LEXTREME: A Multi-Lingual and Multi-Task Benchmark for the Legal Domain

Joel Niklaus [1,2,6*] Veton Matoshi [2*] Pooja Rani [3]

Andrea Galassi [4] Matthias Stürmer [1,2] Ilias Chalkidis [5]

[1]University of Bern [2]Bern University of Applied Sciences [3]University of Zurich
[4]University of Bologna [5]University of Copenhagen [6]Stanford University

Abstract

Lately, propelled by the phenomenal advances around the transformer architecture, the legal NLP field has enjoyed spectacular growth. To measure progress, well curated and challenging benchmarks are crucial. However, most benchmarks are English only and in legal NLP specifically there is no multilingual benchmark available yet. Additionally, many benchmarks are saturated, with the best models clearly outperforming the best humans and achieving near perfect scores. We survey the legal NLP literature and select 11 datasets covering 24 languages, creating LEXTREME. To provide a fair comparison, we propose two aggregate scores, one based on the datasets and one on the languages. The best baseline (XLM-R large) achieves both a dataset aggregate score a language aggregate score of 61.3. This indicates that LEXTREME is still very challenging and leaves ample room for improvement. To make it easy for researchers and practitioners to use, we release LEXTREME on huggingface together with all the code required to evaluate models and a public Weights and Biases project with all the runs.

1 Introduction

In the last decade, the discipline of Natural Language Processing (NLP) has become more and more relevant for Legal Artificial Intelligence, leading to a shift from symbolic to subsymbolic techniques (Villata et al., 2022). Such a change can be motivated partially by the nature of legal resources, which appear mostly in a textual format (legislation, legal proceedings, contracts, etc.).

Following closely the advances in the development of NLP technologies, the legal NLP literature (Zhong et al., 2020; Aletras et al., 2022; Katz et al., 2023) is flourishing with the release of many new resources, including large legal corpora (Henderson et al., 2022), task-specific datasets (Chalkidis et al., 2021a; Shen et al., 2022), and pre-trained legal-oriented language models (PLMs) (Chalkidis et al., 2020; Zheng et al., 2021; Xiao et al., 2021; Niklaus and Giofré, 2022).

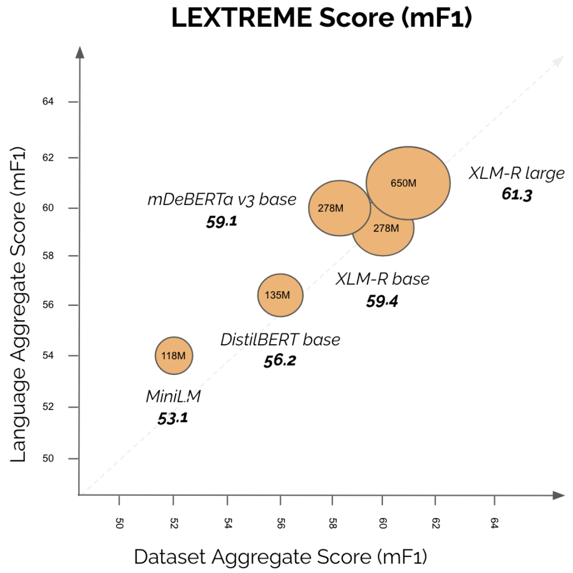

Figure 1: Overview of the multilingual models on the LEXTREME benchmark. The bubble size and text inside indicate the parameter count.

In particular, the development and spread of the so-called Foundation Models (Bommasani et al., 2022), large neural networks trained on vast corpora, led to massive performance improvements on popular benchmarks such as GLUE (Wang et al., 2019b) or SuperGLUE (Wang et al., 2019a). This exemplifies the need for more challenging benchmarks to continually measure progress. Legal benchmark suites (Chalkidis et al., 2022a; Hwang et al., 2022) to evaluate the performance of PLMs in a more systematic way have been also developed, showcasing the superiority of legal-oriented PLMs over generic ones on downstream tasks.

However, general-purpose models, trained on resources such as Wikipedia, may be insufficient to address tasks in the legal domain. Indeed, such a domain is strongly characterized both by its lexicon and by specific knowledge typically not available

[*] Equal contribution.

outside of specialized domain resources. Laypeople even sometimes call the language used in legal documents “legalese” or “legal jargon”, emphasizing its complexity. It is therefore necessary to develop specialized Legal Language Models, to be trained on large collections of legal documents, and to be evaluated on proper legal benchmarks.

Existing benchmarks, such as GLUE, often tackle linguistic tasks, such as semantic textual similarity or natural language inference, with no direct application in mind. There is a need for benchmarks that tackle use cases as close as possible to the real world to align model development with practical deployment needs.

The rising need to build NLP systems for languages different from English, the scarcity of textual resources for those languages and the spread of code-switching in many cultures (Torres Cacoullos, 2020) has pushed researchers to design new multilingual learning approaches. This, in turn, has brought the necessity to develop proper multilingual benchmarks to evaluate multilingual language models (Conneau et al., 2020). This is of paramount importance for legal NLP, especially in case of inherently multinational (European Union, Council of Europe), or multilingual (Canada, Switzerland) legal systems.

In this work, we propose a challenging multilingual benchmark for the legal domain containing datasets with valuable use cases, calling it LEXTREME. We survey the literature and select 11 datasets out of 108 papers based on our exclusion and inclusion criteria. We evaluate five popular multilingual encoder-based language models and find that model size correlates well with performance on LEXTREME. For easy evaluation, we release the aggregate dataset on the huggingface hub[1] and the code to run experiments on GitHub.[2]

Contributions

The contributions of this paper are two-fold:

1. We review the literature for suitable legal datasets and compile a multilingual legal benchmark of 11 datasets in 24 languages.

2. We evaluate various baselines on LEXTREME to provide a reference point for researchers and practitioners to compare to.

2 Related Work

2.1 Benchmarks for Language Models

GLUE (Wang et al., 2019b) is one of the first benchmarks for the evaluation of general-purpose neural language models. It is a set of supervised sentence understanding predictive tasks in the English language that was created through aggregation and curation of already existing datasets. GLUE became quickly obsolete with the advent of advanced contextual language models such as BERT (Devlin et al., 2019), which performed extremely well on most of them. SUPERGLUE (Wang et al., 2019a) was later proposed as an updated version of GLUE, including new predictive tasks that are solvable by humans but are difficult for machines. Both benchmarks proposed an evaluation computed as the aggregation of the scores obtained by the same model on each task. They are also agnostic regarding the pre-training of the model, and do not provide a specific corpus for it. Following this trend, many other benchmarks have been proposed, Table 1 provides an overview of the most popular ones.

MMLU (Hendrycks et al., 2021) is specifically designed to evaluate the knowledge acquired during pre-training of the model by including only zero-shot and few-shot learning tasks. It contains about 16K multiple-choice questions divided into 57 subtasks, covering subjects in the humanities, social sciences, hard sciences, and other areas.

SUPERB (Yang et al., 2021) and SUPERB-SG (Tsai et al., 2022) were proposed for speech data, unifying popular datasets. They mainly differ in SUPERB-SG not including only predictive tasks but also generative ones, a characteristic that makes it different from all the other benchmarks discussed in this section. Another important difference is that SUPERB-SG includes tasks such as speech translation and cross-lingual automatic speech recognition, for which knowledge of languages other than English is beneficial. Neither of the two proposes an aggregated score.

