D, Movshovitz-Attias D, Ko J, Cowen AS, Nemade G, Ravi S (2020) Goemotions: A dataset of fine-grained emotions. In: Jurafsky D, Chai J, Schluter N, Tetreault JR (eds) *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020*, Online, July 5–10, 2020, Association for Computational Linguistics, pp 4040–4054, DOI 10.18653/V1/2020.ACL-MAIN.372, URL <https://doi.org/10.18653/v1/2020.acl-main.372>
- Deroy A, Bhattacharya P, Ghosh K, Ghosh S (2021) An analytical study of algorithmic and expert summaries of legal cases. In: *Proc. of the Conf. on Legal Knowledge and Information Systems (JURIX)*, IOS Press, *Frontiers in Artificial Intelligence and Applications*, vol 346, pp 90–99
- Devlin J, Chang M, Lee K, Toutanova K (2019) BERT: pre-training of deep bidirectional transformers for language understanding. In: Burstein J, Doran C, Solorio T (eds) *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019*, Minneapolis, MN, USA, June 2–7, 2019, Volume 1 (Long and Short Papers), Asso-

- ciation for Computational Linguistics, pp 4171–4186, DOI 10.18653/V1/N19-1423, URL <https://doi.org/10.18653/v1/n19-1423>
- Doddapaneni S, Ramesh G, Kunchukuttan A, Kumar P, Khapra MM (2021) A primer on pretrained multilingual language models. CoRR abs/2107.00676
- Dognin PL, Rios J, Luss R, Padhi I, Riemer MD, Liu M, Sattigeri P, Nagireddy M, Varshney KR, Bouneffouf D (2024) Contextual moral value alignment through context-based aggregation. CoRR abs/2403.12805, DOI 10.48550/ARXIV.2403.12805, URL <https://doi.org/10.48550/arXiv.2403.12805>
- Dong L, Yang N, Wang W, Wei F, Liu X, Wang Y, Gao J, Zhou M, Hon H (2019) Unified language model pre-training for natural language understanding and generation. In: Proc. of the Annual Conf. on Neural Information Processing Systems (NeurIPS), pp 13042–13054
- Dong Q, Niu S (2021) Legal judgment prediction via relational learning. In: Proc. of the ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR), ACM, pp 983–992
- Douka S, Abdine H, Vazirgiannis M, Hamdani RE, Amariles DR (2021) JuriBERT: A Masked-Language Model Adaptation for French Legal Text. In: Proc. of the 2021 Workshop on Natural Legal Language Processing (NLLP), ACL, pp 95–101
- Drawzeski K, Galassi A, Jablonowska A, Lagioia F, Lippi M, Micklitz HW, Sartor G, Tagiuri G, Torroni P (2021) A corpus for multilingual analysis of online terms of service. In: Proc. of the 2021 Workshop on Natural Legal Language Processing (NLLP), Association for Computational Linguistics, pp 1–8
- Du Z, Qian Y, Liu X, Ding M, Qiu J, Yang Z, Tang J (2022) GLM: General Language Model Pretraining with Autoregressive Blank Infilling. In: Proc. ACL
- Dubey A, Jauhri A, Pandey A, Kadian A, Al-Dahle A, Letman A, Mathur A, Schelten A, Yang A, Fan A, et al. (2024) The llama 3 herd of models. arXiv preprint arXiv:2407.21783
- Dumitrescu SD, Avram A, Pyysalo S (2020) The birth of Romanian BERT. In: Findings of the Association for Computational Linguistics: EMNLP, Association for Computational Linguistics, pp 4324–4328, DOI 10.18653/v1/2020.findings-emnlp.387
- Dunn M, Sagun L, Higgins M, Güney VU, Cirik V, Cho K (2017) SearchQA: A New Q&A Dataset Augmented with Context from a Search Engine. CoRR abs/1704.05179
- Eddine MK, Tixier AJ, Vazirgiannis M (2021) BARThez: a Skilled Pretrained French Sequence-to-Sequence Model. In: Proc. of the Conf. on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, pp 9369–9390
- Elkins Z, Ginsburg T, Melton J, Shaffer R, Sequeda JF, Miranker DP (2014) Constitute: The world’s constitutions to read, search, and compare. J Web Semant 27-28:10–18
- Ellemers N (2018) Morality and Social Identity. In: van Zomeren M, Dovidio JF (eds)

- The Oxford Handbook of the Human Essence, Oxford University Press, p 0, DOI 10.1093/oxfordhb/9780190247577.013.5, URL <https://doi.org/10.1093/oxfordhb/9780190247577.013.5>
- Elwany E, Moore D, Oberoi G (2019) BERT goes to law school: Quantifying the competitive advantage of access to large legal corpora in contract understanding. CoRR abs/1911.00473, 1911.00473
- Farzindar A, Lapalme G (2004) Letsum, an automatic legal text summarizing system. *Jurix* pp 11–18
- Feijó D, Moreira V (2018) RulingBR: A Summarization Dataset for Legal Texts. In: Proc. of the Conf. on Computational Processing of the Portuguese Language (PROPOR), Springer, Lecture Notes in Computer Science, vol 11122, pp 255–264
- Feijó D, Moreira V (2019) Summarizing legal rulings: Comparative experiments. In: Proc. of the Int. Conf. on Recent Advances in Natural Language Processing (RANLP), pp 313–322
- Fellbaum C (1990) English verbs as a semantic net. *International journal of lexicography* 3(4):278–301, URL <https://doi.org/10.1093/ijl/3.4.278>
- Feng Y, Li C, Ng V (2022) Legal judgment prediction via event extraction with constraints. In: Proc. of the Annual Meeting of the Association for Computational Linguistics (ACL), Association for Computational Linguistics, pp 648–664
- Fisch A, Talmor A, Jia R, Seo M, Choi E, Chen D (2019) MRQA 2019 shared task: Evaluating generalization in reading comprehension. In: Proc. of the Workshop on Machine Reading for Question Answering (MRQA@EMNLP), Association for Computational Linguistics, pp 1–13, DOI 10.18653/v1/D19-5801
- Forbes M, Hwang JD, Shwartz V, Sap M, Choi Y (2020) Social chemistry 101: Learning to reason about social and moral norms. In: Webber B, Cohn T, He Y, Liu Y (eds) Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16–20, 2020, Association for Computational Linguistics, pp 653–670, DOI 10.18653/V1/2020.EMNLP-MAIN.48, URL <https://doi.org/10.18653/v1/2020.emnlp-main.48>
- Francesconi E (2022) The winter, the summer and the summer dream of artificial intelligence in law. *Artif Intell Law* 30(2):147–161
- Franco M, Gaggi O, Palazzi CE (2023) Analyzing the use of large language models for content moderation with chatgpt examples. In: Guidi B, Michienzi A, Ricci L (eds) Proceedings of the 3rd International Workshop on Open Challenges in Online Social Networks, OASIS 2023, Rome, Italy, September 4–8, 2023, ACM, pp 1–8, DOI 10.1145/3599696.3612895, URL <https://doi.org/10.1145/3599696.3612895>
- Fraser KC, Kiritchenko S, Balkir E (2022) Does moral code have a moral code? probing delphi’s moral philosophy. In: Verma A, Pruksachatkun Y, Chang KW, Galstyan A, Dhamala J, Cao YT (eds) Proceedings of the 2nd Workshop on Trustworthy Natural Language Processing (TrustNLP 2022), Association for Computational Linguistics,

- Seattle, U.S.A., pp 26–42, DOI 10.18653/v1/2022.trustnlp-1.3, URL <https://aclanthology.org/2022.trustnlp-1.3>
- Frimer J (2019) Moral foundations dictionary 2.0. URL <https://osf.io/ezn37/>
- Fu X, Zhang J, Meng Z, King I (2020) MAGNN: metapath aggregated graph neural network for heterogeneous graph embedding. In: Proc. of the ACM Web Conf. (WWW), ACM, pp 2331–2341
- Furniturewala S, Jain R, Kumari V, Sharma Y (2021) Legal Text Classification and Summarization using Transformers and Joint Text Features. In: Working Notes of FIRE 2021 - Forum for Information Retrieval Evaluation, CEUR-WS.org, CEUR Workshop Proceedings, vol 3159, pp 541–546
- Gain B, Bandyopadhyay D, Saikh T, Ekbal A (2019) IITP@COLIEE 2019: Legal Information Retrieval Using BM25 and BERT. In: Proc. of COLIEE 2019 workshop: Competition on Legal Information Extraction/Entailment
- Gallegos IO, Rossi RA, Barrow J, Tanjim MM, Kim S, DERNONCOURT F, Yu T, Zhang R, Ahmed NK (2023) Bias and fairness in large language models: A survey. CoRR abs/2309.00770, DOI 10.48550/ARXIV.2309.00770, URL <https://doi.org/10.48550/arXiv.2309.00770>
- Gan L, Kuang K, Yang Y, Wu F (2021) Judgment prediction via injecting legal knowledge into neural networks. In: Proc. of the AAAI Conference on Artificial Intelligence (AAAI), the Conf. on Innovative Applications of Artificial Intelligence (IAAI), the AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI), AAAI Press, pp 12866–12874
- Ganin Y, Ustinova E, Ajakan H, Germain P, Larochelle H, Laviolette F, March M, Lempitsky V (2016) Domain-adversarial training of neural networks. Journal of machine learning research 17(59):1–35
- Gao J, Ning H, Han Z, Kong L, Qi H (2020) Legal text classification model based on text statistical features and deep semantic features. In: Working Notes of FIRE 2020 - Forum for Information Retrieval Evaluation, CEUR-WS.org, CEUR Workshop Proceedings, vol 2826, pp 35–41
- Gao L, Biderman S, Black S, Golding L, Hoppe T, Foster C, Phang J, He H, Thite A, Nabeshima N, Presser S, Leahy C (2021) The pile: An 800gb dataset of diverse text for language modeling. CoRR abs/2101.00027
- García-Ferrero I, Altuna B, Álvarez J, Gonzalez-Dios I, Rigau G (2023) This is not a dataset: A large negation benchmark to challenge large language models. In: Bouamor H, Pino J, Bali K (eds) Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6–10, 2023, Association for Computational Linguistics, pp 8596–8615, URL <https://aclanthology.org/2023.emnlp-main.531>
- Garneau N, Gaumond E, Lamontagne L, Déziel P (2021) Criminelbart: a french canadian legal language model specialized in criminal law. In: Proc. of the Int. Conf. on

- Artificial Intelligence and Law (ICAAIL), ACM, pp 256–257
- Garten J, Hoover J, Johnson KM, Boghrati R, Iskiwitch C, Dehghani M (2018) Dictionaries and distributions: Combining expert knowledge and large scale textual data content analysis: Distributed dictionary representation. *Behavior research methods* 50:344–361
- Geffet M, Dagan I (2005) The distributional inclusion hypotheses and lexical entailment. In: *Procs. 43rd Annual Meeting of the Association for Computational Linguistics*, pp 107–114, DOI 10.3115/1219840.1219854
- Gentner D, France I (1988) The verb mutability effect: Studies of the combinatorial semantics of nouns and verbs. In: *Lexical Ambiguity Resolution*, Morgan Kaufmann
- Ghifary M, Kleijn WB, Zhang M, Balduzzi D, Li W (2016) Deep reconstruction-classification networks for unsupervised domain adaptation. In: *Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, Springer, pp 597–613
- Gloor PA (2022) *Measuring moral values from facial expressions*. Edward Elgar Publishing
- Goebel R, Kano Y, Kim MY, Rabelo J, Satoh K, Yoshioka M (eds) (2021) *Proc. of the Eighth Int. Competition on Legal Information Extraction/Entailment (COLIEE 2021)*
- González-Fernández C, Fernández-Isabel A, Martín de Diego I, Fernández RR, Viseu Pinheiro J (2021) Experts perception-based system to detect misinformation in health websites. *Pattern Recognition Letters* 152:333–339
- Goodman JK, Cryder CE, Cheema A (2013) Data collection in a flat world: The strengths and weaknesses of mechanical turk samples. *Journal of Behavioral Decision Making* 26(3):213–224
- Goyal N, Du J, Ott M, Anantharaman G, Conneau A (2021) Larger-Scale Transformers for Multilingual Masked Language Modeling. In: *Proc. of the Workshop on Representation Learning for NLP, RepL4NLP@ACL-IJCNLP 2021*, Association for Computational Linguistics, pp 29–33, DOI 10.18653/v1/2021.repl4nlp-1.4
- Graham J, Nosek BA, Haidt J, Iyer R, Spassena K, Ditto PH (2008) Moral foundations questionnaire. *Journal of Personality and Social Psychology*
- Graham J, Haidt J, Nosek BA (2009) Liberals and conservatives rely on different sets of moral foundations. *Journal of personality and social psychology* 96(5):1029
- Graham J, Nosek BA, Haidt J, Iyer R, Koleva S, Ditto PH (2011) Mapping the moral domain. *Journal of personality and social psychology* 101(2):366
- Graham J, Haidt J, Koleva S, Motyl M, Iyer R, Wojcik SP, Ditto PH (2013) Moral Foundations Theory: The pragmatic validity of moral pluralism. In: *Advances in Experimental Social Psychology*, vol 47, Elsevier, pp 55–130, DOI 10.1016/B978-0-12-407236-7.00002-4, URL <https://linkinghub.elsevier.com/retrieve/pii/B9780124072367000024>

- Greco CM, Tagarelli A (2023) Topic similarities in rights and duties across european constitutions using transformer-based language models. In: Wehnert S, Fiorelli M, Picca D, Luca EWD, Stellato A (eds) Proceedings of the 1st Legal Information Retrieval meets Artificial Intelligence Workshop LIRAI 2023 co-located with the 34th ACM Hypertext Conference HT 2023, Rome, Italy, September 04, 2023, CEUR-WS.org, CEUR Workshop Proceedings, vol 3594, pp 47–62, URL <https://ceur-ws.org/Vol-3594/paper4.pdf>
- Greco CM, Tagarelli A (2024a) Bringing order into the realm of transformer-based language models for artificial intelligence and law. *Artif Intell Law* 32(4):863–1010, DOI 10.1007/S10506-023-09374-7, URL <https://doi.org/10.1007/s10506-023-09374-7>
- Greco CM, Tagarelli A (2024b) Evaluating gpt-3.5’s awareness and summarization abilities for european constitutional texts with shared topics. *CoRR* abs/2401.14524, DOI 10.48550/ARXIV.2401.14524, URL <https://doi.org/10.48550/arXiv.2401.14524>, 2401.14524
- Greco CM, Tagarelli A, Zumpano E (2022) A comparison of transformer-based language models on NLP benchmarks. In: Rosso P, Basile V, Martínez R, Métais E, Meziane F (eds) Natural Language Processing and Information Systems - 27th International Conference on Applications of Natural Language to Information Systems, NLDB 2022, Valencia, Spain, June 15-17, 2022, Proceedings, Springer, Lecture Notes in Computer Science, vol 13286, pp 490–501, DOI 10.1007/978-3-031-08473-7_45, URL https://doi.org/10.1007/978-3-031-08473-7_45
- Greco CM, Simeri A, Tagarelli A, Zumpano E (2023) Transformer-based language models for mental health issues: A survey. *Pattern Recognit Lett* 167:204–211, DOI 10.1016/J.PATREC.2023.02.016, URL <https://doi.org/10.1016/j.patrec.2023.02.016>
- Greco CM, Cava LL, Tagarelli A (2024a) Talking the talk does not entail walking the walk: On the limits of large language models in lexical entailment recognition. In: Al-Onaizan Y, Bansal M, Chen Y (eds) Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024, Association for Computational Linguistics, pp 14991–15011, URL <https://aclanthology.org/2024.findings-emnlp.881>
- Greco CM, Zangari L, Picca D, Tagarelli A (2024b) E2mocache: A dataset for emotional, event and moral observations in news articles on high-impact legal cases. *CoRR* abs/2409.09001, DOI 10.48550/ARXIV.2409.09001, URL <https://doi.org/10.48550/arXiv.2409.09001>, 2409.09001
- Guillou P (2020) GPortuguese-2 (Portuguese GPT-2 small): a Language Model for Portuguese text generation (and more NLP tasks...). Tech. rep.
- Guo J, Fan Y, Ai Q, Croft WB (2016) A deep relevance matching model for ad-hoc retrieval. In: Proc. of the ACM Conf. on Information and Knowledge Management (CIKM), ACM, pp 55–64

- Guo S, Mokherian N, Lerman K (2023a) A data fusion framework for multi-domain morality learning. In: Lin Y, Cha M, Quercia D (eds) Proceedings of the Seventeenth International AAAI Conference on Web and Social Media, ICWSM 2023, Limassol, Cyprus, June 5-8, 2023, AAAI Press, pp 281–291, DOI 10.1609/ICWSM.V17I1.22145, URL <https://doi.org/10.1609/icwsml.v17i1.22145>
- Guo Z, Jin R, Liu C, Huang Y, Shi D, Supryadi, Yu L, Liu Y, Li J, Xiong B, Xiong D (2023b) Evaluating large language models: A comprehensive survey. DOI 10.48550/ARXIV.2310.19736, URL <https://doi.org/10.48550/arXiv.2310.19736>
- Gururangan S, Marasovic A, Swayamdipta S, Lo K, Beltagy I, Downey D, Smith NA (2020) Don't stop pretraining: Adapt language models to domains and tasks. In: Proc. of the Annual Meeting of the Association for Computational Linguistics (ACL), Association for Computational Linguistics, pp 8342–8360
- Gutiérrez-Fandiño A, Armengol-Estapé J, Gonzalez-Agirre A, Villegas M (2021a) Spanish Legalese Language Model and Corpora. CoRR abs/2110.12201
- Gutiérrez-Fandiño A, Armengol-Estapé J, Pàmies M, Llop-Palao J, Silveira-Ocampo J, Carrino CP, Gonzalez-Agirre A, Armentano-Oller C, Penagos CR, Villegas M (2021b) Spanish language models. CoRR abs/2107.07253
- Haidt J, Graham J (2007) When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize. *Social justice research* 20(1):98–116
- Haidt J, Joseph C (2004a) Intuitive ethics: how innately prepared intuitions generate culturally variable virtues. *Daedalus* 133(4):55–66, DOI 10.1162/0011526042365555, URL <https://doi.org/10.1162/0011526042365555>
- Haidt J, Joseph C (2004b) Intuitive ethics: How innately prepared intuitions generate culturally variable virtues. *Daedalus* 133(4):55–66
- Hämmerl K, Deiseroth B, Schramowski P, Libovický J, Rothkopf CA, Fraser A, Kersting K (2023) Speaking multiple languages affects the moral bias of language models. In: Rogers A, Boyd-Graber JL, Okazaki N (eds) Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023, Association for Computational Linguistics, pp 2137–2156, DOI 10.18653/V1/2023.FINDINGS-S-ACL.134, URL <https://doi.org/10.18653/v1/2023.findings-acl.134>
- Haque F, Nur RU, Jahan SA, Mahmud Z, Shah FM (2020) A transformer based approach to detect suicidal ideation using pre-trained language models. In: Proc. ICCIT, pp 1–5
- He P, Liu X, Gao J, Chen W (2021) Deberta: decoding-enhanced bert with disentangled attention. In: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021, OpenReview.net, URL <https://openreview.net/forum?id=XPZiaotutsD>
- He P, Gao J, Chen W (2023) Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. In: The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May

- 1-5, 2023, OpenReview.net, URL <https://openreview.net/forum?id=sE7-XhLxHA>
- He Z, Guo S, Rao A, Lerman K (2024) Whose emotions and moral sentiments do language models reflect? In: Ku LW, Martins A, Srikumar V (eds) Findings of the Association for Computational Linguistics ACL 2024, Association for Computational Linguistics, Bangkok, Thailand and virtual meeting, pp 6611–6631, DOI 10.18653/v1/2024.findings-acl.395, URL <https://aclanthology.org/2024.findings-acl.395>
- Henderson P, Krass MS, Zheng L, Guha N, Manning CD, Jurafsky D, Ho DE (2022) Pile of Law: Learning Responsible Data Filtering from the Law and a 256GB Open-Source Legal Dataset. In: Proc. of the Annual Conf. on Neural Information Processing Systems (NeurIPS)
- Hendrycks D, Burns C, Basart S, Zou A, Mazeika M, Song D, Steinhardt J (2021a) Measuring massive multitask language understanding. In: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021, OpenReview.net, URL <https://openreview.net/forum?id=d7KBjmI3GmQ>
- Hendrycks D, Burns C, Chen A, Ball S (2021b) CUAD: an expert-annotated NLP dataset for legal contract review. In: Proc. of the Annual Conf. on Neural Information Processing Systems (NeurIPS)
- Hendrycks D, Burns C, Kadavath S, Arora A, Basart S, Tang E, Song D, Steinhardt J (2021c) Measuring mathematical problem solving with the MATH dataset. In: Vanschoren J, Yeung S (eds) Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual, URL <https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/be83ab3ecd0db773eb2dc1b0a17836a1-Abstract-round2.html>
- Hermann KM, Kociský T, Grefenstette E, Espeholt L, Kay W, Suleyman M, Blunsom P (2015) Teaching machines to read and comprehend. In: Proc. NIPS, pp 1693–1701
- Heron L, Kim J, Lee M, Haddad KE, Dupont S, Dutoit T, Truong KP (2018) A dyadic conversation dataset on moral emotions. In: IEEE International Conference on Automatic Face & Gesture Recognition
- Hofmann W, Wisneski DC, Brandt MJ, Skitka LJ (2014) Morality in everyday life. *Science* 345(6202):1340–1343
- Holzenberger N, Durme BV (2021) Factoring statutory reasoning as language understanding challenges. In: Proc. of the Annual Meeting of the Association for Computational Linguistics and the Int. Joint Conf. on Natural Language Processing (ACL/IJCNLP), Association for Computational Linguistics, pp 2742–2758
- Holzenberger N, Blair-Stanek A, Durme BV (2020) A dataset for statutory reasoning in tax law entailment and question answering. In: Proc. of the 2020 Workshop on Natural Legal Language Processing (NLLP), CEUR Workshop Proceedings, vol 2645, pp 31–38

- Hong J, Chong D, Manning CD (2021) Learning from limited labels for long legal dialogue. In: Proc. of the 2021 Workshop on Natural Legal Language Processing (NLLP), ACL, pp 190–204
- Hoover J, Portillo-Wightman G, Yeh L, Havaladar S, Davani AM, Lin Y, Kennedy B, Atari M, Kamel Z, Mendlen M, Moreno G, Park C, Chang TE, Chin J, Leong C, Leung JY, Mirinjian A, Dehghani M (2020a) Moral Foundations Twitter Corpus: A Collection of 35k Tweets Annotated for Moral Sentiment. *Social Psychological and Personality Science* 11(8):1057–1071, DOI 10.1177/1948550619876629, URL <http://journals.sagepub.com/doi/10.1177/1948550619876629>
- Hoover J, Portillo-Wightman G, Yeh L, Havaladar S, Davani AM, Lin Y, Kennedy B, Atari M, Kamel Z, Mendlen M, et al. (2020b) Moral foundations twitter corpus: A collection of 35k tweets annotated for moral sentiment. *Social Psychological and Personality Science* 11(8):1057–1071
- Hopp FR, Fisher JT, Cornell D, Huskey R, Weber R (2021) The extended moral foundations dictionary (emfd): Development and applications of a crowd-sourced approach to extracting moral intuitions from text. *Behavior research methods* 53:232–246
- Horne Z, Powell D (2016) How large is the role of emotion in judgments of moral dilemmas? *PloS one* 11(7):e0154780
- Houlsby N, Giurciu A, Jastrzebski S, Morrone B, de Laroussilhe Q, Gesmundo A, Attariyan M, Gelly S (2019) Parameter-efficient transfer learning for NLP. In: Proc. of the Int. Conf. on Machine Learning (ICML), PMLR, Proceedings of Machine Learning Research, vol 97, pp 2790–2799
- Howe JST, Khang LH, Chai IE (2019) Legal area classification: A comparative study of text classifiers on singapore supreme court judgments. *CoRR abs/1904.06470*
- Hromic H, Hayes C (2014) Constructing twitter datasets using signals for event detection evaluation. URL <https://aran.library.nuigalway.ie/bitstream/10379/4828/1/hromichayes-synthdatasignals.pdf>
- Hu J, Ruder S, Siddhant A, Neubig G, Firat O, Johnson M (2020) XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. *CoRR abs/2003.11080*
- Huang K, Altosaar J, Ranganath R (2019) Clinicalbert: Modeling clinical notes and predicting hospital readmission. *CoRR abs/1904.05342*
- Huang Y, Yu Z, Guo J, Xiang Y, Xian Y (2021) Element graph-augmented abstractive summarization for legal public opinion news with graph transformer. *Neurocomputing* 460:166–180
- Hudzina J, Madan K, Chinnappa D, Harmouche J, Bretz H, Vold A, Schilder F (2020) Information extraction/entailment of common law and civil code. In: *New Frontiers in Artificial Intelligence - JSAI-isAI 2020 Workshops, JURISIN, LENLS 2020 Workshops*, Springer, Lecture Notes in Computer Science, vol 12758, pp 254–268