XTREME (Hu et al., 2020) is a benchmark specifically designed to evaluate the ability of cross-lingual generalization of models. It includes 6 cross-lingual predictive tasks over 10 datasets of miscellaneous texts, covering a total of 40 languages. While some original datasets were already designed for cross-lingual tasks, others were extended by translating part of the data through human professionals and automatic methods.

[1] https://huggingface.co/datasets/joelito/lextreme

[2] https://github.com/JoelNiklaus/LEXTREME

Name	Source	Domain	Tasks	Datasets	Languages	Agg. Score
GLUE	(Wang et al., 2019b)	Misc. Texts	7	9	English	Yes
SUPERGLUE	(Wang et al., 2019a)	Misc. Texts	8	8	English	Yes
CLUE	(Xu et al., 2020)	Misc. Texts	9	9	Chinese	Yes
XTREME	(Hu et al., 2020)	Misc. Texts	6	9	40	Yes
BLUE	(Peng et al., 2019)	Biomedical Texts	5	10	English	Yes
CBLUE	(Zhang et al., 2022)	Biomedical Texts	9	9	Chinese	Yes
MMLU	(Hendrycks et al., 2021)	Misc. Texts	1	57	English	Yes
LexGLUE	(Chalkidis et al., 2022b)	Legal Texts	7	6	English	Yes
LBOX	(Hwang et al., 2022)	Legal Texts	5	5	Korean	Yes
LEXTREME	(our work)	Legal Texts	18	11	25	Yes
SUPERB	(Yang et al., 2021)	Speech	10	10	English	No
SUPERB-SG	(Tsai et al., 2022)	Speech	5	5	English	No
TAPE	(Rao et al., 2019)	Proteins	5	5	n/a	No

Table 1: Characteristics of popular existing NLP benchmarks.

2.2 Benchmarks for Legal Language Models

LEXGLUE (Chalkidis et al., 2022b) is the first benchmark for the legal domain and covers 6 predictive tasks over 5 datasets made of textual documents in English from the US, EU, and CoE. While some tasks may not require specific legal knowledge to be solved, others would probably need, or at least benefit from, information regarding the EU or US legislation on the specific topic. Among the main limitations of their benchmark, Chalkidis et al. highlight its monolingual nature and remark that "*there is an increasing need for developing models for other languages*". Our work is strongly inspired by LEXGLUE and our purpose is to propose a benchmark that, we hope, will help the development of multilingual models for the legal domain.

In a similar direction, Hwang et al. (2022) released the LBOX benchmark. It covers 3 downstream tasks: two legal judgement prediction (LJP) tasks, and one summarization task in Korean.

The LEGALBENCH initiative (Guha et al., 2022) aims to create an open and collaborative legal reasoning benchmark where legal practitioners and other domain experts can contribute by submitting tasks that will be addressed using language models. At its creation, the authors have already added 44 lightweight tasks. While most tasks require legal reasoning based on the common law system, there is also a clause classification task.

Concerning language models specifically trained for the legal domain, many have been proposed for specific languages but, to the best of our knowledge, no multilingual model has been proposed yet. Legal language models have been proposed for English (Chalkidis et al., 2020; Ying and Habernal, 2022), French (Douka et al., 2021), Romanian (Masala et al., 2021), Italian (Tagarelli and Simeri, 2022; Licari and Comandè, 2022), Chinese (Xiao et al., 2021), Arabic (Al-Qurishi et al., 2022), Korean (Hwang et al., 2022), and Portuguese (Ciurlino, 2021). For an overview of the many tasks related to the automatic analysis of legal texts, we suggest reading the works of Chalkidis et al. (2022b) and Zhong et al. (2020).

3 LEXTREME Tasks and Datasets

3.1 LEXTREME Dataset Selection

To select the datasets for the LEXTREME benchmark, we formulate various criteria. We first systematically explore the literature via the ACL anthologyto find relevant datasets for the legal domain. We identify various venues, such as ACL, EACL, NAACL, EMNLP, LREC, ICAIL, and the NLLP workshop. We search the literature of these venues for the years 2010 to 2022. We search for some common keywords (case insensitive) that are related to legal datasets, e.g., *criminal*, *judicial*, *judgment*, *jurisdictions*, *law*, *legal*, *legislation*, *dataset*, and *corpus*. These keywords help to select potentially relevant papers, i.e., 108 papers. Then, three authors analyze these papers based on the inclusion and exclusion criteria given below to ensure that they indeed propose a legal dataset.

Inclusion criteria

I1: It is about legal text (e.g., patents are not considered part of legal text), I2: It performs legal tasks (e.g., judgment prediction) and not other linguistic tasks such as Part-of-Speech (POS) tagging,

I3: It performs NLU tasks (e.g., information retrieval tasks are not considered due to their evaluation complexity),

I4: The tasks are in one of the European languages (e.g., China has its own large legal Natural Language Processing (NLP) community and likely would not benefit much from multilingual models), and

I5: The dataset is annotated by humans directly or indirectly (e.g., judgement labels are extracted with regexes)

Exclusion criteria

E1: The dataset is not publicly available,

E2: The dataset does not contain a public license,

E3: The dataset contains labels that are generated with ML systems.

E4: It is not a peer-reviewed paper

Task	# Examples	# Labels
BCD-J	3234 / 404 / 405	3 / 3 / 3
BCD-U	1715 / 211 / 204	2 / 2 / 2
GAM	19271 / 2726 / 3078	4 / 4 / 4
GLC-V	28536 / 9511 / 9516	47 / 47 / 47
GLC-C	28536 / 9511 / 9516	386 / 377 / 374
GLC-S	28536 / 9511 / 9516	2143 / 1679 / 1685
SJP	59709 / 8208 / 17357	2 / 2 / 2
OTS-UL	2074 / 191 / 417	3 / 3 / 3
OTS-CT	19942 / 1690 / 4297	9 / 8 / 9
C19	3312 / 418 / 418	8 / 8 / 8
MEU-1	817239 / 112500 / 115000	21 / 21 / 21
MEU-2	817239 / 112500 / 115000	127 / 126 / 127
MEU-3	817239 / 112500 / 115000	500 / 454 / 465
GLN	17699 / 4909 / 4017	17 / 17 / 17
LNR	7552 / 966 / 907	11 / 9 / 11
LNB	7828 / 1177 / 1390	13 / 13 / 13
MAP-C	27823 / 3354 / 10590	13 / 11 / 11
MAP-F	27823 / 3354 / 10590	44 / 26 / 34

Table 2: Overview of datasets and their tasks. The fields *# Examples* and *# Labels* provide the values for the splits train, validation, test. For a detailed overview of for the language-specific subsets of each multilingual task, see Table 7 and 8.