- Hui K, Yates A, Berberich K, de Melo G (2017) PACRR: A position-aware neural IR model for relevance matching. In: Proc. of the Conf. on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, pp 1049–1058
- Hulpus I, Kobbe J, Stuckenschmidt H, Hirst G (2020) Knowledge graphs meet moral values. In: Gurevych I, Apidianaki M, Faruqui M (eds) Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics, *SEM@COLING 2020, Barcelona, Spain (Online), December 12–13, 2020, Association for Computational Linguistics, pp 71–80, URL <https://aclanthology.org/2020.starsem-1.8/>
- Hutto CJ, Gilbert E (2014) VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. In: Proc. ICWSM
- Iandola FN, Shaw AE, Krishna R, Keutzer K (2020) SqueezeBERT: What can computer vision teach NLP about efficient neural networks? In: Moosavi NS, Fan A, Shwartz V, Glavas G, Joty SR, Wang A, Wolf T (eds) Proceedings of SustainNLP: Workshop on Simple and Efficient Natural Language Processing, SustainNLP@EMNLP 2020, Online, November 20, 2020, Association for Computational Linguistics, pp 124–135, DOI 10.18653/V1/2020.SUSTAINLP-1.17, URL <https://doi.org/10.18653/v1/2020.sustainlp-1.17>
- II MJB, Katz DM (2022) GPT takes the bar exam. CoRR abs/2212.14402
- Inkpen D, Skaik R, Buddhitha P, Angelov D, Fredenburgh MT (2021) uOttawa at eRisk 2021: Automatic Filling of the Beck’s Depression Inventory Questionnaire using Deep Learning. In: Working Notes of CLEF, CEUR Workshop Proceedings, vol 2936, pp 966–980
- Iyer R, Koleva S, Graham J, Ditto P, Haidt J (2012) Understanding libertarian morality: The psychological dispositions of self-identified libertarians
- Jain D, Borah MD, Biswas A (2021) Summarization of Indian Legal Judgement Documents via Ensembling of Contextual Embedding based MLP Models. In: Working Notes of FIRE 2021 - Forum for Information Retrieval Evaluation, CEUR-WS.org, CEUR Workshop Proceedings, vol 3159, pp 553–561
- Ji S, Yu CP, Fung S, Pan S, Long G (2018) Supervised learning for suicidal ideation detection in online user content. Complex 2018:6157249:1–6157249:10
- Ji S, Zhang T, Ansari L, Fu J, Tiwari P, Cambria E (2021) MentalBERT: Publicly Available Pretrained Language Models for Mental Healthcare. CoRR abs/2110.15621
- Jiang AQ, Sablayrolles A, Mensch A, Bamford C, Chaplot DS, de Las Casas D, Bressand F, Lengyel G, Lample G, Saulnier L, Lavaud LR, Lachaux M, Stock P, Scao TL, Lavril T, Wang T, Lacroix T, Sayed WE (2023) Mistral 7b. CoRR abs/2310.06825
- Jiang J, Ferrara E (2024) Social-llm: Modeling user behavior at scale using language models and social network data. CoRR abs/2401.00893, DOI 10.48550/ARXIV.2401.00893, URL <https://doi.org/10.48550/arXiv.2401.00893>

- Jiang J, Luceri L, Ferrara E (2024) Moral values underpinning COVID-19 online communication patterns. CoRR abs/2401.08789, DOI 10.48550/ARXIV.2401.08789, URL <https://doi.org/10.48550/arXiv.2401.08789>
- Jiang L, Hwang JD, Bhagavatula C, Bras RL, Liang J, Dodge J, Sakaguchi K, Forbes M, Borchardt J, Gabriel S, et al. (2021) Can machines learn morality? the delphi experiment. arXiv preprint arXiv:211007574
- Jiang Z, Levitan SI, Zomick J, Hirschberg J (2020) Detection of mental health from Reddit via deep contextualized representations. In: Proc. LOUHI@EMNLP, pp 147–156
- Jin Z, Levine S, Adauto FG, Kamal O, Sap M, Sachan M, Mihalcea R, Tenenbaum J, Schölkopf B (2022) When to make exceptions: Exploring language models as accounts of human moral judgment. In: Koyejo S, Mohamed S, Agarwal A, Belgrave D, Cho K, Oh A (eds) Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022, URL http://papers.nips.cc/paper_files/paper/2022/hash/b654d6150630a5ba5df7a55621390daf-Abstract-Conference.html
- Johnson K, Goldwasser D (2018) Classification of moral foundations in microblog political discourse. In: Gurevych I, Miyao Y (eds) Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, pp 720–730, DOI 10.18653/v1/P18-1067, URL <https://aclanthology.org/P18-1067>
- Johnson RL, Pistilli G, Menéndez-González N, Duran LDD, Panai E, Kalpokiene J, Bertulfo DJ (2022) The ghost in the machine has an american accent: value conflict in GPT-3. CoRR abs/2203.07785, DOI 10.48550/ARXIV.2203.07785, URL <https://doi.org/10.48550/arXiv.2203.07785>
- Joshi M, Choi E, Weld DS, Zettlemoyer L (2017) TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. In: Proc. of the Annual Meeting of the Association for Computational Linguistics (ACL), Association for Computational Linguistics, pp 1601–1611, DOI 10.18653/v1/P17-1147
- Joshi M, Chen D, Liu Y, Weld DS, Zettlemoyer L, Levy O (2020) Spanbert: Improving pre-training by representing and predicting spans. Trans Assoc Comput Linguistics 8:64–77, DOI 10.1162/TACL_A_00300, URL https://doi.org/10.1162/tac1_a_00300
- Joulin A, Grave E, Bojanowski P, Mikolov T (2017) Bag of tricks for efficient text classification. In: Proc. of the Conf. of the European Chapter of the Association for Computational Linguistics (EACL), Association for Computational Linguistics, pp 427–431
- Kalamkar P, Tiwari A, Agarwal A, Karn S, Gupta S, Raghavan V, Modi A (2022) Corpus for automatic structuring of legal documents. In: Proc. of the Language Resources and Evaluation Conference (LREC), European Language Resources Association, pp

4420–4429

- Kang G, Jiang L, Yang Y, Hauptmann AG (2019) Contrastive adaptation network for unsupervised domain adaptation. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp 4893–4902
- Karami B, Bakouie F, Gharibzadeh S (2023) A transformer-based deep learning model for persian moral sentiment analysis. *Journal of Information Science* DOI 10.1177/01655515231188344
- Kennedy B, Atari M, Davani AM, Hoover J, Omrani A, Graham J, Dehghani M (2021) Moral concerns are differentially observable in language. *Cognition* 212:104696
- Kepplinger HM, Geiss S, Siebert S (2012) Framing scandals: Cognitive and emotional media effects. *Journal of Communication* 62(4):659–681
- Khazaeli S, Punuru J, Morris C, Sharma S, Staub B, Cole M, Chiu-Webster S, Sakalley D (2021) A Free Format Legal Question Answering System. In: *Proc. of the 2021 Workshop on Natural Legal Language Processing (NLLP)*, ACL, pp 107—113
- Kim MY, Rabelo J, Goebel R (2021) BM25 and Transformer-based Legal Information Extraction and Entailment. In: *Proc. of the Eighth Int. Competition on Legal Information Extraction/Entailment (COLIEE 2021)*, pp 25–30
- Klaus S, Hecke RV, Naini KD, Altingovde IS, Bernabé-Moreno J, Herrera-Viedma E (2022) Summarizing legal regulatory documents using transformers. In: Amigó E, Castells P, Gonzalo J, Carterette B, Culpepper JS, Kazai G (eds) *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Madrid, Spain, July 11 - 15, 2022, ACM, pp 2426–2430, DOI 10.1145/3477495.3531872, URL <https://doi.org/10.1145/3477495.3531872>
- Kobbe J, Rehbein I, Hülpuş I, Stuckenschmidt H (2020) Exploring morality in argumentation. In: Cabrio E, Villata S (eds) *Proceedings of the 7th Workshop on Argument Mining*, Association for Computational Linguistics, Online, pp 30–40, URL <https://aclanthology.org/2020.argmining-1.4>
- Kojima T, Gu SS, Reid M, Matsuo Y, Iwasawa Y (2022) Large language models are zero-shot reasoners. In: *Proc. of the Annual Conf. on Neural Information Processing Systems (NeurIPS)*
- Köpf A, Kilcher Y, von Rütte D, Anagnostidis S, Tam ZR, Stevens K, Barhoum A, Nguyen D, Stanley O, Nagyfi R, ES S, Suri S, Glushkov D, Dantuluri A, Maguire A, Schuhmann C, Nguyen H, Mattick A (2023) Openassistant conversations - democratizing large language model alignment. In: Oh A, Naumann T, Globerson A, Saenko K, Hardt M, Levine S (eds) *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, URL http://papers.nips.cc/paper_files/paper/2023/hash/949f0f8f32267d297c2d4e3ee10a2e7e-Abstract-Datasets_and_Benchmarks.html
- Koreeda Y, Manning CD (2021) Contractnli: A dataset for document-level natural

- language inference for contracts. In: Findings of the Association for Computational Linguistics: EMNLP 2021, Association for Computational Linguistics, pp 1907–1919
- Koutsikakis J, Chalkidis I, Malakasiotis P, Androutsopoulos I (2020) GREEK-BERT: The Greeks visiting Sesame Street. In: Proc. of the Hellenic Conf. on Artificial Intelligence, ACM, pp 110–117
- Kruiper R, Konstas I, Gray A, Sadeghineko F, Watson R, Kumar B (2021) SPAR.txt, a cheap Shallow Parsing approach for Regulatory texts. In: Proc. of the 2021 Workshop on Natural Legal Language Processing (NLLP), ACL, pp 129–143
- Kudo T, Richardson J (2018) Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In: Proc. of the Conf. on Empirical Methods in Natural Language Processing (EMNLP): System Demonstrations, Association for Computational Linguistics, pp 66–71, DOI 10.18653/v1/d18-2012
- Kumar A, Sharma K, Sharma A (2021) Hierarchical deep neural network for mental stress state detection using iot based biomarkers. *Pattern Recognition Letters* 145:81–87
- Kwiatkowski T, Palomaki J, Redfield O, Collins M, Parikh AP, Alberti C, Epstein D, Polosukhin I, Devlin J, Lee K, Toutanova K, Jones L, Kelcey M, Chang M, Dai AM, Uszkoreit J, Le Q, Petrov S (2019) Natural questions: a benchmark for question answering research. *Trans Assoc Comput Linguistics* 7:452–466
- Kádár DZ, Parvaresh V, Ning P (2019) Morality, moral order, and language conflict and aggression: A position paper. *Journal of Language Aggression and Conflict* 7(1):6–31, DOI 10.1075/jlac.00017.kad, URL <https://www.jbe-platform.com/content/journals/10.1075/jlac.00017.kad>
- Lage-Freitas A, Allende-Cid H, Santana O, de Oliveira-Lage L (2022) Predicting Brazilian court decisions. *PeerJ Comput Sci* 8:e904, DOI 10.7717/peerj-cs.904
- Lai G, Xie Q, Liu H, Yang Y, Hovy EH (2017) RACE: large-scale reading comprehension dataset from examinations. In: Palmer M, Hwa R, Riedel S (eds) *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9–11, 2017*, Association for Computational Linguistics, pp 785–794, DOI 10.18653/V1/D17-1082, URL <https://doi.org/10.18653/v1/d17-1082>
- Lam JT, Liang D, Dahan S, Zulkernine FH (2020) The gap between deep learning and law: Predicting employment notice. In: Proc. of the 2020 Workshop on Natural Legal Language Processing (NLLP), CEUR Workshop Proceedings, vol 2645, pp 52–56
- Lan Z, Chen M, Goodman S, Gimpel K, Sharma P, Soricut R (2020) ALBERT: A lite BERT for self-supervised learning of language representations. In: 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26–30, 2020, OpenReview.net, URL <https://openreview.net/forum?id=H1eA7AEtvS>

- Le H, Vial L, Frej J, Segonne V, Coavoux M, Lecouteux B, Allauzen A, Crabbé B, Besacier L, Schwab D (2020) FlauBERT: Unsupervised Language Model Pre-training for French. In: Proc. of the Language Resources and Evaluation Conference (LREC), European Language Resources Association, pp 2479–2490
- Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, Kang J (2020) Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinform* 36(4):1234–1240
- Lei T, Barzilay R, Jaakkola TS (2016) Rationalizing neural predictions. In: Proc. of the Conf. on Empirical Methods in Natural Language Processing (EMNLP), The Association for Computational Linguistics, pp 107–117
- Lei Y, Miah MMM, Qamar A, Reddy SR, Tong J, Xu H, Huang R (2024) EMONA: event-level moral opinions in news articles. In: Duh K, Gómez-Adorno H, Bethard S (eds) Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16–21, 2024, Association for Computational Linguistics, pp 5239–5251, DOI 10.18653/V1/2024.NAACL-LONG.293, URL <https://doi.org/10.18653/v1/2024.naacl-long.293>
- Leivaditi S, Rossi J, Kanoulas E (2020) A benchmark for lease contract review. *CoRR abs/2010.10386*
- Lewis M, Liu Y, Goyal N, Ghazvininejad M, Mohamed A, Levy O, Stoyanov V, Zettlemoyer L (2020a) BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In: Proc. of the Annual Meeting of the Association for Computational Linguistics (ACL), Association for Computational Linguistics, pp 7871–7880, DOI 10.18653/v1/2020.acl-main.703
- Lewis PSH, Oguz B, Rinott R, Riedel S, Schwenk H (2020b) MLQA: evaluating cross-lingual extractive question answering. In: Jurafsky D, Chai J, Schluter N, Tetreault JR (eds) Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5–10, 2020, Association for Computational Linguistics, pp 7315–7330, DOI 10.18653/V1/2020.ACL-MAIN.653, URL <https://doi.org/10.18653/v1/2020.acl-main.653>
- Li D, Yang K, Zhang L, Yin D, Peng D (2021a) CLASS: A novel method for chinese legal judgments summarization. In: Proc. of the Int. Conf. on Computer Science and Application Engineering, ACM, pp 86:1–86:5
- Li J, Monroe W, Jurafsky D (2016a) Understanding neural networks through representation erasure. *CoRR abs/1612.08220*
- Li J, Zhao X, Liu J, Wen J, Yang M (2021b) SIAT@COLIEE-2021: Combining Statistics Recall and Semantic Ranking for Legal Case Retrieval and Entailment. In: Proc. of the Eighth Int. Competition on Legal Information Extraction/Entailment (COLIEE 2021), pp 31–37
- Li L, Dai Y, Tang D, Feng Z, Zhou C, Qiu X, Xu Z, Shi S (2022) MarkBERT: Marking Word Boundaries Improves Chinese BERT. *CoRR abs/2203.06378*

- Li M, Wang D, Lu Q, Long Y (2016b) Event based emotion classification for news articles. In: Park JC, Chung J (eds) Proceedings of the 30th Pacific Asia Conference on Language, Information and Computation, PACLIC 30, Seoul, Korea, October 28 - October 30, 2016, ACL, URL <https://aclanthology.org/Y16-2013/>
- Li X, Sun X, Meng Y, Liang J, Wu F, Li J (2020) Dice loss for data-imbalanced NLP tasks. In: Proc. ACL, pp 465–476
- Li Y, Su H, Shen X, Li W, Cao Z, Niu S (2017) Dailydialog: A manually labelled multi-turn dialogue dataset. In: Kondrak G, Watanabe T (eds) Proceedings of the Eighth International Joint Conference on Natural Language Processing, IJCNLP 2017, Taipei, Taiwan, November 27 - December 1, 2017 - Volume 1: Long Papers, Asian Federation of Natural Language Processing, pp 986–995, URL <https://aclanthology.org/I17-1099/>
- Liao J, Chen X, Du L (2023) Concept understanding in large language models: An empirical study. In: Maughan K, Liu R, Burns TF (eds) The First Tiny Papers Track at ICLR 2023, Tiny Papers @ ICLR 2023, Kigali, Rwanda, May 5, 2023, OpenReview.net, URL <https://openreview.net/pdf?id=losgEa0WIL7>
- Licari D, Comandè G (2022) ITALIAN-LEGAL-BERT: A Pre-trained Transformer Language Model for Italian Law. In: Proc. of the KM4LAW Workshop with the 23rd Int. Conf on Knowledge Engineering and Knowledge Management, CEUR, CEUR Workshop Proceedings, vol 3256
- Likert R (1932) A technique for the measurement of attitudes. Archives of psychology
- Limsopatham N (2021) Effectively Leveraging BERT for Legal Document Classification. In: Proc. of the 2021 Workshop on Natural Legal Language Processing (NLLP), ACL, pp 210—216
- Lin CY (2004a) ROUGE: A package for automatic evaluation of summaries. In: Text Summarization Branches Out, Association for Computational Linguistics, pp 74–81
- Lin CY (2004b) ROUGE: A package for automatic evaluation of summaries. In: Text Summarization Branches Out, pp 74–81
- Lin J (1991) Divergence measures based on the shannon entropy. IEEE Trans Inf Theory 37(1):145–151
- Lippi M, Palka P, Contissa G, Lagioia F, Micklitz H, Sartor G, Torroni P (2019) CLAUDETTE: an automated detector of potentially unfair clauses in online terms of service. Artif Intell Law 27(2):117–139
- Liscio E, Dondera A, Geadau A, Jonker CM, Murukannaiah PK (2022) Cross-domain classification of moral values. In: Carpuat M, de Marneffe M, Ruíz IVM (eds) Findings of the Association for Computational Linguistics: NAACL 2022, Seattle, WA, United States, July 10-15, 2022, Association for Computational Linguistics, pp 2727–2745, DOI 10.18653/v1/2022.FINDINGS-NAACL.209, URL <https://doi.org/10.18653/v1/2022.findings-naacl.209>
- Liu C, Chen K (2019) Extracting the gist of chinese judgments of the supreme court. In:

- Proc. of the Int. Conf. on Artificial Intelligence and Law (ICAIL), ACM, pp 73–82
- Liu D, Yan Y, Gong Y, Qi W, Zhang H, Jiao J, Chen W, Fu J, Shou L, Gong M, Wang P, Chen J, Jiang D, Lv J, Zhang R, Wu W, Zhou M, Duan N (2021a) GLGE: A new general language generation evaluation benchmark. In: Proc. ACL/IJCNLP, pp 408–420
- Liu L, Zhang W, Liu J, Shi W, Huang Y (2021b) Interpretable charge prediction for legal cases based on interdependent legal information. In: Proc. of the Joint Conf. on Neural Networks (IJCNN), IEEE, pp 1–8
- Liu P, Yuan W, Fu J, Jiang Z, Hayashi H, Neubig G (2023a) Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput Surv* 55(9):195:1–195:35
- Liu X, Cheng H, He P, Chen W, Wang Y, Poon H, Gao J (2020a) Adversarial training for large neural language models. *CoRR* abs/2004.08994
- Liu X, Yin D, Feng Y, Wu Y, Zhao D (2021c) Everything has a cause: Leveraging causal inference in legal text analysis. In: Proc. of the Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), Association for Computational Linguistics, pp 1928–1941
- Liu X, Wang J, Sun J, Yuan X, Dong G, Di P, Wang W, Wang D (2023b) Prompting frameworks for large language models: A survey. *arXiv preprint arXiv:2311.12785*
- Liu Y, Lapata M (2019) Text summarization with pretrained encoders. In: Proc. of the Conf. on Empirical Methods in Natural Language Processing and the Int. Joint Conf. on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, pp 3728–3738
- Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov V (2019) Roberta: A robustly optimized BERT pretraining approach. *CoRR* abs/1907.11692, URL <http://arxiv.org/abs/1907.11692>
- Liu Y, Gu J, Goyal N, Li X, Edunov S, Ghazvininejad M, Lewis M, Zettlemoyer L (2020b) Multilingual denoising pre-training for neural machine translation. *Trans Assoc Comput Linguistics* 8:726–742, DOI 10.1162/TACL_A_00343, URL https://doi.org/10.1162/tacl_a_00343
- Liu Y, Zhang XF, Wegsman D, Beauchamp N, Wang L (2022) POLITICS: pretraining with same-story article comparison for ideology prediction and stance detection. In: Carpuat M, de Marneffe M, Ruíz IVM (eds) Findings of the Association for Computational Linguistics: NAACL 2022, Seattle, WA, United States, July 10-15, 2022, Association for Computational Linguistics, pp 1354–1374, DOI 10.18653/V1/2022.FINDINGS-NAACL.101, URL <https://doi.org/10.18653/v1/2022.findings-naacl.101>
- Liu Z, Peach RL, Lawrance EL, Noble A, Ungless MA, Barahona M (2021d) Listening to mental health crisis needs at scale: Using natural language processing to understand and evaluate a mental health crisis text messaging service. *Frontiers in Digital*

Health 3

- Liu Z, Yang K, Zhang T, Xie Q, Yu Z, Ananiadou S (2024) Emollms: A series of emotional large language models and annotation tools for comprehensive affective analysis. CoRR abs/2401.08508, DOI 10.48550/ARXIV.2401.08508, URL <https://doi.org/10.48550/arXiv.2401.08508>, 2401.08508
- Locke D, Zuccon G (2022) Case law retrieval: problems, methods, challenges and evaluations in the last 20 years. CoRR abs/2202.07209
- Lokala U, Srivastava A, Dastidar TG, Chakraborty T, Akhtar MS, Panahiazar M, Sheth AP (2022) A Computational Approach to Understand Mental Health from Reddit: Knowledge-Aware Multitask Learning Framework. In: Proc. ICWSM, pp 640–650
- Lopez-Avila A, Suárez-Paniagua V (2023) Combining denoising autoencoders with contrastive learning to fine-tune transformer models. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pp 2021–2032
- Louis A, Spanakis G (2022) A statutory article retrieval dataset in french. In: Proc. of the Annual Meeting of the Association for Computational Linguistics (ACL), Association for Computational Linguistics, pp 6789–6803
- Louis A, van Dijk G, Spanakis G (2023) Finding the Law: Enhancing Statutory Article Retrieval via Graph Neural Networks. In: Proc. of the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2023), Association for Computational Linguistics, pp 2753–2768
- Loureiro D, Jorge A (2019) Language modelling makes sense: Propagating representations through wordnet for full-coverage word sense disambiguation. In: Korhonen A, Traum DR, Màrquez L (eds) Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers, Association for Computational Linguistics, pp 5682–5691, DOI 10.18653/V1/P19-1569, URL <https://doi.org/10.18653/v1/p19-1569>
- Lourie N, Bras RL, Bhagavatula C, Choi Y (2021) UNICORN on RAINBOW: A universal commonsense reasoning model on a new multitask benchmark. In: Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021, AAAI Press, pp 13480–13488, DOI 10.1609/AAAI.V35I15.17590, URL <https://doi.org/10.1609/aaai.v35i15.17590>
- Lovón-Melgarejo J, Moreno JG, Besançon R, Ferret O, Tamine L (2024) Probing pre-trained language models with hierarchy properties. In: Goharian N, Tonello N, He Y, Lipani A, McDonald G, Macdonald C, Ounis I (eds) Advances in Information Retrieval - 46th European Conference on Information Retrieval, ECIR 2024, Glasgow, UK, March 24-28, 2024, Proceedings, Part II, Springer, Lecture Notes in Computer Science, vol 14609, pp 126–142, DOI 10.1007/978-3-031-56060-6_9, URL https://doi.org/10.1007/978-3-031-56060-6_9
- Luceri L, Boniardi E, Ferrara E (2024) Leveraging large language models to detect

- influence campaigns on social media. In: Chua T, Ngo C, Lee RK, Kumar R, Lauw HW (eds) *Companion Proceedings of the ACM on Web Conference 2024*, WWW 2024, Singapore, Singapore, May 13-17, 2024, ACM, pp 1459–1467, DOI 10.1145/3589335.3651912, URL <https://doi.org/10.1145/3589335.3651912>
- Lyu Y, Wang Z, Ren Z, Ren P, Chen Z, Liu X, Li Y, Li H, Song H (2022) Improving legal judgment prediction through reinforced criminal element extraction. *Inf Process Manag* 59(1):102780
- Ma Y, Shao Y, Liu B, Liu Y, Zhang M, Ma S (2021) Retrieving Legal Cases from a Large-scale Candidate Corpus. In: *Proc. of the Eighth Int. Competition on Legal Information Extraction/Entailment (COLIEE 2021)*, pp 38–42
- Maas AL, Daly RE, Pham PT, Huang D, Ng AY, Potts C (2011) Learning word vectors for sentiment analysis. In: Lin D, Matsumoto Y, Mihalcea R (eds) *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference*, 19-24 June, 2011, Portland, Oregon, USA, The Association for Computer Linguistics, pp 142–150, URL <https://aclanthology.org/P11-1015/>
- MacAvaney S, Yates A, Cohan A, Goharian N (2019) CEDR: contextualized embeddings for document ranking. In: *Proc. of the ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR)*, ACM, pp 1101–1104
- Mahari RZ (2021) AutoLAW: Augmented Legal Reasoning through Legal Precedent Prediction. In: *Proc. of the 2021 Workshop on Natural Legal Language Processing (NLLP)*
- Majumder SB, Das D (2020) Rhetorical role labelling for legal judgements using ROBERTA. In: *Working Notes of FIRE 2020 - Forum for Information Retrieval Evaluation*, CEUR-WS.org, CEUR Workshop Proceedings, vol 2826, pp 22–25
- Malik V, Sanjay R, Nigam SK, Ghosh K, Guha SK, Bhattacharya A, Modi A (2021) ILDC for CJPE: indian legal documents corpus for court judgment prediction and explanation. In: *Proc. of the Annual Meeting of the Association for Computational Linguistics and the Int. Joint Conf. on Natural Language Processing (ACL/IJCNLP)*, Association for Computational Linguistics, pp 4046–4062
- Malviya K, Roy B, Saritha S (2021) A transformers approach to detect depression in social media. In: *Proc. ICAIS*, pp 718–723
- Mamakas D, Tsotsi P, Androutsopoulos I, Chalkidis I (2022) Processing Long Legal Documents with Pre-trained Transformers: Modding LegalBERT and Longformer. *CoRR abs/2211.00974*
- Manor L, Li JJ (2019) Plain english summarization of contracts. *CoRR abs/1906.00424*
- Martin L, Müller B, Suárez PJO, Dupont Y, Romary L, de la Clergerie É, Seddah D, Sagot B (2020) CamemBERT: a Tasty French Language Model. In: *Proc. of the Annual Meeting of the Association for Computational Linguistics (ACL)*, Association for Computational Linguistics, pp 7203–7219

- Martinez-Castaño R, Htait A, Azzopardi L, Moshfeghi Y (2020) Early Risk Detection of Self-Harm and Depression Severity using BERT-based Transformers. In: Working Notes of CLEF
- Martino GD, Pio G, Ceci M (2022) PRILJ: an efficient two-step method based on embedding and clustering for the identification of regularities in legal case judgments. *Artif Intell Law* 30(3):359–390
- Masala M, Ruseti S, Dascalu M (2020) RoBERT - A Romanian BERT Model. In: Proc. of the Int. Conf. on Computational Linguistics (COLING), Int. Committee on Computational Linguistics, pp 6626–6637
- Masala M, Iacob RCA, Uban AS, Cidota M, Velicu H, Rebedea T, Popescu M (2021) jurBERT: A Romanian BERT Model for Legal Judgement Prediction. In: Proc. of the 2021 Workshop on Natural Legal Language Processing (NLLP), ACL, pp 86–94
- Matero M, Hung A, Schwartz HA (2022) Evaluating contextual embeddings and their extraction layers for depression assessment. In: Proc. 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis, pp 89–94
- Maupomé D, Armstrong MD, Rancourt F, Soulas T, Meurs M (2021) Early Detection of Signs of Pathological Gambling, Self-Harm and Depression through Topic Extraction and Neural Networks. In: Working Notes of CLEF, CEUR Workshop Proceedings, vol 2936, pp 1031–1045
- Mauriello ML, Lincoln T, Hon G, Simon D, Jurafsky D, Paredes P (2021) SAD: A Stress Annotated Dataset for Recognizing Everyday Stressors in SMS-like Conversational Systems. In: Proc. CHI, ACM, pp 399:1–399:7
- McCurrie CH, Crone DL, Bigelow F, Laham SM (2018) Moral and affective film set (maafs): A normed moral video database. *PloS one* 13(11):e0206604
- McInnes L, Healy J, Melville J (2018a) Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:180203426*
- McInnes L, Healy J, Saul N, Grossberger L (2018b) Umap: Uniform manifold approximation and projection. *The Journal of Open Source Software* 3(29):861
- Meng Y, Speier W, Ong MK, Arnold CW (2021) Bidirectional representation learning from transformers using multimodal electronic health record data to predict depression. *IEEE J Biomed Health Informatics* 25(8):3121–3129
- Mialon G, Dessì R, Lomeli M, Nalmpantis C, Pasunuru R, Raileanu R, Rozière B, Schick T, Dwivedi-Yu J, Celikyilmaz A, Grave E, LeCun Y, Scialom T (2023) Augmented language models: a survey. *CoRR abs/2302.07842*, DOI 10.48550/arXiv.2302.07842
- Mihalcea R, Tarau P (2004) TextRank: Bringing order into text. In: Proc. of the Conf. on Empirical Methods in Natural Language Processing (EMNLP), ACL, pp 404–411
- Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient estimation of word representations in vector space. In: Proc. of the Workshops of the Int. Conf. on Learning Representations (ICLR)

- Miller GA, Beckwith R, Fellbaum C, Gross D, Miller KJ (1990) Introduction to word-net: An on-line lexical database. *International journal of lexicography* 3(4):235–244
- Min B, Ross H, Sulem E, Veyseh APB, Nguyen TH, Sainz O, Agirre E, Heintz I, Roth D (2024) Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Comput Surv* 56(2):30:1–30:40, DOI 10.1145/3605943, URL <https://doi.org/10.1145/3605943>
- Mistica M, Lau JH, Merrifield B, Fazio K, Baldwin T (2021) Semi-automatic triage of requests for free legal assistance. In: *Proc. of the 2021 Workshop on Natural Legal Language Processing (NLLP)*, ACL, pp 217–227
- Miyato T, Dai AM, Goodfellow IJ (2017) Adversarial training methods for semi-supervised text classification. In: *Proc. of the Int. Conf. on Learning Representations (ICLR)*
- Mohammad SM, Bravo-Marquez F, Salameh M, Kiritchenko S (2018) Semeval-2018 task 1: Affect in tweets. In: Apidianaki M, Mohammad SM, May J, Shutova E, Bethard S, Carpuat M (eds) *Proceedings of The 12th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2018*, New Orleans, Louisiana, USA, June 5-6, 2018, Association for Computational Linguistics, pp 1–17, DOI 10.18653/V1/S18-1001, URL <https://doi.org/10.18653/v1/s18-1001>
- Mokhberian N, Hopp FR, Harandizadeh B, Morstatter F, Lerman K (2022) Noise audits improve moral foundation classification. In: *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2022*, Istanbul, Turkey, November 10-13, 2022, IEEE, pp 147–154, DOI 10.1109/ASONAM55673.2022.10068681, URL <https://doi.org/10.1109/ASONAM55673.2022.10068681>
- Morris C, Ritzert M, Fey M, Hamilton WL, Lenssen JE, Rattan G, Grohe M (2019) Weisfeiler and leman go neural: Higher-order graph neural networks. In: *Proc. of the AAAI Conference on Artificial Intelligence (AAAI), the Conf. on Innovative Applications of Artificial Intelligence (IAAI), the AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI)*, AAAI Press, pp 4602–4609
- Moskvoretskii V, Neminova E, Lobanova A, Panchenko A, Nikishina I (2024) TaxoLLaMA: WordNet-based model for solving multiple lexical semantic tasks. In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, Bangkok, Thailand, pp 2331–2350, DOI 10.18653/v1/2024.acl-long.127, URL <https://aclanthology.org/2024.acl-long.127>
- Mukherjee S, Mitra A, Jawahar G, Agarwal S, Palangi H, Awadallah A (2023) Orca: Progressive learning from complex explanation traces of GPT-4. *CoRR* abs/2306.02707
- Murarka A, Radhakrishnan B, Ravichandran S (2020) Detection and classification of mental illnesses on social media using RoBERTa. *arXiv preprint arXiv:2011.11226*
- Nallapati R, Zhou B, dos Santos CN, Gülçehre Ç, Xiang B (2016) Abstractive text

- summarization using sequence-to-sequence rnns and beyond. In: Proc. CoNLL, pp 280–290
- Nallapati R, Zhai F, Zhou B (2017) Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In: Proc. of the AAAI Conference on Artificial Intelligence (AAAI), AAAI Press, pp 3075–3081
- Narayan S, Cohen SB, Lapata M (2018) Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In: Proc. EMNLP, pp 1797–1807
- Naseem U, Lee B, Khushi M, Kim J, Dunn AG (2022) Benchmarking for Public Health Surveillance tasks on Social Media with a Domain-Specific Pretrained Language Model. CoRR abs/2204.04521
- Navigli R, Conia S, Ross B (2023) Biases in large language models: Origins, inventory, and discussion. ACM J Data Inf Qual 15(2):10:1–10:21, DOI 10.1145/3597307, URL <https://doi.org/10.1145/3597307>
- Nguyen H (2023) A Brief Report on LawGPT 1.0: A Virtual Legal Assistant Based on GPT-3. CoRR abs/2302.05729
- Nguyen H, Nguyen L (2021) Sublanguage: A serious issue affects pretrained models in legal domain. CoRR abs/2104.07782
- Nguyen H, Vuong HT, Nguyen PM, Dang TB, Bui QM, Sinh VT, Nguyen CM, Tran VD, Satoh K, Nguyen ML (2020) JNLP team: Deep learning for legal processing in COLIEE 2020. In: Proc. of the 14th Int. Workshop on Juris-Informatics (JURISIN2020), The Japanese Society of Artificial Intelligence, pp 195–208
- Nguyen H, Nguyen PM, Vuong T, Bui QM, Nguyen CM, Dang TB, Tran V, Nguyen ML, Satoh K (2021a) JNLP team: Deep learning approaches for legal processing tasks in COLIEE 2021. In: Proc. of the Eighth Int. Competition on Legal Information Extraction/Entailment (COLIEE 2021), pp 46–53
- Nguyen H, Tran V, Nguyen PM, Vuong T, Bui QM, Nguyen CM, Dang TB, Nguyen ML, Satoh K (2021b) Paralaw nets - cross-lingual sentence-level pretraining for legal text processing. CoRR abs/2106.13403
- Nguyen TD, Chen Z, Carroll NG, Tran A, Klein C, Xie L (2024a) Measuring moral dimensions in social media with mformer. In: Proceedings of the International AAAI Conference on Web and Social Media, vol 18, pp 1134–1147
- Nguyen TD, Chen Z, Carroll NG, Tran A, Klein C, Xie L (2024b) Measuring moral dimensions in social media with mformer. In: Proceedings of the Eighteenth International AAAI Conference on Web and Social Media, ICWSM, AAAI Press, DOI 10.1609/ICWSM.V18I1.31378
- Niklaus J, Chalkidis I, Stürmer M (2021) Swiss-Judgment-Prediction: A Multilingual Legal Judgment Prediction Benchmark. In: Proc. of the 2021 Workshop on Natural Legal Language Processing (NLLP), ACL, pp 19–35
- Niklaus J, Stürmer M, Chalkidis I (2022) An empirical study on cross-x transfer for

- legal judgment prediction. In: Proc. of the Conf. of the Asia-Pacific Chapter of the Association for Computational Linguistics and the Int. Joint Conf. on Natural Language Processing (ACL/IJCNLP), Association for Computational Linguistics, pp 32–46
- Nisa QU, Muhammad R (2021) Towards transfer learning using bert for early detection of self-harm of social media users. In: Working Notes of CLEF, CEUR Workshop Proceedings, vol 2936, pp 1059–1070
- Nunes JL, Almeida GFCF, de Araújo M, Barbosa SDJ (2024) Are large language models moral hypocrites? A study based on moral foundations. CoRR abs/2405.11100, DOI 10.48550/ARXIV.2405.11100, URL <https://doi.org/10.48550/arXiv.2405.11100>, 2405.11100
- Oliveira HG (2023) On the acquisition of wordnet relations in portuguese from pre-trained masked language models. In: Proceedings of the 12th Global Wordnet Conference, pp 41–49
- Oliver Wendell Holmes J (1897) The Path of the Law. Harvard Law Review 10:457–478
- OpenAI (2023) GPT-4 technical report. CoRR abs/2303.08774, DOI 10.48550/ARXIV.2303.08774, URL <https://doi.org/10.48550/arXiv.2303.08774>, 2303.08774
- Ostendorff M, Blume T, Ostendorff S (2020) Towards an open platform for legal information. In: Proc. of the ACM/IEEE Joint Conf. on Digital Libraries (JCDL), pp 385–388, DOI 10.1145/3383583.3398616
- Ostendorff M, Ash E, Ruas T, Gipp B, Schneider JM, Rehm G (2021) Evaluating document representations for content-based legal literature recommendations. In: Proc. of the Int. Conf. on Artificial Intelligence and Law (ICAIL), ACM, pp 109–118, DOI 10.1145/3462757.3466073
- Ouyang L, Wu J, Jiang X, Almeida D, Wainwright CL, Mishkin P, Zhang C, Agarwal S, Slama K, Ray A, Schulman J, Hilton J, Kelton F, Miller L, Simens M, Askeel A, Welinder P, Christiano PF, Leike J, Lowe R (2022) Training language models to follow instructions with human feedback. In: Proc. of the Annual Conf. on Neural Information Processing Systems (NeurIPS)
- Pacheco ML, Goldwasser D (2021) Modeling content and context with deep relational learning. Trans Assoc Comput Linguistics 9:100–119, DOI 10.1162/TACL_A_00357, URL https://doi.org/10.1162/tacl_a_00357
- Pacheco ML, Islam T, Mahajan M, Shor A, Yin M, Ungar L, Goldwasser D (2022) A holistic framework for analyzing the COVID-19 vaccine debate. In: Carpuat M, de Marneffe MC, Meza Ruiz IV (eds) Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Seattle, United States, pp 5821–5839, DOI 10.18653/v1/2022.naacl-main.427, URL <https://aclanthology.org/2022.naacl-main.427>

- Pagliardini M, Gupta P, Jaggi M (2018) Unsupervised learning of sentence embeddings using compositional n-gram features. In: Proc. of the Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), Association for Computational Linguistics, pp 528–540
- Pais V, Mitrofan M, Gasan CL, Coneschi V, Ianov A (2021) Named entity recognition in the Romanian legal domain. In: Proc. of the 2021 Workshop on Natural Legal Language Processing (NLLP), Association for Computational Linguistics, pp 9–18
- Papaloukas C, Chalkidis I, Athinaios K, Pantazi D, Koubarakis M (2021) Multi-granular Legal Topic Classification on Greek Legislation. In: Proc. of the 2021 Workshop on Natural Legal Language Processing (NLLP), ACL, pp 63–75
- Paperno D, Kruszewski G, Lazaridou A, Pham QN, Bernardi R, Pezzelle S, Baroni M, Boleda G, Fernández R (2016) The LAMBADA dataset: Word prediction requiring a broad discourse context. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7–12, 2016, Berlin, Germany, Volume 1: Long Papers, The Association for Computer Linguistics, DOI 10.18653/V1/P16-1144, URL <https://doi.org/10.18653/v1/p16-1144>
- Paul S, Goyal P, Ghosh S (2022a) Lesicin: A heterogeneous graph-based approach for automatic legal statute identification from indian legal documents. In: Proc. of the AAAI Conference on Artificial Intelligence (AAAI), the Conf. on Innovative Applications of Artificial Intelligence (IAAI), the AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI), AAAI Press, pp 11139–11146
- Paul S, Mandal A, Goyal P, Ghosh S (2022b) Pre-training Transformers on Indian Legal Text. CoRR abs/2209.06049
- Pearce K, Zhan T, Komanduri A, Zhan J (2021) A comparative study of transformer-based language models on extractive question answering. CoRR abs/2110.03142
- Pennington J, Socher R, Manning CD (2014) Glove: Global vectors for word representation. In: Proc. of the Conf. on Empirical Methods in Natural Language Processing (EMNLP), ACL, pp 1532–1543
- Perçin S, Galassi A, Lagioia F, Ruggeri F, Santin P, Sartor G, Torroni P (2022) Combining wordnet and word embeddings in data augmentation for legal texts. In: Proceedings of the Natural Legal Language Processing Workshop 2022, pp 47–52
- Pérez JM, Furman DA, Alemany LA, Luque FM (2022) RoBERTuito: a pre-trained language model for social media text in Spanish. In: Proc. of the Language Resources and Evaluation Conf. (LREC), European Language Resources Association, pp 7235–7243
- Peters ME, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, Zettlemoyer L (2018) Deep contextualized word representations. In: Walker MA, Ji H, Stent A (eds) Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1–6, 2018, Volume 1 (Long Papers), Association for Computational Linguistics, pp 2227–2237, DOI 10.18653/V1/N18-1202, URL

<https://doi.org/10.18653/v1/n18-1202>

- Peters ME, Ruder S, Smith NA (2019) To tune or not to tune? adapting pretrained representations to diverse tasks. In: Augenstein I, Gella S, Ruder S, Kann K, Can B, Welbl J, Conneau A, Ren X, Rei M (eds) *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, Association for Computational Linguistics, Florence, Italy, pp 7–14, DOI 10.18653/v1/W19-4302, URL <https://aclanthology.org/W19-4302>
- Pfeiffer J, Vulic I, Gurevych I, Ruder S (2020) MAD-X: an adapter-based framework for multi-task cross-lingual transfer. In: *Proc. of the Conf. on Empirical Methods in Natural Language Processing (EMNLP)*, Association for Computational Linguistics, pp 7654–7673
- Plutchik R (2001) The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American scientist* 89(4):344–350
- Van de Poel I, Royakkers L (2023) *Ethics, technology, and engineering: An introduction*. John Wiley & Sons
- Polignano M, Basile P, de Gemmis M, Semeraro G, Basile V (2019) ALBERTo: Italian BERT Language Understanding Model for NLP Challenging Tasks Based on Tweets. In: *Proc. of the Sixth Italian Conference on Computational Linguistics, CEUR-WS.org, CEUR Workshop Proceedings*, vol 2481
- Polo FM, Mendonça GCF, Parreira KCJ, de Godoy Gianvechio L, Cordeiro P, Ferreira JB, de Lima LMP, do Amaral Maia AC, Vicente R (2021) LegalNLP - Natural Language Processing methods for the Brazilian Legal Language. *CoRR abs/2110.15709*
- Polsley S, Jhunjhunwala P, Huang R (2016) CaseSummarizer: A System for Automated Summarization of Legal Texts. In: *Proc. of the Int. Conf. on Computational Linguistics (COLING): System Demonstrations, ACL*, pp 258–262
- Ponte JM, Croft WB (1998) A language modeling approach to information retrieval. In: *Proc. of the ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR)*, ACM, pp 275–281
- Pradhan S, Moschitti A, Xue N, Uryupina O, Zhang Y (2012) CoNLL-2012 Shared Task: Modeling Multilingual Unrestricted Coreference in OntoNotes. In: *Proc. of the Joint Conf. on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL): Modeling Multilingual Unrestricted Coreference in OntoNotes, ACL*, pp 1–40
- Preteni V, Ghinassi I, Ive J, Saitis C, Kalimeri K (2024) Moralbert: A fine-tuned language model for capturing moral values in social discussions. In: *Procs. of the 2024 International Conference on Information Technology for Social Good, Association for Computing Machinery, New York, NY, USA, GoodIT '24*, p 433–442, DOI 10.1145/3677525.3678694, URL <https://doi.org/10.1145/3677525.3678694>
- Putra IMS, Siahaan D, Saikhu A (2023) Recognizing textual entailment: A review

- of resources, approaches, applications, and challenges. *ICT Express* DOI <https://doi.org/10.1016/j.ict.2023.08.012>
- Pyysalo S, Kanerva J, Virtanen A, Ginter F (2021) WikiBERT Models: Deep Transfer Learning for Many Languages. In: *Proc. of the Nordic Conf. on Computational Linguistics (NoDaLiDa)*, pp 1–10
- Quemy A, Wrembel R (2022) ECHR-OD: on building an integrated open repository of legal documents for machine learning applications. *Inf Syst* 106:101822
- Rabelo J, Kim M, Goebel R (2019a) Combining similarity and transformer methods for case law entailment. In: *Proc. of the Int. Conf. on Artificial Intelligence and Law (ICAIL)*, ACM, pp 290–296, DOI 10.1145/3322640.3326741
- Rabelo J, Kim MY, Goebel R, Yoshioka M, Kano Y, Satoh K (2019b) A Summary of the COLIEE 2019 Competition. In: *New Frontiers in Artificial Intelligence: JSAI-isAI Int. Workshops, JURISIN, AI-Biz, LENLS, Kansei-AI*, pp 34–49, DOI https://doi.org/10.1007/978-3-030-58790-1_3
- Rabelo J, Kim M, Goebel R (2020) The application of text entailment techniques in COLIEE 2020. In: *New Frontiers in Artificial Intelligence - JSAI-isAI 2020 Workshops, JURISIN, LENLS 2020 Workshops*, Springer, Lecture Notes in Computer Science, vol 12758, pp 240–253, DOI 10.1007/978-3-030-79942-7_16
- Rabelo J, Goebel R, Kim MY, Kano Y, Yoshioka M, Satoh K (2022) Overview and Discussion of the Competition on Legal Information Extraction/Entailment (COLIEE) 2021. *The Review of Socionetwork Strategies* DOI 10.1007/s12626-022-00105-z
- Radford A, Narasimhan K, Salimans T, Sutskever I (2018) Improving language understanding by generative pre-training. *OpenAI blog* URL <https://openai.com/research/language-unsupervised>
- Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I, et al. (2019) Language models are unsupervised multitask learners. *OpenAI blog* 1(8):9
- Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, Zhou Y, Li W, Liu PJ (2020) Exploring the limits of transfer learning with a unified text-to-text transformer. *J Mach Learn Res* 21:140:1–140:67, URL <https://jmlr.org/papers/v21/20-074.html>
- Rajpurkar P, Zhang J, Lopyrev K, Liang P (2016) Squad: 100, 000+ questions for machine comprehension of text. In: Su J, Carreras X, Duh K (eds) *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, The Association for Computational Linguistics, pp 2383–2392, DOI 10.18653/V1/D16-1264, URL <https://doi.org/10.18653/v1/d16-1264>
- Rana SS, Solaiman M (2022) Moral identity, consumption values and green purchase behaviour. *Journal of Islamic Marketing* 14(10):2550–2574, DOI 10.1108/JIMA-01-2021-0030, URL <https://doi.org/10.1108/JIMA-01-2021-0030>
- Ranaldi L, Fallucchi F, Santilli A, Zanzotto FM (2021) KERMITviz: Visualizing Neural