After applying the above criteria, we reduce from 108 to 11 datasets. We provide the list of all these datasets in the online repository.[3]

Dataset	Jurisdiction	Languages
BCD	BR	pt
GAM	DE	de
GLC	GR	el
SJP	CH	de, fr, it
OTS	EU	de, en, it, pl
C19	BE, FR, HU, IT, NL, PL, UK	en, fr, hu, it, nb, nl, pl
MEU	EU	24 EU langs
GLN	GR	el
LNR	RO	ro
LNB	BR	pt
MAP	EU	24 EU langs

Table 3: Overview of datasets and the jurisdiction as well as the languages that they cover. The 24 EU languages are: bg, cs, da, de, el, en, es, et, fi, fr, ga, hu, it, lt, lv, mt, nl, pt, ro, sk, sv

3.2 LEXTREME Tasks

LEXTREME constist of three classification task types: Single Label Text Classification (SLTC), Multi Label Text Classification (MLTC), and Named Entity Recognition (NER). We use the existing train, validation, and test splits if present. In the other cases we split the data ourselves (80% train, 10% validation and test each). In the following, we briefly describe the selected datasets. For more information about the number of examples and label classes per split for each task, see Table 2, 7 and 8. For a detailed overview of the jurisdictions as well as the number of languages covered by each dataset, see Table 3.

3.3 LEXTREME Datasets

Each dataset can be either monolingual or multilingual and can have several configurations or (fine-tuning) tasks, which are the basis of our analyses, i.e., the pretrained models have always been fine-tuned on a single task.

Brazilian Court Decisions (BCD) Legal systems are often huge and complex, and the information is scattered across various sources. Thus, predicting case outcomes from multiple vast volumes of litigation is a difficult task. Lage-Freitas et al. (2022) propose an approach to predict Brazilian legal decisions to support legal practitioners. We use their dataset from the State Supreme Court of Alagoas (Brazil). The input to the models is always the case description. We perform two SLTC

[3] https://github.com/JoelNiklaus/LEXTREME

tasks: One (BCD-J) is to predict the approval or dismissal of the case or appeal with three labels *no, partial, yes*, and another (BCD-U) is to predict the unanimity on the decision alongside two labels *unanimity, not-unanimity*.

German Argument Mining (GAM) Identifying arguments in court decisions is an important and challenging task for legal practitioners. Urchs. et al. (2021) compiled a dataset of 200 German court decisions for classifying sentences according to their argumentative function. We use their dataset to perform an MLTC task. The input to the models is a sentence and the output is labeled according to four categories: *conclusion, definition, subsumption, other*.

Greek Legal Code (GLC) Legal documents can cover a wide variety of topics, which makes accurate topic classification all the more important. Papaloukas et al. (2021) compiled a dataset for topic classification of Greek legislation documents. The documents cover 47 main thematic topics which are called *volumes*. Each of them is divided into thematic sub categories which are called *chapters* and subsequently, each chapter breaks down to *subjects*. Therefore, the dataset is used to perform three different SLTC tasks along volume level (GLC-V), chapter level (GLC-C), and subject level (GLC-S). The input to the models is the entire document, and the output is one of the several topic categories.

Swiss Judgment Prediction (SJP) Niklaus et al. (2021, 2022b), focus on predicting the judgment outcome of the cases from the Swiss Federal Supreme Court (FSCS). We use their dataset of 85k cases. The input to the models is the appeal description, and the output is whether the appeal is approved or dismissed. It is also a SLTC task.

Online Terms of Service (OTS) While the benefits of multilingualism in the EU legal world are well known, creating an official version of every legal act in 24 languages raises interpretative challenges. Drawzeski et al. (2021), attempt to automatically detect unfair clauses in Terms of Service. We use their dataset of 100 contracts to perform a SLTC and MLTC task. In the SLTC task (OTS-UL), the input to the models is a sentence, and the output presents the sentence classified into three levels of unfairness. In the MLTC task (OTS-CT), the model identifies the sentence for various clause topics.

COVID19 Emergency Event (C19) The COVID-19 pandemic showed various exceptional measures governments around the world have taken to contain the virus. Tziafas et al. (2021), presented a dataset, also known as EXCEPTIUS, that contains legal documents with sentence-level annotation from several European countries to automatically identify the measures. We use their dataset to perform only one task, i.e., the MLTC task of identifying the type of measure described in a sentence. The input to the models are the sentences, and the output is neither or at least one of the measurement types.

MultiEURLEX (MEU) Multilingual transfer learning has gained significant attention recently due to its increasing applications in NLP tasks. Chalkidis et al. (2021b), explored the cross-lingual transfer for legal NLP and presented a corpus of 65K EU laws. They annotated each law document with multiple labels from the EUROVOC taxonomy. We perform a MLTC task to identify labels (given in the taxonomy) for each document. Since the taxonomy exists on multiple levels, we prepare configurations according to three levels (MEU-1, MEU-2, MEU-2).

Greek Legal NER (GLN) Identifying various named entities from natural language text plays an important role for Natural Language Understanding (NLU). Papaloukas et al. (2021) compiled an annotated dataset for NER in Greek legal documents. The source material are 254 daily issues of the Greek Government Gazette over the period 2000-2017. In *all* NER tasks of LEXTREME the input to the models is the list of tokens, and the output is an entity label for each token.

LegalNERo (LNR) Similar to GLN, Pais et al. (2021) manually annotated Romanian legal documents for various named entities. The dataset is derived from 370 documents from the larger MARCELL Romanian legislative subcorpus[4].

LeNER BR (LNB) Luz de Araujo et al. (2018) compiled a dataset for NER for Brazilian legal documents. To compose the dataset, 66 legal documents from several Brazilian Courts were collected. Additionally, four legislation documents were collected, resulting a total of 70 documents that were annotated for named entities.

[4] https://marcell-project.eu/deliverables.html

Model	Source	Params	Vocab	Specs	Corpora	# Langs
MiniLM	Wang et al. (2020)	118M	250K	1M steps / BS 256	2.5T CC100 data	100
DistilBert	Sanh et al. (2019)	135M	120K	BS up to 4000	Wikipedia	104
mDeberta-v3	He et al. (2020, 2021)	278M	128K	500K steps / BS 8192	2.5T CC100 data	100
XLM-R base	Conneau et al. (2020)	278M	250K	1.5M steps / BS 8192	2.5T CC100 data	100
XLM-R large	Conneau et al. (2020)	560M	250K	1.5M steps / BS 8192	2.5T CC100 data	100

Table 4: Multilingual Models: All models can process up to 512 tokens. BS is short for batch size. Params is the total number of parameters (including the embedding layer).

MAPA (MAP) de Gibert et al. (2022), built a multilingual corpus based on EUR-Lex (Baisa et al., 2016) for NER. The dataset comes in two configurations, i.e., two NER tasks, as it has been annotated at a coarse-grained (MAP-C) and fine-grained (MAP-F) level. The structure of the dataset is the same as the other datasets for NER.

4 Models Considered

Since our benchmark only contains NLU tasks, we consider encoder only models for simplicity.

MiniLM MiniLM (Wang et al., 2020) is the result of a novel task-agnostic compression technique, also called distillation, in which a compact model — the so-called student — is trained to reproduce the behaviour of a larger pre-trained model — the so-called teacher. This is achieved by deep self-attention distillation, i.e. only the self-attention module of the last Transformer layer of the teacher, which stores a lot of contextual information (Jawahar et al., 2019), is distilled. The student is trained by closely imitating the teacher's final Transformer layer's self-attention behavior. To aid the learner in developing a better imitation, (Wang et al., 2020) also introduce the self-attention value-relation transfer in addition to the self-attention distributions. The addition of a teacher assistant results in further improvements. For the training of multilingual MiniLM, XLM-R$_{\text{BASE}}$ was used.

DistilBERT DistilBERT (Sanh et al., 2019) is a more compressed version of BERT (Devlin et al., 2019) using teacher-student learning, similar to MiniLM. DistilBERT is distilled from BERT, thus both share a similar overall architecture. The pooler and token-type embeddings are eliminated, and the number of layers is decreased by a factor of 2 in DistilBERT. DistilBERT is distilled in very large batches while utilizing gradient accumulation and dynamic masking, but without the next sentence prediction objective. DistilBERT was trained on the same corpus as the original BERT.

mDEBERTa He et al. (2020) suggest a new model architecture called DeBERTa (Decoding-enhanced BERT with disentangled attention), which employs two novel methods to improve the BERT and RoBERTa models. The first is the disentangled attention mechanism, in which each word is represented by two vectors that encode its content and position, respectively, and the attention weights between words are calculated using disentangled matrices on their respective contents and relative positions. To predict the masked tokens during pre-training, an enhanced mask decoder is utilized, which incorporates absolute positions in the decoding layer. Additionally, the generalization of models is enhanced through fine-tuning using a new virtual adversarial training technique. He et al. (2021) introduce mDEBERTa-v3 by further improving the efficiency of pre-training by replacing Masked-Language Modeling (MLM) in DeBERTa with the task of replaced token detection (RTD) where the model is trained to predict whether a token in the corrupted input is either original or replaced by agenerator. Further improvements are achieved via *gradient-disentangled embedding sharing* (GDES).

XLM-RoBERTa XLM-R (Conneau et al., 2020) is a multilingual language model which has the same pretraining objectives as RoBERTa (Liu et al., 2019), such as dynamic masking, but not next sentence prediction. It is pre-trained on a large corpus comprising 100 languages. The authors report a significant performance gain over multilingual BERT (mBERT) in a variety of tasks with results competitive with state-of-the-art monolingual models (Conneau et al., 2020).

4.1 Hierarchical Variants

A significant part of the datasets consists of very long documents, the best examples being all variants of MultiEURLEX, cf. Figure 12. However, Transformer-based models usually allow a maximum input length of 512 tokens. It is possible to use the models without further ado for documents that exceed this length by far. However, this can only be achieved by a massive truncation of the original document. This procedure has the consequence that only the first section of a document is available for classification tasks. This is the reason why we used hierarchical variants of pretraining models for finetuning on data sets with particularly long documents (cf. histograms).

The hierarchical variants used in the study are broadly equivalent to those in (Chalkidis et al., 2021c; Niklaus et al., 2022a). First, we convert each document into a list of equal-length paragraphs. Afterward, we use a pre-trained Transformer-based model to encode each of these paragraphs separately and to obtain the [CLS] embedding of each paragraph which can be used as a context-unaware paragraph representation. In order to make them context-aware, i.e. aware of the surrounding paragraphs, the paragraph representations are fed into a 2-layered Transformer encoder with varying specifications depending on the model type. Finally, max-pooling over the context-aware paragraph representations is deployed, which results in a document representation that is fed to a classification layer.

5 Experimental Setup

Some datasets were highly imbalanced, one of the best examples being BCD-U with a proportion of the minority class of about 2%. Therefore, we applied random oversampling on all tasks of the SLTC datasets, except for GLC, since all its subsets have too many labels, which would have led to a drastic increase in the data size and thus in the computational costs for finetuning. For each run, we used the same hyperparameters, as described in Section A.2.

As described in section 4.1, some tasks contain very long documents, which required the usage of hierarchical variants with sequence lengths that go beyond 512. Based on the distribution of the sequence length per example for each task (cf. section D), we decided on suitable sequence lengths for each task before finetuning. A list of suitable sequence lengths can be found in A.1. Tasks with a maximum sequence length of over 512 required the usage of hierarchical variants.

Evaluation Metrics We use the macro-F1 score for all datasets to ensure comparability across the entire benchmark, since it can be computed for both text classification and NER tasks. Mathew's Correlation Coefficient (MCC) is a suitable score for evaluating text classification tasks but its applicability to NER tasks is unclear. For brevity, we do not display additional scores, but more detailed (such as precision and recall, and scores per seed) and additional scores (such as MCC) can be found online on our Weights and Biases project[5].

Aggregate Score We acknowledge that the datasets included in LEXTREME are diverse and hard to compare due to variations in the number of samples and task complexity (Raji et al., 2021a). This is why we always report the scores for each dataset subset, enabling a fine-grained analysis. However, we believe that by taking the following three measures, an aggregate score can provide more benefits than drawbacks, encouraging the community to evaluate multilingual legal models on a curated benchmark facilitating comparisons.

We (a) evaluate all datasets with the same score (macro-F1) making aggregation more intuitive and easier to interpret, (b) aggregating the F1 scores again using the harmonic mean, since F1 scores are already rates and obtained using the harmonic mean over precision and recall, following Tatiana and Valentin (2021), and (c) basing our final aggregate score on two intermediate aggregate scores — the dataset aggregate and language aggregate score – thus weighing datasets and languages equally promoting model fairness and robustness.

The final LEXTREME score is computed using the harmonic mean of the dataset and the language aggregate score. We compute the dataset aggregate score by taking the successive harmonic mean of (1.) the languages inside the configurations (e.g., de,fr,it within SJP), (2.) the configurations inside the datasets (e.g., OTS-UL, OTS-CT within OTS), and (3.) the datasets inside LEXTREME (BCD, GAM, etc.). We compute the language aggregate score by taking the successive harmonic mean of (1.) the configurations inside the datasets, (2.) the datasets for the given language (e.g., MAP and MEU for lv), and (3.) the languages inside LEXTREME (bg,cs, etc.).

[5] https://wandb.ai/lextreme/paper_results

Model	**BCD**	**GAM**	**GLC**	**SJP**	**OTS**	**C19**	**MEU**	**GLN**	**LNR**	**LNB**	**MAP**	**Agg.**
MiniLM	53.0	**73.3**	42.1	67.7	44.1	2.6	62.0	40.5	46.8	86.0	55.5	52.2
DistilBERT	54.5	69.5	**62.8**	66.8	56.1	22.2	63.6	38.1	48.4	78.7	55.0	56.0
mDeBERTa v3	57.6	70.9	52.2	69.1	66.5	25.5	65.1	42.2	46.6	87.8	**60.2**	58.5
XLM-R base	**63.5**	72.0	56.8	**69.3**	67.8	26.4	65.6	47.0	47.7	86.0	56.1	59.9
XLM-R large	58.7	73.1	57.4	69.0	**75.0**	**29.0**	**68.1**	**48.0**	**49.5**	**88.2**	58.5	**61.3**

Table 5: Dataset aggregate scores for multilingual models. The best scores are in bold.