- Network Activations on Syntactic Trees. In: Proc. of the Int. Conf. on Metadata and Semantic Research (MTSR), Revised Selected Papers, Springer, Communications in Computer and Information Science, vol 1537, pp 139–147
- Raney AA (2011) The role of morality in emotional reactions to and enjoyment of media entertainment. *J Media Psychol Theor Methods Appl* 23(1):18–23, DOI 10.1027/1864-1105/A000027, URL <https://doi.org/10.1027/1864-1105/a000027>
- Rao A, Guo S, Wang SN, Morstatter F, Lerman K (2023) Pandemic culture wars: Partisan differences in the moral language of COVID-19 discussions. In: He J, Palpanas T, Hu X, Cuzzocrea A, Dou D, Slezak D, Wang W, Gruca A, Lin JC, Agrawal R (eds) *IEEE International Conference on Big Data, BigData 2023*, Sorrento, Italy, December 15–18, 2023, IEEE, pp 413–422, DOI 10.1109/BIGDATA59044.2023.10386084, URL <https://doi.org/10.1109/BigData59044.2023.10386084>
- Ravichander A, Black AW, Wilson S, Norton TB, Sadeh NM (2019) Question answering for privacy policies: Combining computational and legal perspectives. In: Proc. of the Conf. on Empirical Methods in Natural Language Processing and the Int. Joint Conf. on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, pp 4946–4957
- Rawat BPS, Kovaly S, Pigeon WR, Yu H (2022) Scan: Suicide attempt and ideation events dataset. *CoRR* abs/2205.07872
- Reimers N, Gurevych I (2019) Sentence-bert: Sentence embeddings using siamese bert-networks. In: Inui K, Jiang J, Ng V, Wan X (eds) *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019*, Hong Kong, China, November 3–7, 2019, Association for Computational Linguistics, pp 3980–3990, DOI 10.18653/V1/D19-1410, URL <https://doi.org/10.18653/v1/D19-1410>
- Reinig I, Becker M, Rehbein I, Ponzetto S (2024) A survey on modelling morality for text analysis. In: Ku LW, Martins A, Srikumar V (eds) *Findings of the Association for Computational Linguistics ACL 2024*, Association for Computational Linguistics, Bangkok, Thailand and virtual meeting, pp 4136–4155, DOI 10.18653/v1/2024.findings-acl.245, URL <https://aclanthology.org/2024.findings-acl.245>
- Renner J, Denis P, Gilleron R (2023) Wordnet is all you need: A surprisingly effective unsupervised method for graded lexical entailment. In: *Empirical Methods in Natural Language Processing*
- Ribeiro MT, Singh S, Guestrin C (2016) “Why Should I Trust You?”: Explaining the Predictions of Any Classifier. In: Proc. of the ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, ACM, pp 1135–1144, DOI 10.1145/2939672.2939778
- Rogers A, Kovaleva O, Rumshisky A (2020) A primer in bertology: What we know about how BERT works. *Trans Assoc Comput Linguistics* 8:842–866, DOI 10.1162/TACL_A_00349, URL https://doi.org/10.1162/tacl_a_00349

- Rojecki A, Zheleva E, Levine L (2021) The moral imperatives of self-quarantining. In: Annual meeting of the American Political Science Association
- Rosa GM, Rodrigues RC, de Alencar Lotufo R, Nogueira R (2021) To tune or not to tune?: zero-shot models for legal case entailment. In: Proc. of the Int. Conf. on Artificial Intelligence and Law (ICAIL), ACM, pp 295–300, DOI 10.1145/3462757.3466103
- Rossi J, Kanoulas E (2019) Legal Information Retrieval with Generalized Language Models: ILPS Participation to COLIEE 2019. In: Proc. of COLIEE 2019 workshop: Competition on Legal Information Extraction/Entailment, URL <https://hdl.handle.net/11245.1/8f93b374-736d-4a47-971d-bd28cad3f088>
- Roy S, Goldwasser D (2021) Analysis of nuanced stances and sentiment towards entities of US politicians through the lens of moral foundation theory. In: Ku LW, Li CT (eds) Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media, Association for Computational Linguistics, Online, pp 1–13, DOI 10.18653/v1/2021.socialnlp-1.1, URL <https://aclanthology.org/2021.socialnlp-1.1>
- Rozière B, Gehring J, Gloeckle F, Sootla S, Gat I, Tan XE, Adi Y, Liu J, Sauvestre R, Remez T, Rapin J, Kozhevnikov A, Evtimov I, Bitton J, Bhatt M, Ferrer CC, Grattafiori A, Xiong W, Défossez A, Copet J, Azhar F, Touvron H, Martin L, Usunier N, Scialom T, Synnaeve G (2024) Code llama: Open foundation models for code. 2308.12950
- Sadaf A (2011) Public perception of media role. *International Journal of Humanities & Social Sciences* Volume 1:228–236
- Sainz O, García-Ferrero I, Agerri R, de Lacalle OL, Rigau G, Agirre E (2023a) Golli: Annotation guidelines improve zero-shot information-extraction. arXiv preprint arXiv:231003668
- Sainz O, de Lacalle OL, Agirre E, Rigau G (2023b) What do language models know about word senses? zero-shot WSD with language models and domain inventories. In: Rigau G, Bond F, Rademaker A (eds) Proceedings of the 12th Global Wordnet Conference, GWC 2023, University of the Basque Country, Donostia - San Sebastian, Basque Country, Spain, 23 - 27 January 2023, Global Wordnet Association, pp 331–342, URL <https://aclanthology.org/2023.gwc-1.40>
- Sakata W, Shibata T, Tanaka R, Kurohashi S (2019) FAQ retrieval using query-question similarity and bert-based query-answer relevance. In: Proc. of the ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR), ACM, pp 1113–1116, DOI 10.1145/3331184.3331326
- Salaün O, Langlais P, Lou A, Westermann H, Benyekhlef K (2020) Analysis and Multilabel Classification of Quebec Court Decisions in the Domain of Housing Law. In: Proc. of the Int. Conf. on Applications of Natural Language to Information Systems (NLDB), Springer, Lecture Notes in Computer Science, vol 12089, pp 135–143, DOI 10.1007/978-3-030-51310-8_13

- Samy D, Arenas-García J, Pérez-Fernández D (2020) Legal-ES: A Set of Large Scale Resources for Spanish Legal Text Processing. In: Proc. of the Workshop on Language Technologies for Government and Public Administration (LT4Gov@LREC 2020), European Language Resources Association, pp 32–36
- Sanchez L, He J, Manotumruksa J, Albakour D, Martinez M, Lipani A (2020) Easing legal news monitoring with learning to rank and BERT. In: Proc. of the European Conf. on IR Research (ECIR), Springer, Lecture Notes in Computer Science, vol 12036, pp 336–343, DOI 10.1007/978-3-030-45442-5_42
- Sang EFTK (2002) Introduction to the CoNLL-2002 Shared Task: Language-Independent Named Entity Recognition. In: Proc. of the Conf. on Natural Language Learning (CoNLL), ACL
- Sang EFTK, Meulder FD (2003) Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In: Proc. of the Conf. on Natural Language Learning (CoNLL), ACL, pp 142–147
- Sanh V, Debut L, Chaumond J, Wolf T (2019) Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. CoRR abs/1910.01108, URL <http://arxiv.org/abs/1910.01108>
- Sansone C, Sperlí G (2022) Legal information retrieval systems: State-of-the-art and open issues. *Inf Syst* 106:101967
- Santosh TYSS, Xu S, Ichim O, Grabmair M (2022) Deconfounding legal judgment prediction for european court of human rights cases towards better alignment with experts. In: Proc. of the Conf. on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, pp 1120–1138
- Sarkar R, Ojha AK, Megaro J, Mariano J, Herard V, McCrae JP (2021) Few-shot and Zero-shot Approaches to Legal Text Classification: A Case Study in the Financial Sector. In: Proc. of the 2021 Workshop on Natural Legal Language Processing (NLLP), ACL, pp 102—106
- Savelka J, Ashley KD (2021) Discovering explanatory sentences in legal case decisions using pre-trained language models. In: Findings of the Association for Computational Linguistics: EMNLP 2021, Association for Computational Linguistics, pp 4273–4283
- Savelka J, Ashley KD (2022) Legal information retrieval for understanding statutory terms. *Artif Intell Law* 30(2):245–289
- Savelka J, Westermann H, Benyekhlef K (2020) Cross-domain generalization and knowledge transfer in transformers trained on legal data. In: Proc. of the Fourth Workshop on Automated Semantic Analysis of Information in Legal Text (ASAIL) in conjunction with the Int. Conf. on Legal Knowledge and Information Systems (JURIX), CEUR-WS.org, CEUR Workshop Proceedings, vol 2764
- Savelka J, Westermann H, Benyekhlef K, Alexander CS, Grant JC, Amariles DR, Hamdani RE, Meeüs S, Troussel AC, Araszkiewicz M, Ashley KD, Ashley A,

- Branting K, Falduti M, Grabmair M, Harasta J, Novotná T, Tippet E, Johnson S (2021) Lex Rosetta: transfer of predictive models across languages, jurisdictions, and legal domains. In: Proc. of the Int. Conf. on Artificial Intelligence and Law (ICAIL), ACM, pp 129–138
- Schilder F, Chinnappa D, Madan K, Harmouche J, Vold A, Bretz H, Hudzina J (2021) A Pentapus Grapples with Legal Reasoning. In: Proc. of the Eighth Int. Competition on Legal Information Extraction/Entailment (COLIEE 2021), pp 60–68
- Schoene AM, Lacey G, Turner AP, Dethlefs N (2019) Dilated LSTM with attention for classification of suicide notes. In: Proc. LOUHI@EMNLP, pp 136–145
- Schramowski P, Turan C, Andersen N, Rothkopf CA, Kersting K (2022) Large pre-trained language models contain human-like biases of what is right and wrong to do. 2103.11790
- Schulman J, Wolski F, Dhariwal P, Radford A, Klimov O (2017) Proximal policy optimization algorithms. CoRR abs/1707.06347
- Schwartz SH (1992) Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. *Advances in Experimental Social Psychology*, vol 25, Academic Press, pp 1–65, DOI [https://doi.org/10.1016/S0065-2601\(08\)60281-6](https://doi.org/10.1016/S0065-2601(08)60281-6), URL <https://www.sciencedirect.com/science/article/pii/S0065260108602816>
- See A, Liu PJ, Manning CD (2017) Get to the point: Summarization with pointer-generator networks. In: Proc. ACL, pp 1073–1083
- Seker A, Bandel E, Bareket D, Brusilovsky I, Greenfeld RS, Tsarfaty R (2021) Aleph-BERT: A Hebrew Large Pre-Trained Language Model to Start-off your Hebrew NLP Application With. CoRR abs/2104.04052
- Sennrich R, Haddow B, Birch A (2016) Neural machine translation of rare words with subword units. In: Proc. of the Annual Meeting of the Association for Computational Linguistics (ACL), Association for Computational Linguistics, DOI 10.18653/v1/p16-1162
- Serras FR, Finger M (2022) verBERT: Automating Brazilian Case Law Document Multi-label Categorization Using BERT. CoRR abs/2203.06224
- Shaghaghian S, Feng LY, Jafarpour B, Pogrebnikov N (2020) Customizing contextualized language models for legal document reviews. In: Proc. of the IEEE Int. Conf. on Big Data (BigData), IEEE, pp 2139–2148, DOI 10.1109/BigData50022.2020.9378201
- Shaheen Z, Wohlgenannt G, Filtz E (2020) Large scale legal text classification using transformer models. CoRR abs/2010.12871
- Shaheen Z, Wohlgenannt G, Muromtsev D (2021) Zero-shot cross-lingual transfer in legal domain using transformer models. CoRR abs/2111.14192
- Shao H, Chen Y, Huang S (2020a) BERT-Based Ensemble Model for Statute Law Retrieval and Legal Information Entailment. In: *New Frontiers in Artificial Intelligence*

- JSAI-isAI 2020 Workshops, JURISIN, LENLS 2020 Workshops, Springer, Lecture Notes in Computer Science, vol 12758, pp 226–239
- Shao Y, Liu B, Mao J, Liu Y, Zhang M, Ma S (2020b) Thuir@coliee-2020: Leveraging semantic understanding and exact matching for legal case retrieval and entailment. CoRR abs/2012.13102
- Shao Y, Mao J, Liu Y, Ma W, Satoh K, Zhang M, Ma S (2020c) BERT-PLI: modeling paragraph-level interactions for legal case retrieval. In: Proc. of the Int. Joint Conf. on Artificial Intelligence (IJCAI), pp 3501–3507, DOI 10.24963/ijcai.2020/484
- Sharma E, Li C, Wang L (2019) BIGPATENT: A large-scale dataset for abstractive and coherent summarization. In: Proc. ACL, pp 2204–2213
- Shen Z, Lo K, Yu L, Dahlberg N, Schlanger M, Downey D (2022) Multi-LexSum: Real-world Summaries of Civil Rights Lawsuits at Multiple Granularities. In: Proc. of the Annual Conf. on Neural Information Processing Systems (NeurIPS)
- Shliazhko O, Fenogenova A, Tikhonova M, Mikhailov V, Kozlova A, Shavrina T (2022) mgpt: Few-shot learners go multilingual. CoRR abs/2204.07580
- Shorten C, Khoshgoftaar TM, Furht B (2021) Text data augmentation for deep learning. Journal of big Data 8(1):101
- Shukla A, Bhattacharya P, Poddar S, Mukherjee R, Ghosh K, Goyal P, Ghosh S (2022) Legal case document summarization: Extractive and abstractive methods and their evaluation. In: Proc. of the Conf. of the Asia-Pacific Chapter of the Association for Computational Linguistics and the Int. Joint Conf. on Natural Language Processing (AACL/IJCNLP), Association for Computational Linguistics, pp 1048–1064
- Simeri A, Tagarelli A (2023) Exploring domain and task adaptation of LAMBERTa models for article retrieval on the Italian Civil Code. In: Proc. of the Conf. on Information and Research science Connecting to Digital and Library science (IRCDL), CEUR-WS.org, CEUR Workshop Proceedings, vol 3365, pp 130–143
- Simmons G (2023) Moral mimicry: Large language models produce moral rationalizations tailored to political identity. In: Padmakumar V, Vallejo G, Fu Y (eds) Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop), Association for Computational Linguistics, Toronto, Canada, pp 282–297, DOI 10.18653/v1/2023.acl-srw.40, URL <https://aclanthology.org/2023.acl-srw.40>
- Simonsen KB, Widmann T (2023) The politics of right and wrong: Moral appeals in political communication over six decades in ten western democracies. DOI 10.31219/osf.io/m6qkg, URL osf.io/m6qkg
- Singhal P, Walambe R, Ramanna S, Kotecha K (2023) Domain adaptation: challenges, methods, datasets, and applications. IEEE access 11:6973–7020
- Skaik R, Inkpen D (2021) Using social media for mental health surveillance: A review. ACM Comput Surv 53(6):129:1–129:31
- Slonim N (2022) Project debater - an autonomous debating system. In: Proceedings

- of the International Symposium on Artificial Intelligence and Mathematics 2022 (ISAIM 2022), Fort Lauderdale, Florida, USA, January 3-5, 2022, URL https://isaim2022.cs.ou.edu/papers/ISAIM2022_Tishby_Slonim.pdf
- Socher R, Perelygin A, Wu J, Chuang J, Manning CD, Ng AY, Potts C (2013) Recursive deep models for semantic compositionality over a sentiment treebank. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL, ACL, pp 1631–1642
- Song D, Gao S, He B, Schilder F (2022) On the effectiveness of pre-trained language models for legal natural language processing: An empirical study. *IEEE Access* 10:75835–75858
- Song F, Croft WB (1999) A general language model for information retrieval. In: Proc. of the ACM Conf. on Information and Knowledge Management (CIKM), ACM, pp 316–321
- Song K, Tan X, Qin T, Lu J, Liu T (2019) MASS: masked sequence to sequence pre-training for language generation. In: Proc. of the Int. Conf. on Machine Learning (ICML), pp 5926–5936
- Song K, Tan X, Qin T, Lu J, Liu T (2020) MPNet: Masked and Permuted Pre-training for Language Understanding. In: Proc. of the Annual Conference on Neural Information Processing Systems (NeurIPS 2020)
- Souza F, Nogueira R, de Alencar Lotufo R (2020) BERTimbau: Pretrained BERT Models for Brazilian Portuguese. In: Proc. of the Brazilian Conf. on Intelligent Systems (BRACIS), Springer, Lecture Notes in Computer Science, vol 12319, pp 403–417
- Spartalis C, Drosatos G, Arampatzis A (2021) Transfer Learning for Automated Responses to the BDI Questionnaire. In: Working Notes of CLEF, CEUR Workshop Proceedings, vol 2936, pp 1046–1058
- Spinde T, Hinterreiter S, Haak F, Ruas T, Giese H, Meuschke N, Gipp B (2023) The media bias taxonomy: A systematic literature review on the forms and automated detection of media bias. *arXiv preprint arXiv:231216148*
- Steinberger R, Pouliquen B, Widiger A, Ignat C, Erjavec T, Tufis D, Varga D (2006) The JRC-Acquis: A Multilingual Aligned Parallel Corpus with 20+ Languages. In: Proc. of the Int. Conf. on Language Resources and Evaluation (LREC), European Language Resources Association (ELRA), pp 2142–2147
- Strapparava C, Mihalcea R (2007) Semeval-2007 task 14: Affective text. In: Agirre E, i Villodre LM, Wicentowski R (eds) Proceedings of the 4th International Workshop on Semantic Evaluations, SemEval@ACL 2007, Prague, Czech Republic, June 23-24, 2007, The Association for Computer Linguistics, pp 70–74, URL <https://aclanthology.org/S07-1013/>
- Strickson B, Iglesia BDL (2020) Legal judgement prediction for UK courts. In: Proc.

- of the Int. Conf. on Information Science and Systems (ICISS), ACM, pp 204–209
- Su J (2021) T5 pegasus - zhuiyai. Tech. rep., URL <https://github.com/ZhuiyiTechnology/t5-pegasus>
- Su J, Lu Y, Pan S, Wen B, Liu Y (2021) Roformer: Enhanced transformer with rotary position embedding. CoRR abs/2104.09864
- Sun M, Guo Z, Deng X (2021) Intelligent bert-bilstm-crf based legal case entity recognition method. In: Proc. of the ACM Turing Award Celebration Conf., ACM, pp 186–191
- Sundararajan M, Taly A, Yan Q (2017) Axiomatic attribution for deep networks. In: Proc. of the Int. Conf. on Machine Learning (ICML), PMLR, Proceedings of Machine Learning Research, vol 70, pp 3319–3328
- Surden H (2019) Artificial intelligence and law: An overview. 35 GA ST U L REV 1305
- Tagarelli A, Simeri A (2022) Unsupervised law article mining based on deep pre-trained language representation models with application to the Italian civil code. Artif Intell Law 30(3):417–473, DOI 10.1007/s10506-021-09301-8
- Tang L, Clematide S (2021) Searching for legal documents at paragraph level: Automating label generation and use of an extended attention mask for boosting neural models of semantic similarity. In: Proc. of the 2021 Workshop on Natural Legal Language Processing (NLLP), ACL, pp 114–122
- Tao Y, Viberg O, Baker RS, Kizilcec RF (2024) Cultural bias and cultural alignment of large language models. PNAS Nexus 3(9):pgae346, DOI 10.1093/pnasnexus/pgae346, URL <https://doi.org/10.1093/pnasnexus/pgae346>
- Tarasconi F, Botros M, Caserio M, Sportelli G, Giacalone G, Uttini C, Vignati L, Zanetta F (2020) Natural language processing applications in case-law text publishing. In: Proc. of the Conf. on Legal Knowledge and Information Systems (JURIX), IOS Press, Frontiers in Artificial Intelligence and Applications, vol 334, pp 154–163
- Team G, Mesnard T, Hardin C, Dadashi R, Bhupatiraju S, Pathak S, Sifre L, Rivière M, Kale MS, Love J, et al. (2024) Gemma: Open models based on gemini research and technology. arXiv preprint arXiv:240308295
- Tekin M, Ekici S (2023) The relationship between emotions and morality. Frontiers in Psychology DOI 10.3389/fpsyg.2023.1234567
- Tikhomirov M, Loukachevitch NV (2024) Exploring prompt-based methods for zero-shot hypernym prediction with large language models. CoRR abs/2401.04515, DOI 10.48550/ARXIV.2401.04515, URL <https://doi.org/10.48550/arXiv.2401.04515>
- Touvron H, Lavril T, Izacard G, Martinet X, Lachaux M, Lacroix T, Rozière B, Goyal N, Hambro E, Azhar F, Rodriguez A, Joulin A, Grave E, Lample G (2023a) Llama: Open and efficient foundation language models. CoRR abs/2302.13971, DOI 10.48550/ARXIV.2302.13971, URL <https://doi.org/10.48550/arXiv.2302.13971>

- Touvron H, Martin L, Stone K, Albert P, Almahairi A, Babaei Y, Bashlykov N, Batra S, Bhargava P, Bhosale S, Bikel D, Blecher L, Canton-Ferrer C, Chen M, Cucurull G, Esiobu D, Fernandes J, Fu J, Fu W, Fuller B, Gao C, Goswami V, Goyal N, Hartshorn A, Hosseini S, Hou R, Inan H, Kardas M, Kerkez V, Khabsa M, Kloumann I, Korenev A, Koura PS, Lachaux M, Lavril T, Lee J, Liskovich D, Lu Y, Mao Y, Martinet X, Mihaylov T, Mishra P, Molybog I, Nie Y, Poulton A, Reizenstein J, Rungta R, Saladi K, Schelten A, Silva R, Smith EM, Subramanian R, Tan XE, Tang B, Taylor R, Williams A, Kuan JX, Xu P, Yan Z, Zarov I, Zhang Y, Fan A, Kambadur M, Narang S, Rodriguez A, Stojnic R, Edunov S, Scialom T (2023b) Llama 2: Open foundation and fine-tuned chat models. CoRR abs/2307.09288
- Trager J, Ziabari AS, Davani AM, Golazizian P, Karimi-Malekabadi F, Omrani A, Li Z, Kennedy B, Reimer NK, Reyes M, Cheng K, Wei M, Merrifield C, Khosravi A, Alvarez E, Dehghani M (2022) The moral foundations reddit corpus. CoRR abs/2208.05545, DOI 10.48550/ARXIV.2208.05545, URL <https://doi.org/10.48550/arXiv.2208.05545>
- Trautmann D, Petrova A, Schilder F (2022) Legal prompt engineering for multilingual legal judgement prediction. CoRR abs/2212.02199
- Trias F, Wang H, Jaume S, Idreos S (2021) Named entity recognition in historic legal text: A transformer and state machine ensemble method. In: Proc. of the 2021 Workshop on Natural Legal Language Processing (NLLP), ACL, pp 172–179
- Triguero I, García S, Herrera F (2015) Self-labeled techniques for semi-supervised learning: taxonomy, software and empirical study. *Knowl Inf Syst* 42(2):245–284, DOI 10.1007/s10115-013-0706-y
- Trischler A, Wang T, Yuan X, Harris J, Sordoni A, Bachman P, Suleman K (2017) NewsQA: A Machine Comprehension Dataset. In: Proc. of the Workshop on Representation Learning for NLP, Rep4NLP@ACL 2017, Association for Computational Linguistics, pp 191–200, DOI 10.18653/v1/w17-2623
- Trivedi H, Kwon H, Khot T, Sabharwal A, Balasubramanian N (2019) Repurposing entailment for multi-hop question answering tasks. In: Proc. of the Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), Association for Computational Linguistics, pp 2948–2958
- Tsarapatsanis D, Aletras N (2021) On the ethical limits of natural language processing on legal text. In: Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Association for Computational Linguistics, pp 3590–3599, DOI 10.18653/v1/2021.findings-acl.314
- Tseng YH, Ku MC, Chen WL, Chang YL, Hsieh SK (2023) Vec2Gloss: definition modeling leveraging contextualized vectors with Wordnet gloss. In: Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation, Association for Computational Linguistics, Hong Kong, China, pp 679–690, URL <https://aclanthology.org/2023.paclic-1.68>