Model	**bg**	**cs**	**da**	**de**	**el**	**en**	**es**	**et**	**fi**	**fr**	**ga**	**hr**	**hu**	**it**	**lt**	**lv**	**mt**	**nl**	**pl**	**pt**	**ro**	**sk**	**sl**	**sv**	**Agg.**
MiniLM	64.0	57.7	55.4	60.1	48.9	42.8	63.8	59.7	56.6	48.5	41.5	62.2	41.8	45.6	59.8	60.2	55.7	38.8	33.5	63.5	58.4	58.9	62.2	59.4	54.1
DistilBERT	65.3	60.2	57.4	64.1	53.1	54.0	66.9	57.4	55.7	55.8	45.5	63.1	39.9	54.9	58.0	57.7	57.3	42.0	43.6	64.7	57.4	59.0	63.3	59.2	56.5
mDeBERTa v3	61.9	60.6	59.3	66.6	54.0	58.9	66.9	60.3	**61.1**	57.0	50.2	65.0	44.2	59.7	63.7	61.4	**61.2**	**48.1**	50.2	67.9	**60.8**	**65.2**	65.2	**65.4**	59.8
XLM-R base	**68.3**	61.3	58.5	66.0	54.7	58.6	63.8	59.3	57.5	57.7	47.8	65.9	43.3	59.6	60.3	60.8	58.0	45.0	52.0	**68.2**	59.2	60.3	66.2	61.7	58.9
XLM-R large	64.5	**63.3**	**65.1**	**68.3**	**59.6**	**61.9**	**70.0**	**61.3**	60.9	**57.9**	**50.3**	**68.3**	**44.7**	**62.9**	**66.1**	**65.5**	60.1	43.9	**55.0**	68.1	60.2	62.8	**68.2**	62.5	**61.3**

Table 6: Language aggregate scores for multilingual models. The best scores are in bold.

6 Results

In this section, we discuss the main result of our evaluation of the baseline models. Scores on the validation datasets and standard deviations across seeds can be found in Appendix C.

We show the dataset and language aggregated results in Tables 5 and 6 respectively. For both the dataset aggregate and the language aggregate scores, we see a clear trend that larger models perform better. However, when looking at the individual datasets and languages, the scores are more erratic. We notice that on some datasets, such as C19, GLC or OTS, the models vary greatly, with differences as large as 29.2 between the worst performing MiniLM and the best performing XLM-R large. MiniLM seems to struggle greatly with these three datasets, while even achieving the best performance on GAM. On other datasets, such as SJP, MEU, LNR, and MAP the models are very close together (6 points or fewer between best and worst model). SJP, MEU and MAP are the largest datasets in LEXTREME, thus probably decreasing the influence of the pretraining on downstream performance and leveling the playing field. LNR, however, is the smallest NER task, opposing this hypothesis. In contrast to the inconsistent results on the datasets, we notice, that XLM-R performs best on most languages. Additionally, we note that the variability of the models within a language is similar to the variability within a dataset, however, we don't see extreme cases such as GLC or OTS.

7 Conclusions and Future Work

Conclusions

We survey the literature and select 11 datasets out of 108 papers with rigorous criteria to compile the first multilingual benchmark for legal NLP. By open-sourcing both the dataset and the code, we invite researchers and practitioners to evaluate any future multilingual models on our benchmark. We provide baselines for five popular multilingual encoder-based language models of different sizes. We hope that this benchmark will foster the creation of novel legal multilanguage models and therefore contribute to the progress of natural legal language processing. We imagine this work as a living benchmark and invite the community to extend it with new suitable datasets.

Future Work

In future work, we will extend this benchmark with other NLU tasks and also generation tasks such as summarization, simplification, or translation. Another avenue of future work can be the extension with datasets in more languages or from jurisdictions not yet covered in the current version. Finally, we leave the evaluation of other models such as mT5 (Xue et al., 2021) to future work.

Limitations

It is important to not exceed with the enthusiasm for language models and the ambitions of benchmarks: many recent works have addressed the limits of these tools and analyzed the consequences of their misuses. For example, Bender and Koller (2020) argue that language models do not really learn "meaning". Koch et al. (2021) evaluate the use of datasets inside scientific communities and highlight that many machine learning communities focus on very few datasets and that often these dataset are "borrowed" from other communities. Raji et al. (2021b) offer a detailed exploration of the limits of popular "general" benchmarks, such as

GLUE (Wang et al., 2019b) and ImageNET (Deng et al., 2009). Their analysis covers 3 aspects: limited task design, de-contextualized data and performance reporting, inappropriate community use.

The first problem concerns the fact that typically tasks are not chosen considering proper theories and selecting what would be needed to prove generality. Instead, they are limited to what is considered interesting by the community, what is available, or other similar criteria. These considerations hold also for our work. Therefore, we can not claim that our benchmark can be used to assess the "generality" of a model or proving that it "understands natural legal language".

The second point address the fact that any task, data, or metric are limited to their context, therefore "data benchmarks are closed and inherently subjective, localized constructions". In particular, the content of the data can be too different from real data and the format of the tasks can be too homogeneous compared to human activities. Moreover, any dataset inherently contains biases. We tackle this limitation by deciding to include only tasks and data that are based on real world scenarios, in an effort to minimize the difference between the performance of a model on our benchmark and its performance on a real world problem.

The last aspect regards the negative consequences that benchmarks can have. The competitive testing may encourage misbehavior and the aggregated performance evaluation does create a mirage of cross-domain comparability. The presence of popular benchmarks can influence a scientific community up to the point of steering towards techniques that perform well on that specific benchmark, in disfavor of those that do not. Finally, benchmarks can be misused in marketing to promote commercial products while hiding their flaws. Since these behaviour obviously can not be forecasted in advance, but we hope that this analysis of the shortcomings of our work will be sufficient to prevent misuses of our benchmark and will also inspire research directions for complementary future works. For what specifically concerns aggregated evaluations, they provide an intuitive but imprecise understanding of the performance of a model. While we do not deny their potential downsides, we believe that their responsible use is beneficial, especially when compared to the evaluation of a model on only an arbitrarily selected set of datasets. Therefore, we have decided to provide an aggregated performance evaluation and to weight languages and tasks equally to make it as robust and fair as possible.

It is important to remark that while Raji et al. and Koch et al. argument against the misrepresentations and the misuses of benchmarks and datasets, they do not argue against their usefulness. On the contrary, they consider the creation and adoption of novel benchmarks a sign of a healthy scientific community.

Ethics Statement

The scope of this work is to release a unified multilingual legal NLP benchmark to accelerate the development and evaluation of multilingual legal language models. A transparent multilingual and multinational benchmark for NLP in the legal domain might serve as an orientation for scholars and industry researchers by broadening the discussion and helping practitioners to build assisting technology for legal professionals and laypersons. We believe that this is an important application field, where research should be conducted (Tsarapatsanis and Aletras, 2021) to improve legal services and democratize law, while also highlight (inform the audience on) the various multi-aspect shortcomings seeking a responsible and ethical (fair) deployment of legal-oriented technologies.

Nonetheless, irresponsible use (deployment) of such technology is a plausible risk, as in any other application (e.g., online content moderation) and domain (e.g., medical). We believe that similar technologies should only be deployed to assist human experts (e.g., legal scholars in research, or legal professionals in forecasting or assessing legal case complexity) with notices on their limitations.

All datasets included in LEXTREME, are publicly available and have been previously published. We referenced the original work and encourage LEXTREME users to do so as well. In fact, we believe this work should only be referenced, in addition to citing the original work, when experimenting with multiple LEXTREME datasets and using the LEXTREME evaluation infrastructure. Otherwise, only the original work should be cited.

References

Muhammad Al-Qurishi, Sarah AlQaseemi, and Riad Soussi. 2022. Aralegal-bert: A pretrained language model for arabic legal text. In *NLLP*.

Nikolaos Aletras, Leslie Barrett, Catalina Chalkidis Ilias Goanta, and Daniel Preotiuc-Pietro, editors. 2022. *Proceedings of the Natural Legal Language Processing Workshop 2022*. Association for Computational Linguistics, Abu Dhabi, UAE.

Vít Baisa, Jan Michelfeit, Marek Medved', and Miloš Jakubíček. 2016. European Union language resources in Sketch Engine. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2799–2803, Portorož, Slovenia. European Language Resources Association (ELRA).

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: on meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 5185–5198. Association for Computational Linguistics.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. 2022. On the Opportunities and Risks of Foundation Models. ArXiv:2108.07258 [cs].

Ilias Chalkidis, Manos Fergadiotis, and Ion Androutsopoulos. 2021a. MultiEURLEX - a multi-lingual and multi-label legal document classification dataset for zero-shot cross-lingual transfer. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6974–6996, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ilias Chalkidis, Manos Fergadiotis, and Ion Androutsopoulos. 2021b. Multieurlex–a multi-lingual and multi-label legal document classification dataset for zero-shot cross-lingual transfer. *arXiv preprint arXiv:2109.00904*. Dataset URL: https://huggingface.co/datasets/multi_eurlex.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904, Online. Association for Computational Linguistics.

Ilias Chalkidis, Manos Fergadiotis, Dimitrios Tsarapatsanis, Nikolaos Aletras, Ion Androutsopoulos, and Prodromos Malakasiotis. 2021c. Paragraph-level rationale extraction through regularization: A case study on European court of human rights cases. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 226–241, Online. Association for Computational Linguistics.

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Katz, and Nikolaos Aletras. 2022a. LexGLUE: A benchmark dataset for legal language understanding in English. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4310–4330, Dublin, Ireland. Association for Computational Linguistics.

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael J. Bommarito II, Ion Androutsopoulos, Daniel Martin Katz, and Nikolaos Aletras. 2022b. Lexglue: A benchmark dataset for legal language understanding in english. In *ACL (1)*, pages 4310–4330. Association for Computational Linguistics.

Victor Hugo Ciurlino. 2021. Bertbr: a pretrained language model for law texts. Master's thesis, Universidade de Brasília.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Ona de Gibert, A García-Pablos, Montse Cuadros, and Maite Melero. 2022. Spanish datasets for sensitive entity detection in the legal domain. In

Proceedings of the Thirteenth International Conference on Language Resources and Evaluation (LREC'22), Marseille, France, june. European Language Resource Association (ELRA). Dataset URL: https://tinyurl.com/mv65cp66.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pages 248–255. IEEE Computer Society.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Stella Douka, Hadi Abdine, Michalis Vazirgiannis, Rajaa El Hamdani, and David Restrepo Amariles. 2021. JuriBERT: A masked-language model adaptation for French legal text. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 95–101, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kasper Drawzeski, Andrea Galassi, Agnieszka Jablonowska, Francesca Lagioia, Marco Lippi, Hans Wolfgang Micklitz, Giovanni Sartor, Giacomo Tagiuri, and Paolo Torroni. 2021. A corpus for multilingual analysis of online terms of service. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 1–8. Dataset URL: https://claudette.eui.eu/corpus_multilingual_NLLP2021.zip.

Neel Guha, Daniel E. Ho, Julian Nyarko, and Christopher Ré. 2022. Legalbench: Prototyping a collaborative benchmark for legal reasoning. *CoRR*, abs/2209.06120.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing. pages 1–17.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *ArXiv*, abs/2006.03654.

Peter Henderson, Mark S. Krass, Lucia Zheng, Neel Guha, Christopher D. Manning, Dan Jurafsky, and Daniel E. Ho. 2022. Pile of Law: Learning Responsible Data Filtering from the Law and a 256GB Open-Source Legal Dataset. ArXiv:2207.00220 [cs].

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *ICML*, volume 119 of *Proceedings of Machine Learning Research*, pages 4411–4421. PMLR.

Wonseok Hwang, Dongjun Lee, Kyoungyeon Cho, Hanuhl Lee, and Minjoon Seo. 2022. A multi-task benchmark for korean legal language understanding and judgement prediction. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.

Daniel Martin Katz, Dirk Hartung, Lauritz Gerlach, Abhik Jana, and Michael James Bommarito. 2023. Natural Language Processing in the Legal Domain.

Bernard Koch, Emily Denton, Alex Hanna, and Jacob G. Foster. 2021. Reduced, reused and recycled: The life of a dataset in machine learning research. In *NeurIPS Datasets and Benchmarks*.

André Lage-Freitas, Héctor Allende-Cid, Orivaldo Santana, and Lívia Oliveira-Lage. 2022. Predicting brazilian court decisions. *PeerJ Computer Science*, 8:e904. Dataset URL: https://github.com/proflage/predicting-brazilian-court-decisions.

Daniele Licari and Giovanni Comandè. 2022. Italian-legal-bert: A pre-trained transformer language model for italian law. In *EKAW-C*, volume 3256. CEUR-WS.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. (1).

Pedro Henrique Luz de Araujo, Teófilo E de Campos, Renato RR de Oliveira, Matheus Stauffer, Samuel Couto, and Paulo Bermejo. 2018. Lener-br: a dataset for named entity recognition in brazilian legal text. In *International Conference on Computational Processing of the Portuguese Language*, pages 313–323. Springer. Dataset URL: https://huggingface.co/datasets/lener_br.

Mihai Masala, Radu Cristian Alexandru Iacob, Ana Sabina Uban, Marina Cidota, Horia Velicu, Traian Rebedea, and Marius Popescu. 2021. jurBERT: A Romanian BERT model for legal judgement prediction. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 86–94, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Joel Niklaus, Ilias Chalkidis, and Matthias Stürmer. 2021. Swiss-judgment-prediction: a multilingual legal judgment prediction benchmark. *arXiv preprint arXiv:2110.00806*. Dataset URL: https://huggingface.co/datasets/swiss_judgment_prediction.

Joel Niklaus and Daniele Giofré. 2022. BudgetLongformer: Can we Cheaply Pretrain a SotA Legal Language Model From Scratch? ArXiv:2211.17135 [cs].

Joel Niklaus, Matthias Stürmer, and Ilias Chalkidis. 2022a. An empirical study on cross-X transfer for legal judgment prediction. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 32–46, Online only. Association for Computational Linguistics.

Joel Niklaus, Matthias Stürmer, and Ilias Chalkidis. 2022b. An Empirical Study on Cross-X Transfer for Legal Judgment Prediction. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 32–46, Online only. Association for Computational Linguistics.

Vasile Pais, Maria Mitrofan, Carol Luca Gasan, Vlad Coneschi, and Alexandru Ianov. 2021. Named entity recognition in the Romanian legal domain. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 9–18, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Christos Papaloukas, Ilias Chalkidis, Konstantinos Athinaios, Despina-Athanasia Pantazi, and Manolis Koubarakis. 2021. Multi-granular legal topic classification on greek legislation. *arXiv preprint arXiv:2109.15298*. Dataset URL: https://huggingface.co/datasets/greek_legal_code.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer learning in biomedical natural language processing: An evaluation of BERT and elmo on ten benchmarking datasets. In *BioNLP@ACL*, pages 58–65. Association for Computational Linguistics.