- Tuggener D, von Däniken P, Peetz T, Cieliebak M (2020) LEDGAR: A large-scale multi-label corpus for text classification of legal provisions in contracts. In: Proc. of the Int. Conf. on Language Resources and Evaluation (LREC), European Language Resources Association, pp 1235–1241
- Tziafas G, de Saint-Phalle E, de Vries W, Egger C, Caselli T (2021) A Multilingual Approach to Identify and Classify Exceptional Measures Against COVID-19. In: Proc. of the 2021 Workshop on Natural Legal Language Processing (NLLP), ACL, pp 46–62
- Uban AS, Rosso P (2020) Deep Learning Architectures and Strategies for Early Detection of Self-harm and Depression Level Prediction. In: Working Notes of CLEF
- Ugazio G, Lamm C, Singer T (2012) The role of emotions for moral judgments depends on the type of emotion and moral scenario. *Emotion* 12(3):579
- Urchs S, Mitrovic J, Granitzer M (2021) Design and implementation of german legal decision corpora. In: Proc. of the Int. Conf. on Agents and Artificial Intelligence (ICAART), pp 515–521
- Vajre V, Naylor M, Kamath U, Shehu A (2021) PsychBERT: A Mental Health Language Model for Social Media Mental Health Behavioral Analysis. In: Proc. IEEE BIBM, pp 1077–1082
- Vasilyev OV, Dharnidharka V, Bohannon J (2020) Fill in the BLANC: human-free quality estimation of document summaries. CoRR abs/2002.09836
- Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I (2017) Attention is all you need. In: Guyon I, von Luxburg U, Bengio S, Wallach HM, Fergus R, Vishwanathan SVN, Garnett R (eds) *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4–9, 2017, Long Beach, CA, USA*, pp 5998–6008, URL <https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html>
- Verma S, Nidhi V (2018) Extractive summarization using deep learning. *Res Comput Sci* 147(10):107–117
- Vig J (2019) A multiscale visualization of attention in the transformer model. In: *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL): System Demonstrations*
- Vincent P, Larochelle H, Bengio Y, Manzagol PA (2008) Extracting and composing robust features with denoising autoencoders. In: *Proceedings of the 25th International Conference on Machine Learning, Association for Computing Machinery, New York, NY, USA, ICML '08*, p 1096–1103, DOI 10.1145/1390156.1390294, URL <https://doi.org/10.1145/1390156.1390294>
- Vold A, Conrad JG (2021) Using Transformers to improve answer retrieval for legal questions. In: Proc. of the Int. Conf. on Artificial Intelligence and Law (ICAAIL), ACM, pp 245–249, DOI 10.1145/3462757.3466102

- de Vries W, van Cranenburgh A, Bisazza A, Caselli T, van Noord G, Nissim M (2019) Bertje: A dutch BERT model. CoRR abs/1912.09582
- Vulić I, Gerz D, Kiela D, Hill F, Korhonen A (2017) HyperLex: A Large-Scale Evaluation of Graded Lexical Entailment. *Computational Linguistics* 43(4):781–835, DOI 10.1162/COLI.a.00301
- Vuong T, Bui QM, Nguyen H, Nguyen T, Tran V, Phan X, Satoh K, Le NM (2023) SM-BERT-CR: a deep learning approach for case law retrieval with supporting model. *Artif Intell Law* 31(3):601–628, DOI 10.1007/s10506-022-09319-6
- Wachsmuth H, Naderi N, Hou Y, Bilu Y, Prabhakaran V, Thijm TA, Hirst G, Stein B (2017) Computational argumentation quality assessment in natural language. In: Lapata M, Blunsom P, Koller A (eds) *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 1: Long Papers*, Association for Computational Linguistics, pp 176–187, DOI 10.18653/V1/E17-1017, URL <https://doi.org/10.18653/v1/e17-1017>
- Walker VR, Pillaipakkamnatt K, Davidson AM, Linares M, Pesce DJ (2019) Automatic classification of rhetorical roles for sentences: Comparing rule-based scripts with machine learning. In: *Proc. of the Third Workshop on Automated Semantic Analysis of Information in Legal Texts, co-located with the Int. Conf. on Artificial Intelligence and Law (ICAIL)*, CEUR-WS.org, CEUR Workshop Proceedings, vol 2385
- Wang A, Singh A, Michael J, Hill F, Levy O, Bowman SR (2018) GLUE: A multi-task benchmark and analysis platform for natural language understanding. In: Linzen T, Chrupala G, Alishahi A (eds) *Proceedings of the Workshop: Analyzing and Interpreting Neural Networks for NLP, BlackboxNLP@EMNLP 2018*, Brussels, Belgium, November 1, 2018, Association for Computational Linguistics, pp 353–355, DOI 10.18653/V1/W18-5446, URL <https://doi.org/10.18653/v1/w18-5446>
- Wang A, Pruksachatkun Y, Nangia N, Singh A, Michael J, Hill F, Levy O, Bowman SR (2019) Superglue: A stickier benchmark for general-purpose language understanding systems. In: Wallach HM, Larochelle H, Beygelzimer A, d'Alché-Buc F, Fox EB, Garnett R (eds) *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp 3261–3275, URL <https://proceedings.neurips.cc/paper/2019/hash/4496bf24afe7fab6f046bf4923da8de6-Abstract.html>
- Wang K, Reimers N, Gurevych I (2021a) Tsdae: Using transformer-based sequential denoising auto-encoder for unsupervised sentence embedding learning. In: *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp 671–688
- Wang Q, Zhao K, Amor R, Liu B, Wang R (2022) D2GCLF: document-to-graph classifier for legal document classification. In: *Findings of the Association for Computational Linguistics: NAACL, Association for Computational Linguistics*, pp 2208–2221

- Wang S, Khabsa M, Ma H (2020a) To pretrain or not to pretrain: Examining the benefits of pretraining on resource rich tasks. In: Proc. of the Annual Meeting of the Association for Computational Linguistics (ACL), Association for Computational Linguistics, pp 2209–2213
- Wang SH, Scardigli A, Tang L, Chen W, Levkin D, Chen A, Ball S, Woodside T, Zhang O, Hendrycks D (2023a) MAUD: an expert-annotated legal NLP dataset for merger agreement understanding. CoRR abs/2301.00876
- Wang W, Wei F, Dong L, Bao H, Yang N, Zhou M (2020b) MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers. In: Proc. of the Annual Conf. on Neural Information Processing Systems (NeurIPS)
- Wang X, Zhou W, Zu C, Xia H, Chen T, Zhang Y, Zheng R, Ye J, Zhang Q, Gui T, Kang J, Yang J, Li S, Du C (2023b) Instructuie: Multi-task instruction tuning for unified information extraction. 2304.08085
- Wang Y, Xiao C, Ma S, Zhong H, Tu C, Zhang T, Liu Z, Sun M (2021b) Equality before the law: Legal judgment consistency analysis for fairness. CoRR abs/2103.13868
- Wang Y, Zhong W, Li L, Mi F, Zeng X, Huang W, Shang L, Jiang X, Liu Q (2023c) Aligning large language models with human: A survey. CoRR abs/2307.12966, DOI 10.48550/ARXIV.2307.12966, URL <https://doi.org/10.48550/arXiv.2307.12966>
- Weber R, Mangus JM, Huskey R, Hopp FR, Amir O, Swanson R, Gordon A, Khooshabeh P, Hahn L, Tamborini R (2021) Extracting latent moral information from text narratives: Relevance, challenges, and solutions. In: Computational Methods for Communication Science, Routledge, pp 39–59
- Wehnert S, Sudhi V, Dureja S, Kutty L, Shahania S, Luca EWD (2021) Legal norm retrieval with variations of the BERT model combined with TF-IDF vectorization. In: Proc. of the Int. Conf. on Artificial Intelligence and Law (ICAIL), ACM, DOI 10.1145/3462757.3466104
- Wehnert S, Dureja S, Kutty L, Sudhi V, Luca EWD (2022) Applying BERT embeddings to predict legal textual entailment. Rev Socionetwork Strateg 16(1):197–219
- Welbl J, Stenetorp P, Riedel S (2018) Constructing datasets for multi-hop reading comprehension across documents. Trans Assoc Comput Linguistics 6:287–302
- Westermann H, Savelka J, Benyekhlef K (2020) Paragraph similarity scoring and fine-tuned BERT for legal information retrieval and entailment. In: New Frontiers in Artificial Intelligence - JSAI-isAI 2020 Workshops, JURISIN, LENLS 2020 Workshops, Springer, Lecture Notes in Computer Science, vol 12758, pp 269–285
- Williams A, Nangia N, Bowman SR (2018) A broad-coverage challenge corpus for sentence understanding through inference. In: Walker MA, Ji H, Stent A (eds) Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers), Association

- for Computational Linguistics, pp 1112–1122, DOI 10.18653/V1/N18-1101, URL <https://doi.org/10.18653/v1/n18-1101>
- Wilson S, Schaub F, Dara AA, Liu F, Cherivirala S, Leon PG, Andersen MS, Zimmeck S, Sathyendra KM, Russell NC, Norton TB, Hovy EH, Reidenberg JR, Sadeh NM (2016) The creation and analysis of a website privacy policy corpus. In: Proc. of the Annual Meeting of the Association for Computational Linguistics (ACL), Association for Computer Linguistics
- Wrzalik M, Krechel D (2021) GerDaLIR: A German Dataset for Legal Information Retrieval. In: Proc. of the 2021 Workshop on Natural Legal Language Processing (NLLP), ACL, pp 123–128
- Wu SH, Qiu Z (2021) A RoBERTa-based model on measuring the severity of the signs of depression. In: Working Notes of CLEF, CEUR Workshop Proceedings, vol 2936, pp 1071–1080
- Wu Y, Schuster M, Chen Z, Le QV, Norouzi M, Macherey W, Krikun M, Cao Y, Gao Q, Macherey K, Klingner J, Shah A, Johnson M, Liu X, Kaiser L, Gouws S, Kato Y, Kudo T, Kazawa H, Stevens K, Kurian G, Patil N, Wang W, Young C, Smith J, Riesa J, Rudnick A, Vinyals O, Corrado G, Hughes M, Dean J (2016) Google’s Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. CoRR abs/1609.08144, 1609.08144
- Xiang W, Wang B (2019) A survey of event extraction from text. IEEE Access 7:173111–173137
- Xiao C, Zhong H, Guo Z, Tu C, Liu Z, Sun M, Feng Y, Han X, Hu Z, Wang H, Xu J (2018) CAIL2018: A large-scale legal dataset for judgment prediction. CoRR abs/1807.02478
- Xiao C, Zhong H, Guo Z, Tu C, Liu Z, Sun M, Zhang T, Han X, Hu Z, Wang H, Xu J (2019) CAIL2019-SCM: A dataset of similar case matching in legal domain. CoRR abs/1911.08962
- Xiao C, Hu X, Liu Z, Tu C, Sun M (2021) Lawformer: A Pre-trained Language Model for Chinese Legal Long Documents. AI Open 2:79–84, DOI 10.1016/j.aiopen.2021.06.003
- Xie JY, Hirst G, Xu Y (2020) Contextualized moral inference. CoRR abs/2008.10762, URL <https://arxiv.org/abs/2008.10762>, 2008.10762
- Xiong C, Callan J, Liu T (2017) Word-entity duet representations for document ranking. In: Proc. of the ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR), ACM, pp 763–772
- Xiong M, Hu Z, Lu X, Li Y, Fu J, He J, Hooi B (2024) Can LLMs express their uncertainty? an empirical evaluation of confidence elicitation in LLMs. In: The Twelfth International Conference on Learning Representations, URL <https://openreview.net/forum?id=gjeQKFxFpZ>
- Xue L, Constant N, Roberts A, Kale M, Al-Rfou R, Siddhant A, Barua A, Raffel

- C (2021) mt5: A massively multilingual pre-trained text-to-text transformer. In: Toutanova K, Rumshisky A, Zettlemoyer L, Hakkani-Tür D, Beltagy I, Bethard S, Cotterell R, Chakraborty T, Zhou Y (eds) *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021*, Online, June 6-11, 2021, Association for Computational Linguistics, pp 483–498, DOI 10.18653/V1/2021.NAACL-MAI N.41, URL <https://doi.org/10.18653/v1/2021.naacl-main.41>
- Yamada H, Tokunaga T (2019) A performance study on fine-tuned large language models in the Legal Case Entailment Task. In: *Proc. of COLIEE 2019 workshop: Competition on Legal Information Extraction/Entailment*, URL https://www.cl.c.titech.ac.jp/tokunaga/_media/publication/yamada_2019aa.pdf
- Yang J, Jin H, Tang R, Han X, Feng Q, Jiang H, Zhong S, Yin B, Hu XB (2024) Harnessing the power of llms in practice: A survey on chatgpt and beyond. *ACM Trans Knowl Discov Data* 18(6):160:1–160:32, DOI 10.1145/3649506, URL <https://doi.org/10.1145/3649506>
- Yang Z, Yang D, Dyer C, He X, Smola AJ, Hovy EH (2016) Hierarchical attention networks for document classification. In: *Proc. of the Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, Association for Computational Linguistics, pp 1480–1489, DOI 10.18653/v1/n16-1174
- Yang Z, Qi P, Zhang S, Bengio Y, Cohen WW, Salakhutdinov R, Manning CD (2018) HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. In: *Proc. of the Conf. on Empirical Methods in Natural Language Processing (EMNLP)*, Association for Computational Linguistics, pp 2369–2380, DOI 10.18653/v1/d18-1259
- Yang Z, Dai Z, Yang Y, Carbonell JG, Salakhutdinov R, Le QV (2019) Xlnet: Generalized autoregressive pretraining for language understanding. In: Wallach HM, Larochelle H, Beygelzimer A, d'Alché-Buc F, Fox EB, Garnett R (eds) *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019*, December 8-14, 2019, Vancouver, BC, Canada, pp 5754–5764, URL <https://proceedings.neurips.cc/paper/2019/hash/dc6a7e655d7e5840e66733e9ee67cc69-Abstract.html>
- Yoon J, Junaid M, Ali S, Lee J (2022) Abstractive summarization of korean legal cases using pre-trained language models. In: *Proc. of the Int. Conf. on Ubiquitous Information Management and Communication (IMCOM)*, IEEE, pp 1–7
- Yoshioka M, Aoki Y, Suzuki Y (2021) Bert-based ensemble methods with data augmentation for legal textual entailment in COLIEE statute law task. In: Maranhão J, Wyner AZ (eds) *ICAIL '21: Eighteenth International Conference for Artificial Intelligence and Law*, São Paulo Brazil, June 21 - 25, 2021, ACM, pp 278–284, DOI 10.1145/3462757.3466105, URL <https://doi.org/10.1145/3462757.3466105>
- Yu F, Quartey L, Schilder F (2022a) Legal prompting: Teaching a language model to

- think like a lawyer. CoRR abs/2212.01326
- Yu W, Sun Z, Xu J, Dong Z, Chen X, Xu H, Wen J (2022b) Explainable legal case matching via inverse optimal transport-based rationale extraction. In: Proc. of the ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR), ACM, pp 657–668
- Yun S, Jeong M, Kim R, Kang J, Kim HJ (2019) Graph transformer networks. In: Proc. of the Annual Conf. on Neural Information Processing Systems (NeurIPS), pp 11960–11970
- Zaheer M, Guruganesh G, Dubey KA, Ainslie J, Alberti C, Ontañón S, Pham P, Ravula A, Wang Q, Yang L, Ahmed A (2020) Big bird: Transformers for longer sequences. In: Larochelle H, Ranzato M, Hadsell R, Balcan M, Lin H (eds) Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6–12, 2020, virtual, URL <https://proceedings.neurips.cc/paper/2020/hash/c8512d142a2d849725f31a9a7a361ab9-Abstract.html>
- Zangari L, Greco CM, Picca D, Tagarelli A (2024) A survey on moral foundation theory and pre-trained language models: Current advances and challenges. CoRR abs/2409.13521, DOI 10.48550/ARXIV.2409.13521, URL <https://doi.org/10.48550/arXiv.2409.13521>, 2409.13521
- Zangari L, Greco CM, Picca D, Tagarelli A (2025) ME²-BERT: Are Events and Emotions what you need for Moral Foundation Prediction? In: Proceedings of the 31st International Conference on Computational Linguistics (COLING), Association for Computational Linguistics
- Zanzotto FM, Santilli A, Ranaldi L, Onorati D, Tommasino P, Fallucchi F (2020) KERMIT: complementing transformer architectures with encoders of explicit syntactic interpretations. In: Proc. of the Conf. on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, pp 256–267
- Zeberga K, Attique M, Shah B, Ali F, Jembre YZ, Chung TS (2022) A novel text mining approach for mental health prediction using bi-lstm and bert model. Computational Intelligence and Neuroscience 2022
- Zeiler MD, Fergus R (2014) Visualizing and understanding convolutional networks. In: Proc. of the European Conf. on Computer Vision (ECCV), Springer, Lecture Notes in Computer Science, vol 8689, pp 818–833
- Zellers R, Bisk Y, Schwartz R, Choi Y (2018) SWAG: A large-scale adversarial dataset for grounded commonsense inference. In: Riloff E, Chiang D, Hockenmaier J, Tsujii J (eds) Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018, Association for Computational Linguistics, pp 93–104, DOI 10.18653/V1/D18-1009, URL <https://doi.org/10.18653/v1/d18-1009>
- Zhang B, Xiong D, Su J (2018) Accelerating neural transformer via an average attention network. In: Proc. of the Annual Meeting of the Association for Computational

- Linguistics (ACL), Association for Computational Linguistics, pp 1789–1798
- Zhang S, Roller S, Goyal N, Artetxe M, Chen M, Chen S, Dewan C, Diab MT, Li X, Lin XV, Mihaylov T, Ott M, Shleifer S, Shuster K, Simig D, Koura PS, Sridhar A, Wang T, Zettlemoyer L (2022) OPT: open pre-trained transformer language models. CoRR abs/2205.01068, DOI 10.48550/ARXIV.2205.01068, URL <https://doi.org/10.48550/arXiv.2205.01068>, 2205.01068
- Zhang T, Schoene A, Ananiadou S (2021a) Automatic identification of suicide notes with a transformer-based deep learning model. *Internet Interventions* 25:100422
- Zhang X, Chen M, May J (2021b) Salience-aware event chain modeling for narrative understanding. In: Moens MF, Huang X, Specia L, Yih SWt (eds) *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, pp 1418–1428, DOI 10.18653/v1/2021.emnlp-main.107, URL <https://aclanthology.org/2021.emnlp-main.107>
- Zhang XF, Wu W, Beauchamp N, Wang L (2024) MOKA: Moral knowledge augmentation for moral event extraction. In: Duh K, Gomez H, Bethard S (eds) *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Association for Computational Linguistics, Mexico City, Mexico, pp 4481–4502, DOI 10.18653/v1/2024.naacl-long.252, URL <https://aclanthology.org/2024.naacl-long.252>
- Zhang Y, Wu J, Yu F, Xu L (2023) Moral Judgments of Human vs. AI Agents in Moral Dilemmas. *Behavioral Sciences* 13(2):181, DOI 10.3390/bs13020181, URL <https://www.mdpi.com/2076-328X/13/2/181>
- Zheng L, Guha N, Anderson BR, Henderson P, Ho DE (2021) When does pretraining help?: assessing self-supervised learning for law and the CaseHOLD dataset of 53,000+ legal holdings. In: *Proc. of the Int. Conf. on Artificial Intelligence and Law (ICAIL)*, ACM, pp 159–168, DOI 10.1145/3462757.3466088
- Zhong H, Guo Z, Tu C, Xiao C, Liu Z, Sun M (2018) Legal judgment prediction via topological learning. In: *Proc. of the Conf. on Empirical Methods in Natural Language Processing (EMNLP)*, Association for Computational Linguistics, pp 3540–3549, DOI 10.18653/v1/d18-1390
- Zhong H, Zhang Z, Liu Z, Sun M (2019a) Open chinese language pre-trained model zoo. Tech. rep., URL <https://github.com/thunlp/openclap>
- Zhong H, Xiao C, Tu C, Zhang T, Liu Z, Sun M (2020) How does NLP benefit legal system: A summary of legal artificial intelligence. In: *Proc. of the Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, pp 5218–5230, DOI 10.18653/v1/2020.acl-main.466
- Zhong L, Zhong Z, Zhao Z, Wang S, Ashley KD, Grabmair M (2019b) Automatic summarization of legal decisions using iterative masking of predictive sentences. In: *Proc. of the Int. Conf. on Artificial Intelligence and Law (ICAIL)*, ACM, pp 163–172

- Ziems C, Yu JA, Wang Y, Halevy AY, Yang D (2022) The moral integrity corpus: A benchmark for ethical dialogue systems. In: Muresan S, Nakov P, Villavicencio A (eds) Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022, Association for Computational Linguistics, pp 3755–3773, DOI 10.18653/V1/2022.ACL-LONG.261, URL <https://doi.org/10.18653/v1/2022.acl-long.261>
- Zimmeck S, Story P, Smullen D, Ravichander A, Wang Z, Reidenberg JR, Russell NC, Sadeh NM (2019) MAPS: scaling privacy compliance analysis to a million apps. Proceedings on Privacy Enhancing Technologies 2019(3):66–86
- Zlatintsi A, Koutras P, Evangelopoulos G, Malandrakis N, Efthymiou N, Pastra K, Potamianos A, Maragos P (2017) Cognimuse: a multimodal video database annotated with saliency, events, semantics and emotion with application to summarization. EURASIP Journal on Image and Video Processing 2017(1):1–24
- Zmiycharov V, Chechev M, Lazarova GA, Tsonkov T, Koychev I (2021) A comparative study on abstractive and extractive approaches in summarization of european legislation documents. In: Angelova G, Kunilovskaya M, Mitkov R, Nikolova-Koleva I (eds) Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021), Held Online, 1-3 September, 2021, INCOMA Ltd., pp 1645–1651

Appendix A

A.1 RQ1 best matchings

For each macro-topic, we report the best-matching pair of topic descriptions: on the bottom, the description generated by GPT-3.5T, on the top, the one listed in ConstituteProject; the prefix in each description corresponds to the topic name.

Macro-topic M1:

Prohibition of slavery: Forbids the practice of owning or otherwise exploiting human beings. May still allow for compulsory labor as a punishment for a crime, and/or compulsory service in the military in time of war.

Prohibition of slavery and forced labor: Constitutions may include provisions that explicitly prohibit slavery, servitude, or forced labor. This protects individuals from being subjected to any form of forced labor or exploitation.

Macro-topic M2:

Right to just remuneration: Grants individuals the right to just remuneration or fair/equal payment for work. May also regulate wages including the establishment of a minimum or living wage.

Right to fair remuneration: This argument ensures that individuals have the right to fair remuneration for their work, including equal pay for equal work. It protects individuals from unfair wage practices and ensures that they receive a decent income to support themselves and their families.

Macro-topic M3:

Provisions for intellectual property: Gives individuals and businesses the exclusive right to protect and profit from their creative works.

Right to cultural and intellectual property: Some constitutions recognize the right to protect and benefit from one's cultural and intellectual creations, promoting innovation, creativity, and economic incentives for artists, inventors, and creators.

Macro-topic M4:

Duty to pay taxes: Requires all citizens to contribute to the expenses of the state through taxes. This is usually a general statement and may still allow for variation on the basis of income.

Duty to pay taxes: Citizens have a duty to pay taxes to contribute to the financing of public services and the functioning of the state. This duty is essential for the provision of public goods and services.

Macro-topic M5:

Duty to obey the constitution: Requires citizens or public officers to obey the constitution. This primarily applies to individuals in cases where the constitution includes a statement of the duties of citizens or holders of public office.

Maintenance of public order and security: Constitutions typically impose a duty on the state to maintain public order and security within its territory. This duty includes protecting citizens from crime, ensuring the functioning of the justice system, and maintaining law and order.

Macro-topic M6:

Right to privacy: Grants individuals protection against government intrusion into their property, papers, person, or communications without proper authorization. Also protects against such intrusions by private individuals.

Right to privacy: This right protects individuals' privacy, including their personal and family life, home, and correspondence, from arbitrary interference by the state or other individuals

Macro-topic M7:

Right to speedy trial: Grants individuals charged with an offense the right to be brought to trial without any undue delay.

Right to speedy trial: This right ensures that individuals are not subjected to undue delays in the legal process. It aims to prevent prolonged pre-trial detention and to ensure that justice is delivered in a timely manner.

Macro-topic M8:

Ombudsman: A public official tasked with investigating improper or illegal actions on the part of government agencies which may infringe on the rights of individuals. Enables individuals to bring complaints about the actions of the government.

Constitutional ombudsman: Some European countries have constitutional ombudsmen or similar institutions that are responsible for protecting citizens' rights and overseeing the enforcement of the constitution. These independent bodies investigate complaints, mediate disputes, and promote good governance.

Macro-topic M9:

Indigenous right to representation: Grants indigenous peoples the right to representation in the central government. This might, for example, involve a seat quota in the legislature.

Representation and participation: Minority rights include the right to representation and participation in decision-making processes at various levels of government. This can involve reserved seats or quotas to ensure minority groups have a voice in legislative bodies or other institutions.

A.2 RQ2 performances

Tables A.1-A.2 show the summarization performance for all topics, comparing the generated summaries with the original texts or the reference summaries.

Table A.1: Answering RQ2 – Summarization performance for all topics: generated summaries vs. input source texts.

Topic id	#C	CR (%)	Rouge-1	Rouge-2	Rouge-L	Rouge-LSum	N (%)	J	D	SBERT	TF-IDF	BH	BT
Physical Integrity Rights													
slave	29	39.723	0.406	0.329	0.316	0.382	5.65	0.63	0.773	0.8	0.811	0.266	0.230
cruelty	35	50.773	0.536	0.463	0.458	0.519	6.452	0.747	0.855	0.88	0.875	0.371	0.313
cappun	32	31.905	0.328	0.245	0.247	0.289	5.172	0.679	0.809	0.821	0.694	0.407	0.220
life	34	32.498	0.326	0.258	0.255	0.303	4.098	0.527	0.69	0.858	0.887	0.331	0.241
torture	34	43.875	0.445	0.362	0.402	0.415	8.333	0.801	0.89	0.933	0.823	0.456	0.265
corppun	5	76.365	0.718	0.57	0.626	0.632	8.772	0.788	0.881	0.971	0.846	0.511	0.426
Social Rights													
edcomp	31	40.039	0.409	0.313	0.357	0.386	5.357	0.627	0.771	0.916	0.876	0.323	0.263
env	38	34.129	0.344	0.285	0.29	0.323	3.777	0.495	0.662	0.866	0.932	0.124	0.205
finsup3	31	31.333	0.308	0.244	0.267	0.288	8.602	0.426	0.598	0.891	0.842	0.181	0.189
safework	17	53.275	0.527	0.42	0.49	0.49	6.504	0.685	0.813	0.915	0.883	0.323	0.279
provwork	32	64.184	0.647	0.49	0.519	0.592	7.692	0.683	0.811	0.904	0.897	0.255	0.307
conright	12	69.31	0.692	0.514	0.621	0.621	4.587	0.8	0.889	0.957	0.892	0.416	0.348
childwrk	17	48.122	0.495	0.397	0.419	0.46	7.843	0.584	0.737	0.956	0.876	0.309	0.307
standliv	20	117.418	0.773	0.579	0.617	0.725	5.288	0.801	0.889	0.938	0.908	0.255	0.375
edfree	39	48.11	0.491	0.395	0.432	0.466	11.5	0.577	0.731	0.943	0.939	0.245	0.266
finsup1	26	68.413	0.715	0.593	0.669	0.673	6.464	0.734	0.847	0.949	0.962	0.242	0.380
finsup4	24	53.355	0.556	0.441	0.502	0.516	6.186	0.63	0.773	0.804	0.906	0.222	0.323
remuner	22	69.758	0.702	0.562	0.63	0.648	7.742	0.694	0.819	0.951	0.95	0.256	0.303
shelter	16	58.365	0.597	0.456	0.461	0.557	7.282	0.707	0.829	0.839	0.884	0.279	0.328
strike	28	45.619	0.454	0.363	0.416	0.418	9.489	0.593	0.745	0.961	0.935	0.329	0.302
health	33	60.292	0.66	0.587	0.614	0.642	3.361	0.689	0.816	0.896	0.975	0.278	0.344
achighed	20	68.056	0.711	0.584	0.657	0.674	4.118	0.815	0.898	0.972	0.954	0.317	0.325
leisure	22	39.768	0.414	0.331	0.363	0.39	6.122	0.535	0.697	0.928	0.877	0.349	0.280
jointrde	38	19.786	0.191	0.156	0.143	0.187	4.412	0.355	0.524	0.775	0.9	0.284	0.145
finsup2	27	57.268	0.599	0.475	0.534	0.549	5.021	0.67	0.802	0.874	0.925	0.257	0.305
water	3	91.043	0.829	0.716	0.774	0.774	10	0.844	0.915	0.949	0.873	0.522	0.493
scifree	5	64.609	0.52	0.32	0.439	0.415	27.273	0.457	0.627	0.909	0.791	0.321	0.262
Economic Rights													
busines	25	60.023	0.632	0.518	0.582	0.591	6.303	0.719	0.837	0.873	0.923	0.333	0.387
intprop	23	57.029	0.592	0.521	0.561	0.58	2.597	0.771	0.871	0.813	0.878	0.113	0.258
transfer	23	57.566	0.593	0.467	0.536	0.561	5	0.748	0.856	0.959	0.9	0.316	0.338
occupate	33	38.111	0.371	0.283	0.301	0.319	6.25	0.56	0.718	0.957	0.858	0.363	0.248
proprght	40	28.843	0.285	0.223	0.256	0.263	4.785	0.426	0.598	0.687	0.924	0.129	0.172
exprop	42	33.33	0.336	0.285	0.315	0.32	1.976	0.493	0.66	0.806	0.945	0.158	0.203

Continued on the next page

Topic id	#C	CR (%)	Rouge-1	Rouge-2	Rouge-L	Rouge-LSum	N (%)	J	D	SBERT	TF-IDF	BH	BT
freecomp	21	66.533	0.671	0.527	0.608	0.627	6.422	0.747	0.855	0.952	0.889	0.272	0.399
Citizen Duties													
milserv	28	62.349	0.648	0.517	0.565	0.602	9.705	0.679	0.809	0.95	0.955	0.254	0.387
taxes	18	65.903	0.705	0.58	0.648	0.682	4.464	0.77	0.87	0.967	0.936	0.405	0.377
work	5	67.43	0.652	0.475	0.552	0.557	10.294	0.693	0.819	0.85	0.84	0.355	0.290
General Duties													
abide	24	62.595	0.667	0.548	0.619	0.616	5.426	0.691	0.817	0.912	0.941	0.208	0.352
binding	8	79.368	0.8	0.603	0.743	0.746	4.167	0.767	0.868	0.964	0.94	0.422	0.400
Civil and Political Rights													
assoc	41	39.191	0.4	0.326	0.356	0.383	7.054	0.501	0.668	0.896	0.956	0.215	0.258
express	42	40.061	0.413	0.346	0.377	0.388	5.276	0.605	0.754	0.848	0.945	0.106	0.216
voteun	34	64.709	0.699	0.633	0.673	0.68	2.198	0.802	0.89	0.87	0.959	0.14	0.387
assem	43	37.243	0.378	0.318	0.347	0.358	8.854	0.527	0.69	0.872	0.922	0.229	0.251
fndfam	25	64.381	0.691	0.606	0.661	0.67	6.329	0.783	0.878	0.961	0.931	0.362	0.425
freemove	35	26.691	0.256	0.209	0.228	0.245	6.283	0.477	0.646	0.856	0.882	0.261	0.262
libel	26	29.107	0.293	0.236	0.262	0.27	3.191	0.345	0.513	0.739	0.83	0.278	0.215
nomil	23	54.219	0.571	0.47	0.51	0.543	6.154	0.667	0.8	0.933	0.951	0.313	0.328
civmar	5	87.967	0.692	0.509	0.587	0.64	15.625	0.643	0.783	0.95	0.791	0.31	0.345
freerel	43	11.13	0.11	0.087	0.092	0.104	3.306	0.251	0.401	0.864	0.821	0.176	0.117
arms	1	45.528	0.326	0	0.326	0.326	50	0.25	0.4	0.729	0.777	0.083	0.333
citren	12	41.749	0.388	0.319	0.335	0.36	9.524	0.633	0.776	0.957	0.868	0.5	0.431
dignity	35	62.55	0.665	0.555	0.619	0.619	4.432	0.7	0.823	0.923	0.975	0.173	0.304
marriage	24	56.742	0.623	0.543	0.59	0.609	3.311	0.756	0.861	0.903	0.957	0.345	0.345
childpro	32	50.087	0.516	0.402	0.449	0.479	6.061	0.661	0.796	0.753	0.93	0.19	0.266
acfree	28	72.705	0.777	0.686	0.754	0.754	4.211	0.781	0.877	0.931	0.968	0.285	0.370
infoacc	28	46.45	0.476	0.385	0.429	0.449	8.156	0.556	0.714	0.921	0.939	0.144	0.226
press	38	30.197	0.317	0.283	0.297	0.307	2.915	0.432	0.603	0.998	0.91	0.148	0.204
debtors	3	52.941	0.406	0.213	0.308	0.308	29.412	0.5	0.667	0.909	0.605	0.162	0.189
devlpers	13	61.121	0.64	0.514	0.61	0.598	3.185	0.764	0.866	0.949	0.923	0.333	0.358
overthrw	6	89.522	0.851	0.75	0.837	0.842	11.429	0.785	0.879	0.984	0.948	0.562	0.515
petition	33	61.888	0.648	0.525	0.594	0.596	7.627	0.663	0.797	0.914	0.952	0.186	0.312
privacy	41	31.41	0.318	0.262	0.283	0.303	3.909	0.532	0.694	0.617	0.907	0.21	0.258
opinion	38	33.069	0.336	0.288	0.311	0.323	5.677	0.497	0.664	0.841	0.927	0.166	0.196
Legal Procedural Rights													
prerel	16	69.006	0.721	0.589	0.646	0.664	6.548	0.741	0.851	0.928	0.938	0.275	0.282
juvenile	9	51.623	0.544	0.452	0.442	0.512	3.226	0.703	0.826	0.949	0.869	0.451	0.405
presinoc	32	51.769	0.531	0.426	0.465	0.478	4.8	0.708	0.829	0.926	0.919	0.388	0.302
wolaw	35	65.58	0.723	0.641	0.684	0.682	2.804	0.832	0.908	0.911	0.983	0.196	0.352
miranda	19	50.586	0.511	0.401	0.455	0.493	6.63	0.671	0.803	0.727	0.771	0.209	0.205

Continued on the next page

Topic id	#C	CR (%)	Rouge-1	Rouge-2	Rouge-L	Rouge-LSum	N (%)	J	D	SBERT	TF-IDF	BH	BT
jury	19	61.55	0.668	0.611	0.654	0.649	1.156	0.743	0.853	0.84	0.941	0.35	0.363
excrim	26	55.054	0.583	0.499	0.548	0.548	6.335	0.711	0.831	0.953	0.977	0.322	0.358
amparo	1	100.25	1	1	1	1	2.083	0.959	0.979	1	1	0.646	0.600
fairtri	27	48.677	0.527	0.467	0.497	0.509	3.586	0.722	0.839	0.887	0.94	0.239	0.328
falseimp	24	50.873	0.523	0.418	0.47	0.489	5.236	0.656	0.792	0.784	0.862	0.247	0.243
couns	29	55.554	0.588	0.503	0.535	0.558	1.794	0.716	0.834	0.911	0.957	0.235	0.337
dueproc	6	94.937	0.968	0.954	0.968	0.95	0.629	0.963	0.981	0.997	0.991	0.425	0.497
habcorp	42	15.275	0.155	0.13	0.128	0.147	6.383	0.291	0.451	0.761	0.892	0.121	0.174
appeal	29	55.118	0.584	0.489	0.536	0.544	4.56	0.718	0.836	0.861	0.943	0.148	0.286
speedtri	27	42.769	0.437	0.347	0.375	0.403	8.586	0.584	0.737	0.871	0.892	0.227	0.261
examwit	9	67.422	0.673	0.527	0.629	0.64	9.278	0.721	0.838	0.946	0.893	0.318	0.231
trilang	18	55.578	0.55	0.428	0.49	0.502	9.444	0.682	0.811	0.943	0.925	0.244	0.310
expost	32	51.515	0.554	0.473	0.487	0.529	5.59	0.717	0.835	0.907	0.973	0.223	0.281
vicright	5	80.641	0.778	0.589	0.7	0.715	9.639	0.806	0.893	0.961	0.913	0.396	0.325
doubjep	17	41.464	0.412	0.331	0.378	0.383	5.882	0.584	0.737	0.806	0.775	0.158	0.187
pubtri	36	35.398	0.368	0.317	0.338	0.342	4.142	0.572	0.728	0.908	0.927	0.276	0.300
evidence	32	46.777	0.48	0.398	0.438	0.451	5.072	0.667	0.8	0.758	0.893	0.121	0.254
Enforcement													
hr	2	81.758	0.811	0.637	0.733	0.733	6.25	0.818	0.9	0.949	0.883	0.412	0.338
amparo	1	100.25	1	1	1	1	2.083	0.959	0.979	1	1	0.646	0.646
ombuds	26	23.88	0.228	0.183	0.201	0.217	5.665	0.444	0.615	0.661	0.805	0.074	0.141
inalrht	32	54.61	0.589	0.516	0.554	0.574	4.255	0.647	0.786	0.575	0.842	0.193	0.300
Equality_Gender_and_Minority_Rights													
equalgr2	13	73.825	0.771	0.656	0.73	0.733	7.792	0.751	0.858	0.938	0.948	0.244	0.363
equalgr13	20	50.437	0.562	0.484	0.498	0.536	1.667	0.742	0.852	0.916	0.943	0.242	0.261
equalgr5	24	56.295	0.626	0.55	0.592	0.609	3.226	0.765	0.867	0.954	0.959	0.277	0.288
indpolgr1	1	20.686	0.164	0.085	0.153	0.14	5.195	0.33	0.497	0.808	0.595	0.117	0.081
indpolgr6	4	18.614	0.17	0.117	0.147	0.149	6.803	0.321	0.486	0.808	0.768	0.164	0.132
equalgr11	21	51.77	0.521	0.449	0.499	0.496	5.983	0.596	0.747	0.824	0.875	0.17	0.237
equalgr16	14	44.51	0.512	0.432	0.478	0.482	3.261	0.669	0.802	0.959	0.914	0.344	0.247
equalgr8	6	58.895	0.544	0.376	0.518	0.508	7.921	0.669	0.802	0.925	0.804	0.22	0.187
equalgr3	22	52.926	0.528	0.453	0.504	0.509	8.197	0.626	0.77	0.839	0.88	0.159	0.206
equalgr10	13	41.04	0.406	0.342	0.382	0.39	9.581	0.551	0.711	0.843	0.824	0.191	0.228
rightres	6	70.774	0.689	0.549	0.621	0.632	8.759	0.691	0.817	0.888	0.857	0.426	0.328
equalgr1	38	46.728	0.485	0.419	0.455	0.462	4.096	0.583	0.737	0.823	0.936	0.092	0.237
equalgr4	30	45.361	0.471	0.406	0.443	0.449	6.923	0.602	0.752	0.772	0.924	0.152	0.243
socclas	12	87.199	0.901	0.839	0.884	0.886	4.167	0.83	0.907	0.863	0.966	0.21	0.490
selfdet	14	48.189	0.512	0.414	0.472	0.472	2.97	0.634	0.776	0.89	0.902	0.243	0.284
equal	43	49.019	0.532	0.476	0.506	0.513	2.506	0.675	0.806	0.847	0.97	0.159	0.307

Continued on the next page

Topic id	#C	CR (%)	Rouge-1	Rouge-2	Rouge-L	Rouge-LSum	N (%)	J	D	SBERT	TF-IDF	BH	BT
equalgr15	22	47.409	0.475	0.396	0.432	0.446	7.826	0.607	0.756	0.863	0.893	0.155	0.237
equalgr12	22	51.341	0.573	0.492	0.519	0.54	2.4	0.744	0.853	0.967	0.936	0.298	0.266
asylum	28	47.113	0.502	0.427	0.452	0.478	4.329	0.656	0.792	0.912	0.922	0.274	0.314
indpolgr2	2	31.547	0.328	0.309	0.324	0.327	3.077	0.768	0.869	0.915	0.797	0.334	0.261
cultgrht	35	26.83	0.268	0.232	0.248	0.256	2.417	0.395	0.567	0.792	0.869	0.176	0.184
matequal	20	55.574	0.537	0.431	0.467	0.489	5.634	0.698	0.822	0.944	0.879	0.402	0.336
equalgr6	34	25.179	0.262	0.229	0.217	0.254	1.538	0.412	0.583	0.881	0.917	0.242	0.186
equalgr9	10	56.885	0.584	0.432	0.468	0.518	9	0.728	0.843	0.959	0.883	0.309	0.191
indcit	2	12.522	0.1	0.064	0.091	0.093	8	0.223	0.365	0.7	0.501	0.079	0.058
equalgr7	5	31.009	0.262	0.186	0.236	0.248	15.172	0.48	0.649	0.842	0.647	0.176	0.137

Table A.2: Answering RQ2 – Summarization performance for all topics: generated summaries vs. reference summaries.

Topic id	#C	CR (%)	Rouge-1	Rouge-2	Rouge-L	Rouge-LSum	N (%)	J	D	SBERT	TF-IDF	BH	BT
Physical Integrity Rights													
slave		105.81	0.692	0.377	0.246	0.422	31.638	0.528	0.691	0.916	0.781		
cruelty		140.613	0.631	0.382	0.243	0.447	33.548	0.539	0.701	0.902	0.764		
cappun		107.398	0.743	0.469	0.444	0.459	25.862	0.614	0.761	0.943	0.761		
life		72.215	0.609	0.373	0.223	0.439	29.508	0.432	0.604	0.92	0.731		
torture		138.812	0.64	0.384	0.254	0.472	40.152	0.513	0.678	0.939	0.658		
corppun		117.01	0.64	0.358	0.294	0.381	40.351	0.459	0.63	0.936	0.446		
Social Rights													
edcomp		78.58	0.562	0.236	0.282	0.396	48.214	0.374	0.545	0.885	0.717		
env		81.635	0.654	0.314	0.324	0.563	34.99	0.44	0.611	0.885	0.798		
finsup3		61.713	0.505	0.22	0.23	0.405	38.172	0.352	0.52	0.912	0.655		
safework		103.184	0.599	0.255	0.357	0.483	47.154	0.367	0.537	0.906	0.734		
provwork		130.702	0.566	0.259	0.319	0.439	49.798	0.39	0.561	0.882	0.706		
conright		115.465	0.591	0.261	0.406	0.435	44.954	0.408	0.58	0.948	0.682		
childwrk		92.078	0.624	0.292	0.323	0.477	38.235	0.441	0.612	0.949	0.690		
standliv		264.267	0.294	0.128	0.188	0.222	51.442	0.399	0.571	0.926	0.565		
edfree		85.829	0.662	0.336	0.374	0.531	36.5	0.431	0.602	0.955	0.892		
finsup1		129.208	0.577	0.273	0.365	0.449	50.951	0.379	0.55	0.887	0.672		
finsup4		96.42	0.691	0.329	0.384	0.558	37.801	0.465	0.635	0.865	0.771		
remuner		146.191	0.53	0.26	0.306	0.44	54.839	0.338	0.505	0.91	0.715		
shelter		123.558	0.559	0.224	0.275	0.387	48.058	0.399	0.571	0.888	0.659		
strike		97.13	0.662	0.321	0.437	0.54	37.956	0.425	0.596	0.976	0.824		

Continued on the next page

Topic id	CR (%)	Rouge-1	Rouge-2	Rouge-L	Rouge-LSum	N (%)	J	D	SBERT	TF-IDF
health	112.262	0.628	0.328	0.41	0.512	39.496	0.463	0.633	0.952	0.744
achighed	121.638	0.61	0.285	0.391	0.456	50	0.376	0.547	0.958	0.803
leisure	88.943	0.607	0.277	0.332	0.481	47.959	0.315	0.479	0.869	0.775
jointrde	39.705	0.354	0.194	0.187	0.309	35.294	0.301	0.463	0.928	0.728
finsup2	111.301	0.668	0.329	0.347	0.508	41.841	0.424	0.595	0.896	0.783
water	118.229	0.478	0.258	0.372	0.372	51.667	0.382	0.552	0.746	0.489
scifree	91.545	0.573	0.249	0.468	0.468	47.727	0.343	0.511	0.89	0.507
Economic Rights										
busines	124.012	0.625	0.296	0.391	0.504	45.798	0.398	0.57	0.942	0.731
intprop	113.148	0.685	0.415	0.477	0.592	34.74	0.502	0.669	0.914	0.748
transfer	119.953	0.558	0.246	0.36	0.434	63	0.261	0.413	0.913	0.823
occupate	74.528	0.575	0.252	0.314	0.473	41.667	0.357	0.527	0.859	0.645
proprght	60.898	0.512	0.216	0.247	0.423	47.368	0.283	0.441	0.811	0.833
exprop	75.118	0.577	0.243	0.277	0.486	45.059	0.351	0.52	0.882	0.850
freecomp	142.497	0.602	0.287	0.34	0.45	48.165	0.406	0.578	0.921	0.834
Citizen Duties										
milserv	130.015	0.612	0.308	0.342	0.497	51.477	0.371	0.541	0.934	0.764
taxes	123.878	0.518	0.22	0.305	0.387	50	0.359	0.528	0.904	0.568
work	94.389	0.592	0.308	0.413	0.413	48.529	0.35	0.519	0.83	0.546
General Duties										
abide	139.561	0.576	0.328	0.356	0.483	51.938	0.363	0.532	0.898	0.786
binding	106.49	0.578	0.282	0.452	0.501	52.778	0.33	0.496	0.942	0.636
Civil and Political Rights										
assoc	81.877	0.659	0.352	0.362	0.559	36.1	0.41	0.581	0.916	0.866
express	82.521	0.635	0.317	0.316	0.543	43.645	0.379	0.55	0.834	0.780
voteun	138.407	0.628	0.393	0.381	0.523	39.56	0.47	0.64	0.918	0.907
assem	78.834	0.644	0.318	0.329	0.529	34.896	0.434	0.605	0.939	0.820
fndfam	133.167	0.601	0.352	0.437	0.517	49.367	0.392	0.563	0.962	0.732
freemove	64.871	0.538	0.246	0.244	0.455	42.408	0.334	0.501	0.885	0.748
libel	56.608	0.482	0.227	0.253	0.407	41.489	0.241	0.389	0.801	0.652
nomil	112.322	0.596	0.278	0.322	0.484	50.769	0.352	0.52	0.936	0.743
civmar	23.249	0.039	0.009	0.024	0.024	87.5	0.061	0.114	0.235	0.002
freerel	23.44	0.211	0.112	0.109	0.183	30.579	0.235	0.38	0.895	0.745
arms	41.791	0.056	0	0.056	0.056	100	0	0	0.623	0.151
citren	80.539	0.458	0.17	0.324	0.359	54.762	0.275	0.432	0.913	0.551
dignity	125.297	0.644	0.35	0.424	0.52	42.382	0.431	0.602	0.904	0.836
marriage	112.067	0.631	0.366	0.452	0.546	40.397	0.431	0.602	0.932	0.758
childpro	103.579	0.674	0.307	0.336	0.547	44.904	0.415	0.587	0.952	0.824
acfree	152.325	0.534	0.28	0.37	0.415	50	0.383	0.554	0.933	0.658