Inioluwa Deborah Raji, Emily M. Bender, Amandalynne Paullada, Emily Denton, and Alex Hanna. 2021a. AI and the Everything in the Whole Wide World Benchmark. ArXiv:2111.15366 [cs].

Inioluwa Deborah Raji, Emily Denton, Emily M. Bender, Alex Hanna, and Amandalynne Paullada. 2021b. AI and the everything in the whole wide world benchmark. In *NeurIPS Datasets and Benchmarks*.

Roshan Rao, Nicholas Bhattacharya, Neil Thomas, Yan Duan, Xi Chen, John F. Canny, Pieter Abbeel, and Yun S. Song. 2019. Evaluating protein transfer learning with TAPE. In *NeurIPS*, pages 9686–9698.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Zejiang Shen, Kyle Lo, Lauren Yu, Nathan Dahlberg, Margo Schlanger, and Doug Downey. 2022. Multilexsum: Real-world summaries of civil rights lawsuits at multiple granularities. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Andrea Tagarelli and Andrea Simeri. 2022. Lamberta: Law article mining based on bert architecture for the italian civil code. In *ICRDL*.

Shavrina Tatiana and Malykh Valentin. 2021. How not to Lie with a Benchmark: Rearranging NLP Leaderboards. ArXiv:2112.01342 [cs].

Rena Torres Cacoullos. 2020. Code-switching strategies: Prosody and syntax. *Frontiers in Psychology*, 11.

Hsiang-Sheng Tsai, Heng-Jui Chang, Wen-Chin Huang, Zili Huang, Kushal Lakhotia, Shu-Wen Yang, Shuyan Dong, Andy T. Liu, Cheng-I Lai, Jiatong Shi, Xuankai Chang, Phil Hall, Hsuan-Jui Chen, Shang-Wen Li, Shinji Watanabe, Abdelrahman Mohamed, and Hung-yi Lee. 2022. SUPERB-SG: enhanced speech processing universal performance benchmark for semantic and generative capabilities. In *ACL (1)*, pages 8479–8492. Association for Computational Linguistics.

Dimitrios Tsarapatsanis and Nikolaos Aletras. 2021. On the ethical limits of natural language processing on legal text. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3590–3599, Online. Association for Computational Linguistics.

Georgios Tziafas, Eugenie de Saint-Phalle, Wietse de Vries, Clara Egger, and Tommaso Caselli. 2021. A multilingual approach to identify and classify exceptional measures against covid-19. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 46–62. Dataset URL: https://tinyurl.com/ycysvtbm.

Stefanie Urchs., Jelena Mitrović., and Michael Granitzer. 2021. Design and implementation of german legal decision corpora. pages 515–521. SciTePress.

Serena Villata, Michał Araszkiewicz, Kevin D. Ashley, Trevor J. M. Bench-Capon, L. Karl Branting, Jack G. Conrad, and Adam Zachary Wyner. 2022. Thirty years of artificial intelligence and law: the third decade. *Artificial Intelligence and Law*, 30:561–591.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019a. Superglue: A stickier benchmark for general-purpose language understanding systems. In *NeurIPS*, pages 3261–3275.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019b. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *ICLR (Poster)*. OpenReview.net.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers.

Chaojun Xiao, Xueyu Hu, Zhiyuan Liu, Cunchao Tu, and Maosong Sun. 2021. Lawformer: A pre-trained language model for chinese legal long documents. *AI Open*, 2:79–84.

Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Chenjie Cao, Yudong Li, Yechen Xu, Kai Sun, Dian Yu, Cong Yu, Yin Tian, Qianqian Dong, Weitang Liu, Bo Shi, Yiming Cui, Junyi Li, Jun Zeng, Rongzhao Wang, Weijian Xie, Yanting Li, Yina Patterson, Zuoyu Tian, Yiwen Zhang, He Zhou, Shaoweihua Liu, Zhe Zhao, Qipeng Zhao, Cong Yue, Xinrui Zhang, Zhengliang Yang, Kyle Richardson, and Zhenzhong Lan. 2020. CLUE: A Chinese language understanding evaluation benchmark. In *COLING*, pages 4762–4772, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. *arXiv:2010.11934 [cs]*. ArXiv: 2010.11934.

Shu-Wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y. Lin, Andy T. Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, Tzu-Hsien Huang, Wei-Cheng Tseng, Kotik Lee, Da-Rong Liu, Zili Huang, Shuyan Dong, Shang-Wen Li, Shinji Watanabe, Abdelrahman Mohamed, and Hung-yi Lee. 2021. SUPERB: speech processing universal performance benchmark. In *Interspeech*, pages 1194–1198. ISCA.

Yin Ying and Ivan Habernal. 2022. Privacy-Preserving Models for Legal Natural Language Processing. In *NLLP*, page (to appear), Abu Dhabi, UAE.

Ningyu Zhang, Mosha Chen, Zhen Bi, Xiaozhuan Liang, Lei Li, Xin Shang, Kangping Yin, Chuanqi Tan, Jian Xu, Fei Huang, Luo Si, Yuan Ni, Guotong Xie, Zhifang Sui, Baobao Chang, Hui Zong, Zheng Yuan, Linfeng Li, Jun Yan, Hongying Zan, Kunli Zhang, Buzhou Tang, and Qingcai Chen. 2022. CBLUE: A Chinese biomedical language understanding evaluation benchmark. In *ACL*, pages 7888–7915, Dublin, Ireland. Association for Computational Linguistics.

Lucia Zheng, Neel Guha, Brandon R. Anderson, Peter Henderson, and Daniel E. Ho. 2021. When does pretraining help? assessing self-supervised learning for law and the casehold dataset of 53,000+ legal holdings. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, ICAIL '21, page 159–168, New York, NY, USA. Association for Computing Machinery.

Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020. How does NLP benefit legal system: A summary of legal artificial intelligence. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5218–5230, Online. Association for Computational Linguistics.

A Experiment Details

A.1 Maximum Sequence Lengths

Brazilian Court Decisions: 1024 (128 x 8) CoVID19: 256 German Argument Mining: 256 Greek Legal Code: 4096 (if speed is important: 2048) (128 x 32 / 16) Greek Legal NER: 512 (max for non-hierarchical) LegalNERo: 512 (max for non-hierarchical) LeNER: 512 (max for non-hierarchical) MAPA: 512 (max for non-hierarchical) MultiEURLEX: 4096 (or for maximum performance 8192) (128 x 32 / 64) Online Terms of Service: 256 Swiss Judgment Prediction: 2048 (or for maximum performance on fr: 4096) (128 x 16 / 32)

A.2 Hyperparameters

We used learning rate 1e-5 for all models and datasets without tuning. We ran all experiments with 3 random seeds (1-3) We always used batch size 64. In case the GPU memory was insufficient, we additionally used gradient accumulation. We trained using early stopping on the validation loss with patience of 5 epochs. Because MultiEURLEX is very large and the experiment very long, we just train for 1 epoch and evaluated after every 1000th step. We used AMP mixed precision training and evaluation to reduce costs. Mixed precision was not used in combination with microsoft/mdeberta-v3-base because it led to errors. The experiments were run the following NVIDIA GPUs: 24GB RTX3090, 32GB V100 and 80GB A100.