Continued on the next page

Topic id	CR (%)	Rouge-1	Rouge-2	Rouge-L	Rouge-LSum	N (%)	J	D	SBERT TF-IDF
infoacc	93.721	0.646	0.292	0.365	0.549	45.035	0.36	0.53	0.872 0.781
press	64.287	0.56	0.267	0.275	0.489	36.735	0.331	0.497	0.921 0.738
debtors	66.477	0.493	0.232	0.36	0.36	47.059	0.3	0.462	0.88 0.423
devlpers	125.188	0.535	0.247	0.37	0.42	51.592	0.357	0.526	0.924 0.644
overthrw	124.316	0.622	0.356	0.468	0.468	52.857	0.34	0.508	0.879 0.629
petition	132.208	0.645	0.35	0.37	0.495	42.373	0.466	0.636	0.904 0.846
privacy	70.166	0.607	0.306	0.316	0.537	37.459	0.413	0.584	0.894 0.706
opinion	67.496	0.571	0.31	0.318	0.49	37.991	0.387	0.558	0.963 0.761
Legal_Procedural_Rights									
prerel	182.146	0.402	0.194	0.254	0.311	60.119	0.307	0.47	0.864 0.542
juvenile	96.611	0.519	0.193	0.331	0.421	49.462	0.333	0.5	0.894 0.514
presinoc	129.293	0.574	0.28	0.342	0.471	56.8	0.318	0.482	0.924 0.728
wolaw	135.409	0.558	0.292	0.294	0.448	54.206	0.332	0.499	0.864 0.657
miranda	94.069	0.583	0.234	0.313	0.423	47.514	0.347	0.515	0.894 0.558
jury	138.303	0.52	0.234	0.341	0.409	58.382	0.3	0.462	0.933 0.561
excrim	119.397	0.636	0.317	0.364	0.505	38.914	0.429	0.6	0.944 0.732
amparo	167.083	0.461	0.298	0.365	0.365	47.917	0.431	0.602	0.64 0.522
fairtri	109.108	0.649	0.266	0.322	0.515	45.02	0.397	0.568	0.913 0.762
falseimp	102.666	0.609	0.297	0.342	0.477	48.691	0.36	0.53	0.89 0.664
couns	102.592	0.642	0.327	0.324	0.514	42.601	0.412	0.583	0.925 0.684
dueproc	187.612	0.377	0.17	0.261	0.286	62.264	0.299	0.46	0.912 0.464
habcorp	31.587	0.302	0.162	0.131	0.264	27.128	0.307	0.47	0.865 0.663
appeal	120.237	0.648	0.298	0.352	0.509	43.322	0.397	0.569	0.912 0.813
speedtri	92.018	0.661	0.328	0.343	0.537	45.455	0.371	0.541	0.906 0.730
examwit	121.61	0.536	0.215	0.332	0.363	60.825	0.25	0.4	0.938 0.541
trilang	106.076	0.605	0.289	0.33	0.454	51.111	0.344	0.512	0.893 0.583
expost	114.35	0.663	0.342	0.347	0.547	49.689	0.365	0.535	0.904 0.787
vicright	119.224	0.537	0.26	0.393	0.398	46.988	0.393	0.564	0.905 0.551
doubjep	102.986	0.59	0.271	0.31	0.438	44.706	0.367	0.537	0.859 0.605
pubtri	73.183	0.616	0.328	0.348	0.533	36.686	0.423	0.594	0.949 0.779
evidence	97.476	0.681	0.364	0.389	0.57	34.541	0.473	0.642	0.895 0.737
Enforcement									
hr	124.415	0.537	0.261	0.417	0.417	50	0.364	0.533	0.857 0.615
amparo	187.822	0.397	0.226	0.362	0.362	54.167	0.386	0.557	0.65 0.504
ombuds	89.22	0.602	0.257	0.203	0.509	45.074	0.351	0.52	0.766 0.676
inalrht	119.438	0.603	0.269	0.33	0.489	50.355	0.345	0.513	0.823 0.605
Equality_Gender_and_Minority_Rights									
equalgr2	179.841	0.445	0.233	0.314	0.349	53.247	0.379	0.55	0.855 0.667
equalgr13	122.005	0.589	0.338	0.406	0.489	45	0.402	0.574	0.916 0.689

Continued on the next page

Topic id	CR (%)	Rouge-1	Rouge-2	Rouge-L	Rouge-LSum	N (%)	J	D	SBERT	TF-IDF
equalgr5	110.203	0.632	0.338	0.431	0.541	38.065	0.459	0.63	0.921	0.706
indpolgr1	48.789	0.34	0.156	0.277	0.27	25.974	0.422	0.594	0.934	0.690
indpolgr6	43.304	0.365	0.153	0.175	0.275	34.694	0.294	0.454	0.819	0.702
equalgr11	113.069	0.585	0.266	0.337	0.488	42.735	0.395	0.567	0.871	0.722
equalgr16	85.617	0.607	0.239	0.325	0.463	43.478	0.397	0.568	0.937	0.640
equalgr8	103.328	0.552	0.237	0.307	0.343	51.485	0.355	0.524	0.921	0.552
equalgr3	122.126	0.547	0.274	0.346	0.467	41.393	0.407	0.579	0.833	0.646
equalgr10	119.321	0.526	0.202	0.26	0.407	47.904	0.345	0.513	0.886	0.530
rightres	117.322	0.617	0.291	0.392	0.441	48.175	0.364	0.534	0.905	0.587
equalgr1	105.941	0.627	0.323	0.358	0.535	37.201	0.435	0.606	0.881	0.802
equalgr4	105.303	0.605	0.288	0.354	0.514	40.769	0.401	0.572	0.781	0.709
socclas	212.569	0.343	0.156	0.216	0.281	68.182	0.265	0.419	0.671	0.447
selfdet	109.671	0.704	0.407	0.46	0.545	40.924	0.431	0.603	0.888	0.861
equal	108.89	0.658	0.302	0.325	0.574	44.612	0.39	0.561	0.889	0.791
equalgr15	106.839	0.62	0.285	0.34	0.506	38.696	0.438	0.609	0.904	0.696
equalgr12	121.749	0.584	0.309	0.362	0.455	44	0.407	0.579	0.894	0.736
asylum	105.377	0.62	0.277	0.378	0.504	49.351	0.36	0.529	0.952	0.683
indpolgr2	78.48	0.649	0.413	0.447	0.532	32.308	0.484	0.652	0.938	0.781
cultright	61.098	0.541	0.285	0.226	0.48	32.628	0.342	0.51	0.827	0.814
matequal	97.511	0.586	0.308	0.342	0.468	47.887	0.33	0.497	0.891	0.662
equalgr6	54.697	0.461	0.231	0.252	0.418	37.692	0.345	0.513	0.858	0.677
equalgr9	110.431	0.577	0.234	0.284	0.391	45	0.42	0.591	0.958	0.620
indcit	31.42	0.112	0.04	0.068	0.09	70.667	0.152	0.263	0.676	0.020
equalgr7	80.148	0.531	0.197	0.285	0.395	50.345	0.314	0.478	0.909	0.551

A.3 Statistics on term anonymization.

Table A.3 shows statistics related to the presence of the anonymized terms in the summaries. In the first three columns, we report the ratio of the total length of anonymized terms to the length of the full text of both the summary (aS) and the original texts (oS) and the ratio between such values (aS/aO). We note that, in both the summary and the original text, the presence of these terms is mostly low. In less than 15% of cases, the ratio between these two values is below 60%. By contrast, for

the remaining topics, there is some alignment between the presence of the terms in the summary and in the original text.

In the last three columns, we report the number of anonymized terms in the summary (#aS), the total number of the anonymized terms in the country-level summaries obtained in Stage 1 (#aC) and the ratio between such values (#aS/#aC). We observe a certain coherence in the number of terms in the topic-level summary and the country-level summaries. In more than 75% of cases, the ratio between these values is above 60%.

A.4 Experimental environment

We carried out all our experiments using a 8x NVIDIA A30 GPU server with 24 GB of RAM each, 764 GB of system RAM, a Double Intel Xeon Gold 6248R with a total of 96 cores, and Ubuntu Linux 20.04.6 LTS as operating system.

Table A.3: Statistics on the anonymized terms. In order, the columns correspond to the ratio of the total length of anonymized terms in the summary to the length of the full summary (aS), the ratio of the total length of anonymized terms in the original text to the length of the full text (aO), the ratio between aS and aO (aS/aO), the number of anonymized terms in the summary (#aS), the total number of the anonymized terms in the country-level summaries obtained in Stage 1 (#aC) and the ratio between #aS and #aC (#aS/#aC).

M	aS	aO	aS/aO	#aS	#aC	#aS/#aC	M	aS	aO	aS/aO	#aS	#aC	#aS/#aC
Physical Integrity Rights							infoacc	0.031	0.036	0.882	16	24	0.667
slave	0.011	0.020	0.569	4	14	0.286	press	0.027	0.026	1.040	21	27	0.778
cruelty	0.010	0.011	0.862	5	7	0.714	debtors	0.000	0.000	0.000	0	0	0.000
cappun	0.057	0.074	0.776	2	6	0.333	devlpers	0.031	0.039	0.780	5	13	0.385
life	0.053	0.040	1.309	3	7	0.429	overthrw	0.069	0.079	0.878	3	4	0.750
torture	0.011	0.008	1.358	4	5	0.800	petition	0.037	0.034	1.103	21	20	1.050
corppun	0.008	0.006	1.309	1	1	1.000	privacy	0.019	0.036	0.523	11	71	0.155
Social Rights							opinion	0.013	0.010	1.334	8	10	0.800
edcomp	0.056	0.034	1.665	3	5	0.600	Legal and Procedural Rights						
env	0.049	0.046	1.070	34	47	0.723	prerel	0.054	0.056	0.954	16	20	0.800
finstup3	0.025	0.062	0.395	13	33	0.394	juvenile	0.046	0.039	1.172	3	5	0.600
safework	0.009	0.007	1.150	2	2	1.000	presinoc	0.005	0.002	1.932	2	2	1.000
provwork	0.041	0.033	1.250	12	14	0.857	wolaw	0.012	0.011	1.083	6	6	1.000
conright	0.065	0.054	1.211	7	7	1.000	miranda	0.015	0.021	0.711	5	7	0.714
childwrk	0.041	0.028	1.455	5	5	1.000	jury	0.013	0.019	0.700	8	15	0.533
standliv	0.012	0.013	0.930	6	5	1.200	excrim	0.130	0.177	0.736	42	57	0.737
edfree	0.026	0.026	0.991	5	7	0.714	amparo	0.044	0.044	0.998	5	5	1.000
finstup1	0.055	0.062	0.877	23	28	0.821	fairtri	0.009	0.010	0.922	3	5	0.600
finstup4	0.030	0.058	0.512	17	21	0.810	falseimp	0.018	0.056	0.329	7	10	0.700
remuner	0.005	0.003	1.434	2	2	1.000	couns	0.009	0.010	0.878	8	10	0.800
shelter	0.034	0.052	0.663	8	23	0.348	dueproc	0.004	0.004	1.053	2	2	1.000
strike	0.006	0.004	1.543	2	2	1.000	habcorp	0.031	0.034	0.910	12	51	0.235
health	0.020	0.019	1.095	9	9	1.000	appeal	0.107	0.112	0.962	64	69	0.928
achighed	0.044	0.038	1.151	7	9	0.778	speedtri	0.020	0.012	1.697	5	7	0.714
leisure	0.000	0.009	0.000	0	1	0.000	examwit	0.007	0.010	0.742	2	2	1.000
jointrde	0.031	0.031	0.990	5	25	0.200	trilang	0.034	0.034	0.997	12	13	0.923
finstup2	0.022	0.048	0.466	10	13	0.769	expost	0.003	0.004	0.813	2	3	0.667
water	0.066	0.060	1.098	2	2	1.000	vicright	0.014	0.015	0.906	2	2	1.000
scifree	0.000	0.000	0.000	0	0	0.000	doubjep	0.003	0.018	0.174	1	3	0.333
Economic Rights							pubtri	0.009	0.014	0.655	1	7	0.143
business	0.040	0.079	0.512	16	23	0.696	evidence	0.012	0.021	0.553	14	18	0.778
intprop	0.022	0.038	0.589	9	13	0.692	Enforcement						
transfer	0.022	0.015	1.503	2	2	1.000	hr	0.118	0.165	0.718	6	7	0.857
occupate	0.029	0.022	1.337	5	5	1.000	ombuds	0.093	0.110	0.852	95	137	0.693
proprght	0.027	0.024	1.137	8	15	0.533	inalrht	0.013	0.056	0.239	5	57	0.088
exprop	0.028	0.031	0.887	8	13	0.615	Equality Gender and Minority Rights						
freecomp	0.067	0.078	0.859	13	16	0.812	equalgr2	0.039	0.036	1.065	9	9	1.000
Citizen Duties							equalgr13	0.049	0.058	0.846	7	13	0.538
milserv	0.075	0.105	0.715	21	26	0.808	equalgr5	0.050	0.042	1.199	11	13	0.846
taxes	0.028	0.024	1.138	3	4	0.750	indpolgr1	0.011	0.024	0.473	1	1	1.000
work	0.000	0.087	0.000	0	2	0.000	indpolgr6	0.087	0.093	0.941	22	22	1.000
General Duties							equalgr11	0.036	0.054	0.671	10	10	1.000
abide	0.098	0.119	0.825	38	43	0.884	equalgr16	0.082	0.063	1.291	8	10	0.800
binding	0.056	0.067	0.840	4	4	1.000	equalgr8	0.000	0.006	0.000	0	0	0.000
Civil and Political Rights							equalgr3	0.044	0.051	0.867	11	11	1.000
assoc	0.040	0.038	1.055	17	26	0.654	equalgr10	0.059	0.065	0.915	11	11	1.000
express	0.011	0.013	0.830	8	12	0.667	rightres	0.106	0.104	1.021	12	12	1.000
voteun	0.108	0.120	0.897	68	89	0.764	equalgr1	0.042	0.053	0.791	20	26	0.769
assem	0.015	0.017	0.879	7	11	0.636	equalgr4	0.057	0.060	0.947	18	24	0.750
fndfam	0.018	0.038	0.477	3	5	0.600	socclas	0.064	0.074	0.864	18	21	0.857
freemove	0.073	0.158	0.464	14	40	0.350	selfdet	0.093	0.163	0.574	33	35	0.943
libel	0.020	0.020	0.991	2	9	0.222	equal	0.050	0.068	0.745	35	41	0.854
nomil	0.039	0.051	0.766	6	9	0.667	equalgr15	0.050	0.062	0.804	16	18	0.889
civmar	0.013	0.011	1.137	1	1	1.000	equalgr12	0.058	0.058	1.000	8	14	0.571
freerel	0.024	0.017	1.454	3	16	0.188	asylum	0.079	0.121	0.650	23	39	0.590
arms	0.000	0.000	0.000	0	0	0.000	indpolgr2	0.124	0.305	0.408	15	23	0.652
citren	0.147	0.230	0.641	6	11	0.545	cultgrht	0.055	0.077	0.721	30	70	0.429
dignity	0.052	0.052	0.998	23	27	0.852	matequal	0.028	0.015	1.799	1	1	1.000
marriage	0.024	0.027	0.881	5	6	0.833	equalgr6	0.049	0.049	1.001	7	23	0.304
childpro	0.003	0.014	0.227	3	10	0.300	equalgr9	0.051	0.050	1.027	4	5	0.800
acfree	0.024	0.028	0.851	7	8	0.875	indcit	0.108	0.036	2.967	5	6	0.833
							equalgr7	0.020	0.062	0.320	2	2	1.000

Appendix B

B.1 Zero-shot Prompts on WordNet data

System Prompt

You are a linguistic expert who responds to user questions in a concise way.

Direct Prompt

Given the verb v_1 defined as $def(v_1)$ and the verb v_2 defined as $def(v_2)$, what is the verb that entails the other? Answer must be either of the form “ X entails Y ” or “there is no entailment”.

Answer: `select{“ $v_1$  entails  $v_2$ ”, “ $v_2$  entails  $v_1$ ”, “there is not entailment”}.`

*Confidence of the answer (1 is the lowest, 10 is the highest):
`select{“1: No Confidence”, “2: Very Low Confidence”, “3: Low Confidence”, “4: Fair Confidence”, “5: Moderate Confidence”, “6: Good Confidence”, “7: High Confidence”, “8: Very High Confidence”, “9: Extremely High Confidence”, “10: Absolute Certainty”}.`*

Indirect Prompt

Relation F states that given two verbs X and Y , X and Y satisfy F if and only if when doing Y you are also doing X .

Given the verb v_1 defined as $def(v_1)$ and the verb v_2 defined as $def(v_2)$, what is X and what is Y ? Answer must be either a pair (X,Y) or “relation F cannot be satisfied”.

Answer: `select{“( $v_1$ ,  $v_2$ )”, “( $v_2$ ,  $v_1$ )”, “relation  $F$  cannot be satisfied”}.`

Confidence of the answer (1 is the lowest, 10 is the highest):
 select{ ‘1: No Confidence’, ‘2: Very Low Confidence’,
 ‘3: Low Confidence’, ‘4: Fair Confidence’, ‘5:
 Moderate Confidence’, ‘6: Good Confidence’, ‘7: High
 Confidence’, ‘8: Very High Confidence’, ‘9: Extremely
 High Confidence’, ‘10: Absolute Certainty’ }.

Reverse Prompt

Given the verb v_1 defined as $def(v_1)$, and the verb v_2 defined as $def(v_2)$,
 answer YES if ‘not v_2 ’ entails ‘not v_1 ,’ NO otherwise.

Answer: select{ ‘YES’, ‘NO’ }.

Confidence of the answer (1 is the lowest, 10 is the highest):
 select{ ‘1: No Confidence’, ‘2: Very Low Confidence’,
 ‘3: Low Confidence’, ‘4: Fair Confidence’, ‘5:
 Moderate Confidence’, ‘6: Good Confidence’, ‘7: High
 Confidence’, ‘8: Very High Confidence’, ‘9: Extremely
 High Confidence’, ‘10: Absolute Certainty’ }.

B.2 Zero-shot Prompts on HyperLex data

System Message

You are a linguistic expert who responds to user questions in a concise way.

Direct Prompt

Given the verb v_1 and the verb v_2 , what is the verb that entails the other?
 Answer must be either of the form “X entails Y” or “there is no entailment”.

Answer: select{ ‘ v_1 entails v_2 ’, ‘ v_2 entails v_1 ’, ‘there
 is not entailment’ }.

Confidence of the answer (1 is the lowest, 10 is the highest):
 select{ ‘1: No Confidence’, ‘2: Very Low Confidence’,
 ‘3: Low Confidence’, ‘4: Fair Confidence’, ‘5:
 Moderate Confidence’, ‘6: Good Confidence’, ‘7: High
 Confidence’, ‘8: Very High Confidence’, ‘9: Extremely
 High Confidence’, ‘10: Absolute Certainty’ }.

Indirect Prompt

Relation F states that given two verbs X and Y , X and Y satisfy F if and only if when doing Y you are also doing X .

Given the verb v_1 and the verb v_2 , what is X and what is Y ? Answer must be either a pair (X,Y) or “relation F cannot be satisfied”.

Answer: `select{“(v1, v2)”, “(v2, v1)”, “relation F cannot be satisfied”}.`

Confidence of the answer (1 is the lowest, 10 is the highest):
`select{“1: No Confidence”, “2: Very Low Confidence”, “3: Low Confidence”, “4: Fair Confidence”, “5: Moderate Confidence”, “6: Good Confidence”, “7: High Confidence”, “8: Very High Confidence”, “9: Extremely High Confidence”, “10: Absolute Certainty”}.`

Reverse Prompt

Given the verb v_1 and the verb v_2 , answer YES if ‘not v_2 ’ entails ‘not v_1 ,’ NO otherwise.

Answer: `select{“YES”, “NO”}.`

Confidence of the answer (1 is the lowest, 10 is the highest):
`select{“1: No Confidence”, “2: Very Low Confidence”, “3: Low Confidence”, “4: Fair Confidence”, “5: Moderate Confidence”, “6: Good Confidence”, “7: High Confidence”, “8: Very High Confidence”, “9: Extremely High Confidence”, “10: Absolute Certainty”}.`

B.3 Few-shot prompt selection strategy

For the HyperLex-FS prompting, we selected the following verb pairs, one from each of the subranges within [2..6] with step 1 (the associated HyperLex score is also reported within square brackets):

- (warn, advise) [5.75]
- (instruct, inform) [4.31]
- (rationalize, argue) [3.08]
- (take, have) [2.17]

Analogously, for the Fellbaum-FS prompting, we selected the following verb pairs, one from each of the Fellbaum’s categories of verb entailments Fellbaum (1990):

- *Entailment with temporal co-extensiveness*: the activity denoted by the entailing verb implies a more general one, denoted by the entailed verb, in a simultaneous manner (i.e., they are temporally co-extensive). This corresponds to troponymy.
- *Entailment with temporal proper inclusion*: the activity denoted by the entailing verb includes the activity denoted by the entailed verb, and is not temporally co-extensive (i.e., one activity can occur before or after the other).
- *Entailment with backward presupposition*: the activity denoted by the entailed verb always precedes the activity denoted by the entailing verb in time.
- *Cause*: if the activity denoted by verb v_1 causes the activity denoted by verb v_2 , then v_1 also entails v_2 .

It should be noted that no categorization of the WordNet verbs according to the Fellbaum taxonomy was carried out, since WordNet does not provide annotations on the Fellbaum taxonomy types. In addition, evaluating a prediction task on the Fellbaum taxonomy types would be challenging since we are not aware of the existence of lexical databases that provide ground truth labels for these subtypes, thus going beyond the scope of this study.

Below we show the Fellbaum-FS-based verb pairs that were selected to define four response examples to provide to each of the models in a few-shot scenario:

- (limp, walk), as example of troponymy co-extensiveness relation;
- (snore, sleep), as example of troponymy proper inclusion relation;
- (succeed, try), as example of backward presupposition relation;
- (give, have), as example of cause relation.

In the following, we report the pre-prompts (user messages) used for each of the three prompt types (i.e., Direct, Indirect, and Reverse) with HyperLex-FS examples and Fellbaum-FS examples, respectively, over WordNet data. Note that these pre-prompts also apply to HyperLex data apart from the specification of verb definitions.

HyperLex-FS pre-prompt for the Direct prompt

Here are some examples of the task you have to perform:

Start Examples

- Example 1

Input: Given the verb 'warn' defined as 'admonish or counsel in terms of someone's behavior', and the verb 'advise' defined as 'give advice to'. What is the verb that entails the other? Answer must be of the form X entails Y.

Output: 'warn' entails 'advise'

- Example 2

Input: Given the verb 'inform' defined as 'impart knowledge of some fact, state or affairs, or event to', and the verb 'instruct' defined as 'make aware of'. What is the verb that entails the other? Answer must be of the form X entails Y.

Output: 'instruct' entails 'inform'

- Example 3

Input: Given the verb 'argue' defined as 'present reasons and arguments', and the verb 'rationalize' defined as 'defend, explain, clear away, or make excuses for by reasoning'. What is the verb that entails the other? Answer must be of the form X entails Y.

Output: 'rationalize' entails 'argue'

- Example 4

Input: Given the verb 'take' defined as 'experience or feel or submit to', and the verb 'have' defined as 'go through (mental or physical states or experiences)'. What is the verb that entails the other? Answer must be of the form X entails Y.

Output: 'take' entails 'have'

End Examples

HyperLex-FS pre-prompt for the Indirect prompt

Here are some examples of the task you have to perform:

Start Examples

- Example 1

Input: Relation F states that given two verbs X and Y, X and Y satisfy F if and only if when doing Y you are also doing X. Given the verb 'warn' defined as 'admonish or counsel in terms of someone's behavior', and the verb 'advise' defined as 'give advice to' What is X and what is Y? Answer must be a pair (X, Y).

Output: (warn, advise)

- Example 2

Input: Relation F states that given two verbs X and Y, X and Y satisfy F if and only if when doing Y you are also doing X. Given the verb 'inform' defined as 'impart knowledge of some fact, state or affairs, or event to', and the verb 'instruct' defined as 'make aware of' What is X and what is Y? Answer must be a pair (X ,Y).

Output: (instruct, inform)

- Example 3

Input: Relation F states that given two verbs X and Y, X and Y satisfy F if and only if when doing Y you are also doing X. Given the verb 'argue' defined as 'present reasons and arguments', and the verb 'rationalize' defined as 'defend, explain, clear away, or make excuses for by reasoning' What is X and what is Y? Answer must be a pair (X ,Y).

Output: (rationalize, argue)

- Example 4

Input: Relation F states that given two verbs X and Y, X and Y satisfy F if and only if when doing Y you are also doing X. Given the verb 'take' defined as 'experience or feel or submit to', and the verb 'have' defined as 'go through (mental or physical states or experiences)' What is X and what is Y? Answer must be a pair (X ,Y).

Output: (take, have)

End Examples

HyperLex-FS pre-prompt for the Reverse prompt

Here are some examples of the task you have to perform:

Start Examples

- Example 1

Input: Given the verb 'warn' defined as 'admonish or counsel in terms of someone's behavior', and the verb 'advise' defined as 'give advice to' Answer YES if 'not warn' entails 'not advise', NO otherwise.

Output: NO

- Example 2

Input: Given the verb 'inform' defined as 'impart knowledge of some fact,

state or affairs, or event to', and the verb 'instruct' defined as 'make aware of' Answer YES if 'not inform' entails 'not instruct', NO otherwise.

Output: YES

- Example 3

Input: Given the verb 'argue' defined as 'present reasons and arguments', and the verb 'rationalize' defined as 'defend, explain, clear away, or make excuses for by reasoning' Answer YES if 'not argue' entails 'not rationalize', NO otherwise.

Output: YES

- Example 4

Input: Given the verb 'take' defined as 'experience or feel or submit to', and the verb 'have' defined as 'go through (mental or physical states or experiences)' Answer YES if 'not take' entails 'not have', NO otherwise.

Output: NO

End Examples

Fellbaum-FS pre-prompt for the Direct prompt

Here are some examples of the task you have to perform:

Start Examples

- Example 1

Input: Given the verb 'limp' defined as 'walk impeded by some physical limitation or injury', and the verb 'walk' defined as 'use one's feet to advance; advance by steps'. What is the verb that entails the other? Answer must be of the form X entails Y.

Output: 'limp' entails 'walk'

- Example 2

Input: Given the verb 'sleep' defined as 'be asleep', and the verb 'snore' defined as 'breathe noisily during one's sleep'. What is the verb that entails the other? Answer must be of the form X entails Y.

Output: 'snore' entails 'sleep'

- Example 3

Input: Given the verb 'try' defined as 'make an effort or attempt', and the verb 'succeed' defined as 'attain success or reach a desired goal'. What is the verb that entails the other? Answer must be of the form X entails Y.

Output: 'succeed' entails 'try'

- Example 4

Input: Given the verb 'give' defined as 'guide or direct, as by behavior of persuasion', and the verb 'have' defined as 'cause to do; cause to act in a specified manner'. What is the verb that entails the other? Answer must be of the form X entails Y.

Output: 'give' entails 'have'

End Examples

Fellbaum-FS pre-prompt for the Indirect prompt

Here are some examples of the task you have to perform:

Start Examples

- Example 1

Input: Relation F states that given two verbs X and Y, X and Y satisfy F if and only if when doing Y you are also doing X. Given the verb 'limp' defined as 'walk impeded by some physical limitation or injury', and the verb 'walk' defined as 'use one's feet to advance; advance by steps' What is X and what is Y? Answer must be a pair (X, Y).

Output: (limp, walk)

- Example 2

Input: Relation F states that given two verbs X and Y, X and Y satisfy F if and only if when doing Y you are also doing X. Given the verb 'sleep' defined as 'be asleep', and the verb 'snore' defined as 'breathe noisily during one's sleep' What is X and what is Y? Answer must be a pair (X, Y).

Output: (sleep, snore)

- Example 3

Input: Relation F states that given two verbs X and Y, X and Y satisfy F if and only if when doing Y you are also doing X. Given the verb 'try' defined as 'make an effort or attempt', and the verb 'succeed' defined as 'attain

success or reach a desired goal' What is X and what is Y? Answer must be a pair (X ,Y).

Output: (succeed, try)

- Example 4

Input: Relation F states that given two verbs X and Y, X and Y satisfy F if and only if when doing Y you are also doing X. Given the verb 'give' defined as 'guide or direct, as by behavior of persuasion', and the verb 'have' defined as 'cause to do; cause to act in a specified manner' What is X and what is Y? Answer must be a pair (X ,Y).

Output: (give, have)

End Examples

Fellbaum-FS pre-prompt for the Reverse prompt

Here are some examples of the task you have to perform:

Start Examples

- Example 1

Input: Given the verb 'limp' defined as 'walk impeded by some physical limitation or injury', and the verb 'walk' defined as 'use one's feet to advance; advance by steps' Answer YES if 'not limp' entails 'not walk', NO otherwise.

Output: NO

- Example 2

Input: Given the verb 'sleep' defined as 'be asleep', and the verb 'snore' defined as 'breathe noisily during one's sleep'. Answer YES if 'not sleep' entails 'not snore', NO otherwise.

Output: YES

- Example 3

Input: Given the verb 'try' defined as 'make an effort or attempt', and the verb 'succeed' defined as 'attain success or reach a desired goal'. Answer YES if 'not try' entails 'not succeed', NO otherwise.

Output: YES

- Example 4

Input: Given the verb 'give' defined as 'guide or direct, as by behavior of persuasion', and the verb 'have' defined as 'cause to do; cause to act in a specified manner'. Answer YES if 'not give' entails 'not have', NO otherwise.

Output: NO

End Examples

B.4 Experimental environment

We carried out our experiments locally by deploying the models through the open-source *text-generation-webui* framework,¹ using a 8x NVIDIA A30 GPU server with 24 GB of RAM each, 764 GB of system RAM, a Double Intel Xeon Gold 6248R with a total of 96 cores, and Ubuntu Linux 20.04.6 LTS as operating system.

¹ <https://github.com/oobabooga/text-generation-webui>

Appendix C

C.1 Prompt example for case selection in E2MoCase

We found the legal cases through both an extensive manual searching and the use of generative LLM tools with online search capabilities, particularly ChatGPT Plus, Perplexity, and Claude-3-Opus-200K. Below, we report an example of the prompts used on each tool:

Search for newspaper articles related to judicial cases where judges or the media have been influenced by cultural or discriminatory stereotypes. The articles should focus on specific cases, such as those involving individuals who have experienced discrimination. For each case found, provide the name of the individual involved, the year it occurred, a description of the case, and the link to the source.

C.2 SwissDox Query

Figure C.1 shows an example of query in YAML format regarding the legal case of Johnny Depp and The Sun. This query requires the retrieval of Swiss newspaper news in a specific time range (2020-01-01 to 2024-03-16, i.e. the current date at the time of the query). The result of the request is a TSV file containing the content of the article, in xml format, and other information such as document id (id), publication date (pubtime), media code (medium_code) and article language (language).

C.3 Selected cases in E2MoCase

We selected 39 cases (Table C.1) which received high coverage in Switzerland and was covered by at least one Swiss newspaper. Note that we prioritized recent cases, due to the significant developments in the legal system and shifts in public opinion regarding civil rights, discrimination, and biases, over the years. By recent cases, we referred to those occurring within the past 10 years. Out of the 39 selected cases, 25 fall into this category. However, prioritizing recent events does not mean we exclusively included them. The remaining cases are distributed as follows: 5 are between 10-15 years old, 4 are between 15-20 years, 1 is 30 years old, and 1 is 35 years old. This selection aimed to balance the distribution of biases, as illustrated in Figure 6. Additionally, while the temporal distribution varies as outlined, we

```

query:
  dates:
    - from: 2020-01-01
    - to: 2024-03-16
  content:
    AND:
      - Johnny Depp
      - The Sun
  result:
    format: TSV
    columns:
      - id
      - pubtime
      - medium_code
      - language
      - .....

```

Fig. C.1: Swiss-Dox query example.

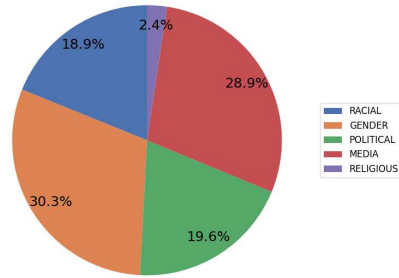


Fig. C.2: Distribution of the biases associated with the paragraphs. The bias of a paragraph is inherited from the legal case it originates from.

Table C.1: Subjects of the selected legal cases.

Britney Spears Conservatorship	Youssef Nada	Tamir Rice
Bill Cosby women accusation	Dominique Strauss-Kahn	Caster Semenya
Johnny Depp and Amber Heard	Michael Brown	Eluana Englaro
Mike Ben Peter	Eric Garner	Mohamed Wa Baile
Harvey Weinstein	Terri Schiavo	Kyle Rittenhouse
Edward Snowden	O.J. Simpson	Starbucks Arrests
Breonna Taylor	Ahmed Mohamed	Robert Kelly
Chelsea Manning	Johnny Depp and The Sun	Patrick Lyoya
Rosemarie Aquilina	Charlie Gard	Cyntoia Brown-Long
Trayvon Martin	Troy Davis	Dieudonné M'bala M'bala
Amanda Knox	Ahmaud Arbery	Nils Fiechter and Adrian Spahr
Sarah Everard	Sandra Bland	Philando Castile
Ethan Couch	Julian Assange political	Jussie Smollett

emphasize that the vast majority of retrieved news items are recent, with 43.8% less than 5 years old, 37% between 5 and 10 years old, 14% over 10 but under 15, 3% over 15 but under 20, and only 1% over 20 years old.

The biases in the legal cases have been manually detected and categorized as follows:

1. **Religious**, which refers to the unfair treatment or attitudes towards individuals or groups based on their religious beliefs.
2. **Media**, i.e., the tendency for media to present information in a way that reflects certain viewpoints, thereby influencing public perception.
3. **Political**, which involves favoring one political party, ideology, or candidate over others.

4. **Gender**, which refers to the preferential treatment or prejudice against individuals based on their gender.
5. **Racial**, which involves discriminatory attitudes or actions towards individuals based on their ethnicity.

The distribution of observed biases across these paragraphs is illustrated in Fig. C.2, with each paragraph inheriting the bias of the source news article.

C.4 Correlation Event in E2MoCase

A Correlation Event is defined through the following Python class, according to the format required by GoLLIE:

```
class CorrelationEvent(Template):
    """A CorrelationEvent refers to any dynamics, state change or
       relationship within a context involving one or more
       entities. Such as discovering, accusing, loving, helping,
       killing,
       firing, finding, cheating, divorcing, reporting.
    """
    mention: str
    """
    The text span that most clearly expresses the CorrelationEvent.
    Such as: "discover", "kill", "find", "love", "accuse", "
           love", "criticize", "discover", "fire", "cheat", "report"
    """
    entities: Dict[str, str] # Dictionary of CorrelationEvents in
                           which the keys correspond to entities and values correspond
                           to the role that entity has in the context of the
                           CorrelationEvent. Such as: "Alex":"victim", "Claire": "
                           police officer", "John":"man", "lawyer":"speaker"
```

Below are some examples of events extracted with GoLLIE based on the definition provided by the CorrelationEvent class:

```
{
  "mention": "trial",
  "entities": {
    "Weinstein": "subject"
  }
}

{
  "mention": "fire",
  "entities": {
    "Wilson": "Officer",
    "Michael Brown": "him"
  }
}

{
  "mention": "murder",
  "entities": {
    "Rittenhouse": "white supremacist"
  }
}
```

C.5 Plutchik’s Wheel of Emotions

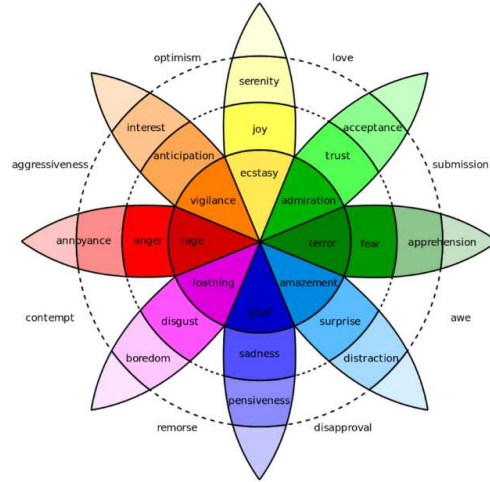


Fig. C.3: Plutchik’s Wheel of Emotions. Plutchik (2001)

Plutchik’s Wheel of Emotions Plutchik (2001) is a model that categorizes human emotions into eight primary types: *joy*, *trust*, *fear*, *surprise*, *sadness*, *disgust*, *anger*, and *anticipation*. These emotions are arranged on a wheel to represent their relationships and intensities, with similar emotions placed close to each other and opposites across from one another. The wheel, illustrated in Fig. C.3, also shows how primary emotions can blend to form complex emotions. This model provides a systematic approach to understanding emotional responses, supporting various fields, including psychology, artificial intelligence, and natural language processing.

C.6 Outlier Removal and Cleaning process in E2MoCase

We follow the strategy described in Alg. 1 to detected and, eventually, remove possible outlier paragraphs. Specifically, given a specific case, we consider all paragraphs from news articles and the associated emotions. The most prevalent emotion E is identified through all the paragraphs of the case. An emotion is prevalent if its scores boxplot has the largest inter quartile range (IQR). For example, considering Fig. C.4a, the prevalent emotion is *fear* as the IQR of its boxplot is the largest w.r.t. all the other boxplots.

Given E , we define its *opposite* and *contrary* emotions. An emotion *opposite* to E is the emotion in the opposite position to E in the Plutchik’s wheel (Fig. C.3). For example, if E is *joy*, its opposite emotion is *sadness*. If E is a positive emotion (*joy*,

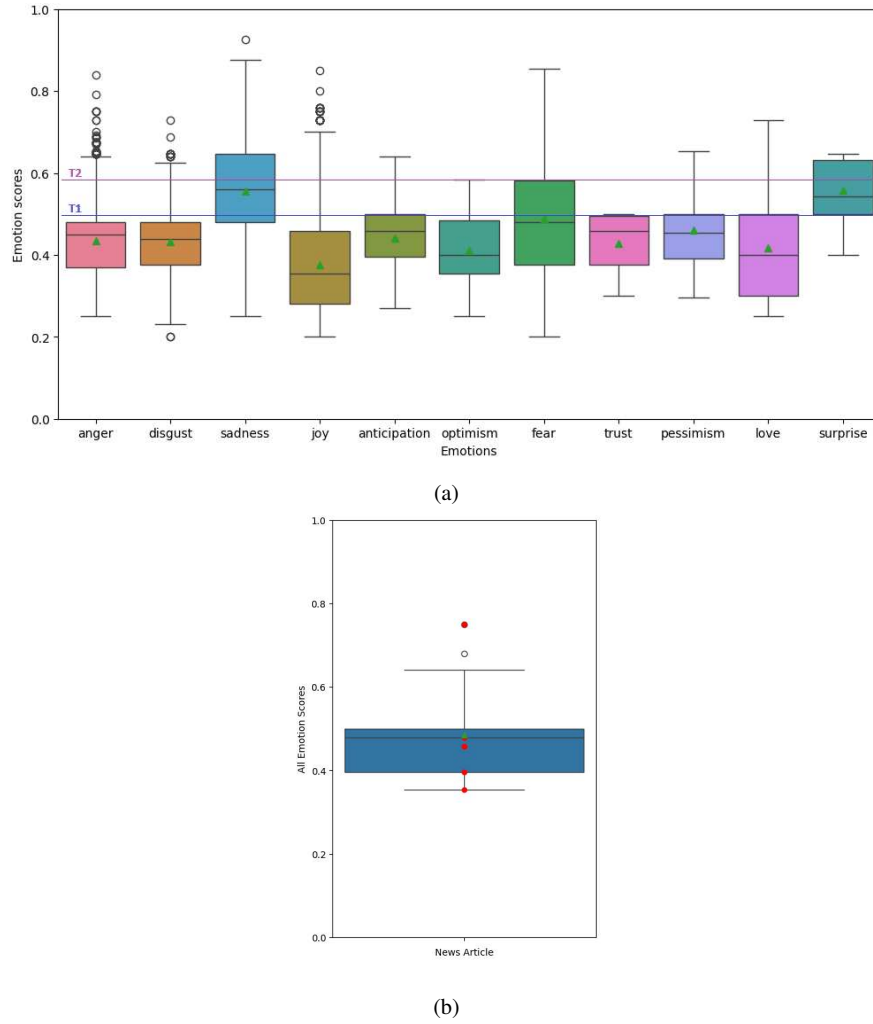


Fig. C.4: (a) Boxplots of the emotions detected in the news article discussing Britney Spears conservatorship. The prevalent emotion is *fear*, from which $T1$ and $T2$ thresholds are retrieved. (b) Boxplot of all the emotions scores in the news article from which the candidate outlier paragraph comes from. The scores of the emotion associated to the candidate outlier are highlighted with red points.

love, optimism, trust), an emotion *contrary* to E is an emotion with negative polarity (*anger, sadness, disgust, pessimism, fear*), and vice versa. Note that if an emotion is both opposite and contrary to E , it is considered solely as opposite. For example, if E is *joy*, its contrary emotions are *anger, disgust, pessimism, fear* (*sadness* is already the opposite emotion).

```

Data: case_paragraphs,
top_emotion=get_emotion_with_max_IQR(case_paragraphs)
opposite_emotion=get_opposite_emotion(top_emotion)
contrary_emotions_list=get_contrary_emotions(top_emotion)
T1 = get_mean_from_max_IQR(case_paragraphs)
T2 = get_Q3_from_max_IQR(case_paragraphs)
Result: case with cleaned emotions
foreach case_paragraph in case_paragraphs do
  foreach emotion in case_paragraph do
    if emotion == opposite_emotion and score(emotion) > T1 then
      remove_if_outlier_in_news(case_paragraph,
        case_paragraphs,
        emotion)
    end
    if emotion in contrary_emotions_list and score(emotion) > T2 then
      remove_if_outlier_in_news(case_paragraph,
        case_paragraphs,
        emotion)
    end
  end
end

```

Algorithm 1: Candidate outlier paragraphs detection pseudocode.

```

Data: case_paragraph,
case_paragraphs,
emotion,
T = 1.5,
news_article = reconstruct_news_article(case_paragraph,
case_paragraphs)
Q1 = get_Q1_from_emotion_scores(news_article)
Q3 = get_Q3_from_emotion_scores(news_article)
Result: case_paragraphs (without outlier emotion)
if score(emotion) < Q1 or score(emotion) > Q3 then
  lof_scores=LocalOutlierFactor(news_article)
  lower = mean(lof_scores) - T * std(lof_scores)
  upper = mean(lof_scores) + T * std(lof_scores)
  outliers = [index_of(s) for s in lof_scores if s < lower or s > upper]
  if index_of(case_paragraph) in outliers then
    remove(case_paragraph, case_paragraphs)
  end
end

```

Algorithm 2: Paragraph removal if outlier pseudocode.

Following Alg. 1, paragraphs with opposite or contrary emotions to E are selected if their scores exceed a certain threshold, i.e. $T1$ for opposite emotions and $T2$ for contrary emotions. $T1$ is the mean value within the IQR of the prevalent emotion E , while $T2$ is the upper quartile of IQR. For example, in Fig. C.4a, we retrieve all the paragraphs with the opposite emotion *anger* whose scores exceed $T1$ and all the paragraphs with contrary emotion *joy*, *love*, *optimism* or *trust* whose scores exceed $T2$. These paragraphs are considered candidates for being outliers,

and they are analyzed following Alg. 2. Firstly, the entire article of the candidate outlier paragraphs P is retrieved, and the emotions of all the article's paragraphs are evaluated. If the score of the opposite/contrary emotions in P is not "common" compared to the scores of the entire article, it is considered as possibly anomalous. For example, Fig. C.4b show the boxplot of the scores of all the emotions of an article containing a candidate outlier paragraph, whose emotion (contrary or opposite) scores are highlighted as red points. If the emotion score of the candidate outlier paragraph is placed outside the IQR of the article boxplot, such paragraph is removed if it is anomalous according to the Local Outlier Factor Breunig et al. (2000). The local outlier factor uses local density estimate with k nearest neighbors. It identifies regions with similar density and points with significantly lower density, identifying outliers by comparing the object's local density with its neighbors. In our case, we considered the embedding of each paragraph obtained with Sentence Transformers,¹ specifically `all-mpnet-base-v1` since it has the best trade-off between maximum manageable text length, encoding speed and model size. The number of neighbors is dynamically calculated with respect to the total number of paragraphs in the current article, so as to ensure that the number of neighbors exceeds the minimum number of samples that a cluster must contain, but remains below the maximum number of neighboring samples that could potentially be local outliers. Following Alg. 2, a paragraph is removed if its emotion score exceeds the upper or lower bounds calculated starting from the mean value of the Local Outlier Factor scores. This approach ensures the robustness of our dataset by minimizing noise and enhancing the relevance of the analyzed content. This process resulted in the removal of about 100 paragraphs, which were manually checked to confirm that they were indeed outliers.

Additionally to the outlier removal process, we cleaned our labels by zeroing the values of morality and emotions with scores below a threshold (we empirically set it to 0.2). If, following this step, there were no more moral/emotional traits with values greater than zero, we assigned the corresponding paragraph the label of moral/emotional neutrality, with score 1. Considering the events, they passed through a lemmatization process, hence removing stop words, punctuation and digits. Then, we removed all repeated events within the same paragraph (i.e., same trigger word and same subjects involved) and all the paragraphs in which no event was detected.

C.7 ME²-BERT Notations

Frequently used symbols and names throughout the main paper are summarized in Table C.2.

¹ https://www.sbert.net/docs/sentence_transformer/pretrained_models.html

Notation	Description
$\mathcal{D}$	Generic dataset with pair of samples (d_i, y_i) .
y_i	Real-value scores associated to MFT values.
d_i	Generic text segment.
$\mathcal{T}$	Space of all possible input texts.
$\mathcal{D}_s$	Labeled source domain dataset.
$\mathcal{D}_t$	Unlabeled target domain dataset.
$d_i^{(s)}, d_j^{(t)}$	Text segment from the source and target domains, respectively.
n	Number of text segments.
n_s, n_t	Number of text segments of source and target domains, respectively.
$y^{(s)}, y^{(t)}$	Moral value scores for source and target domains, respectively.
y_i	Real-value vector for MFT dimensions.
f	Moral value classification function.
g	Domain classification function.
σ	Sigmoid activation function.
P, Q	Marginal and Conditional distributions, respectively.
ψ	Noise generated with Bernoulli distribution.
p	Probability of dropout.
$\odot$	Element-wise multiplication.
$\mathbf{x}_i, \mathbf{x}'_i$	BERT embedding of the i -th text segment, and its corrupted version, respectively.
$\mathbf{x}_i^{(oob)}$	Out-of-the-box BERT embedding
$\hat{\mathbf{x}}_i$	Reconstructed BERT embedding.
$\mathbf{h}_i$	Bottleneck representation learned by the encoder.
f_{enc}, f_{dec}	Encoder and decoder functions of E-DAE.
$\mathbf{h}_p, \mathbf{h}_r$	Positive and negative samples for the contrastive learning function.
λ_{dom}	Weighting parameter for adversarial loss.
$\mathcal{L}_{MSE}$	Mean Squared Error (MSE) loss for reconstruction.
$\mathcal{L}_{tr}^{(s)}, \mathcal{L}_{tr}^{(t)}$	Source and Target triplet margin losses for contrastive learning, respectively.
$\mathcal{L}_{con}$	Contrastive loss.
$\mathcal{L}_{MF}$	Cross-entropy loss for moral value prediction.
$\mathcal{L}_{ADV}$	Loss for adversarial domain classification.
$\mathcal{L}_{E-DAE}$	Loss of the E-DAE module.
$\mathcal{L}$	Overall Loss.
$\tilde{\mathbf{x}}_i$	Gated representation.
$\mathbf{f}_i, \mathbf{q}_i, \mathbf{o}_i$	Forget, input, and output gates.
$y_i^{(D)}$	Ground-truth label indicating which domain the input segment d_i belongs to.
BERT-E2MoCase	BERT model fine-tuned on E2MoCase dataset.
BERT-MFTC	BERT model fine-tuned on MFTC dataset.
BERT-MFRC	BERT model fine-tuned on MFRC dataset.
BERT-MFTRC	BERT model fine-tuned on the aggregation of the MFRC and MFTC datasets.
ME ² -BERT w/o A.	ME ² -BERT without the Adversarial Learning module.
ME ² -BERT w/o A-C.	ME ² -BERT without the Contrastive and Adversarial Learning modules.
ME ² -BERT w/o C.	ME ² -BERT without the Contrastive Learning module.
ME ² -BERT w/o G.	ME ² -BERT without the Gate mechanism.

Table C.2: Table of notations and model names

C.8 Experimental environment

We carried out all our experiments using a 8x NVIDIA A30 GPU server with 24 GB of RAM each, 764 GB of system RAM, a Double Intel Xeon Gold 6248R with a total of 96 cores, and Ubuntu Linux 20.04.6 LTS as operating system.