B Dataset Splits

Language	SJP	OTS-UL	OTS-CT	C19	MEU-1	MEU-2	MEU-3	MAP-C	MAP-F
bg					15986 / 5000 / 5000	15986 / 5000 / 5000	15986 / 5000 / 5000	1411 / 166 / 560	1411 / 166 / 560
cs					23187 / 5000 / 5000	23187 / 5000 / 5000	23187 / 5000 / 5000	1464 / 176 / 563	1464 / 176 / 563
da					55000 / 5000 / 5000	55000 / 5000 / 5000	55000 / 5000 / 5000	1455 / 164 / 550	1455 / 164 / 550
de	35458 / 4705 / 9725	491 / 42 / 103	4480 / 404 / 1027		55000 / 5000 / 5000	55000 / 5000 / 5000	55000 / 5000 / 5000	1457 / 166 / 558	1457 / 166 / 558
el					55000 / 5000 / 5000	55000 / 5000 / 5000	55000 / 5000 / 5000	1529 / 174 / 584	1529 / 174 / 584
en		526 / 49 / 103	5378 / 415 / 1038	648 / 81 / 81	55000 / 5000 / 5000	55000 / 5000 / 5000	55000 / 5000 / 5000	893 / 98 / 408	893 / 98 / 408
es					52785 / 5000 / 5000	52785 / 5000 / 5000	52785 / 5000 / 5000	806 / 248 / 155	806 / 248 / 155
et					23126 / 5000 / 5000	23126 / 5000 / 5000	23126 / 5000 / 5000	1391 / 163 / 516	1391 / 163 / 516
fi					42497 / 5000 / 5000	42497 / 5000 / 5000	42497 / 5000 / 5000	1398 / 187 / 531	1398 / 187 / 531
fr	21179 / 3095 / 6820			1416 / 178 / 178	55000 / 5000 / 5000	55000 / 5000 / 5000	55000 / 5000 / 5000	1297 / 97 / 490	1297 / 97 / 490
ga								1383 / 165 / 515	1383 / 165 / 515
hr					7944 / 2500 / 5000	7944 / 2500 / 5000	7944 / 2500 / 5000		
hu				75 / 10 / 10	22664 / 5000 / 5000	22664 / 5000 / 5000	22664 / 5000 / 5000	1390 / 171 / 525	1390 / 171 / 525
it	3072 / 408 / 812	517 / 50 / 102	4806 / 432 / 1057	742 / 93 / 93	55000 / 5000 / 5000	55000 / 5000 / 5000	55000 / 5000 / 5000	1411 / 162 / 550	1411 / 162 / 550
lt					23188 / 5000 / 5000	23188 / 5000 / 5000	23188 / 5000 / 5000	1413 / 173 / 548	1413 / 173 / 548
lv					23208 / 5000 / 5000	23208 / 5000 / 5000	23208 / 5000 / 5000	1383 / 167 / 553	1383 / 167 / 553
mt					17521 / 5000 / 5000	17521 / 5000 / 5000	17521 / 5000 / 5000	937 / 93 / 442	937 / 93 / 442
nb				221 / 28 / 28					
nl				135 / 18 / 18	55000 / 5000 / 5000	55000 / 5000 / 5000	55000 / 5000 / 5000	1391 / 164 / 530	1391 / 164 / 530
pl		540 / 50 / 109	5278 / 439 / 1175	75 / 10 / 10	23197 / 5000 / 5000	23197 / 5000 / 5000	23197 / 5000 / 5000		
pt					52370 / 5000 / 5000	52370 / 5000 / 5000	52370 / 5000 / 5000	1086 / 105 / 390	1086 / 105 / 390
ro					15921 / 5000 / 5000	15921 / 5000 / 5000	15921 / 5000 / 5000	1480 / 175 / 557	1480 / 175 / 557
sk					22971 / 5000 / 5000	22971 / 5000 / 5000	22971 / 5000 / 5000	1395 / 165 / 526	1395 / 165 / 526
sl					23184 / 5000 / 5000	23184 / 5000 / 5000	23184 / 5000 / 5000		
sv					42490 / 5000 / 5000	42490 / 5000 / 5000	42490 / 5000 / 5000	1453 / 175 / 539	1453 / 175 / 539

Table 7: Overview of the number of examples for each language-specific subset of multilingual tasks. The order of the values is train / validation / test.

Language	SJP	OTS-UL	OTS-CT	C19	MEU-1	MEU-2	MEU-3	MAP-C	MAP-F
bg					21 / 21 / 21	127 / 126 / 127	481 / 454 / 465	11 / 11 / 8	24 / 16 / 13
cs					21 / 21 / 21	127 / 126 / 127	486 / 454 / 465	11 / 11 / 9	30 / 17 / 16
da					21 / 21 / 21	127 / 126 / 127	500 / 454 / 465	11 / 10 / 11	26 / 14 / 14
de	2 / 2 / 2	3 / 3 / 3	9 / 7 / 9		21 / 21 / 21	127 / 126 / 127	500 / 454 / 465	11 / 9 / 10	28 / 14 / 14
el					21 / 21 / 21	127 / 126 / 127	500 / 454 / 465	11 / 11 / 11	31 / 17 / 20
en		3 / 3 / 3	9 / 8 / 9	6 / 6 / 5	21 / 21 / 21	127 / 126 / 127	500 / 454 / 465	11 / 9 / 9	28 / 17 / 18
es					21 / 21 / 21	127 / 126 / 127	497 / 454 / 465	11 / 8 / 11	26 / 13 / 18
et					21 / 21 / 21	127 / 126 / 127	486 / 454 / 465	11 / 11 / 11	25 / 14 / 17
fi					21 / 21 / 21	127 / 126 / 127	493 / 454 / 465	11 / 11 / 10	24 / 19 / 16
fr	2 / 2 / 2			8 / 8 / 7	21 / 21 / 21	127 / 126 / 127	500 / 454 / 465	11 / 11 / 11	32 / 19 / 26
ga								13 / 11 / 11	33 / 17 / 18
hr					21 / 21 / 21	127 / 126 / 127	469 / 437 / 465		
hu				4 / 1 / 1	21 / 21 / 21	127 / 126 / 127	486 / 454 / 465	11 / 10 / 10	20 / 15 / 14
it	2 / 2 / 2	3 / 3 / 3	9 / 8 / 9	7 / 7 / 6	21 / 21 / 21	127 / 126 / 127	500 / 454 / 465	11 / 10 / 11	25 / 15 / 16
lt					21 / 21 / 21	127 / 126 / 127	486 / 454 / 465	11 / 11 / 10	28 / 19 / 21
lv					21 / 21 / 21	127 / 126 / 127	486 / 454 / 465	11 / 11 / 11	31 / 15 / 21
mt					21 / 21 / 21	127 / 126 / 127	485 / 454 / 465	11 / 11 / 11	27 / 15 / 15
nb				7 / 5 / 6					
nl				2 / 2 / 2	21 / 21 / 21	127 / 126 / 127	500 / 454 / 465	10 / 9 / 10	25 / 12 / 14
pl		3 / 3 / 3	9 / 8 / 9	7 / 5 / 3	21 / 21 / 21	127 / 126 / 127	486 / 454 / 465		
pt					21 / 21 / 21	127 / 126 / 127	497 / 454 / 465	11 / 10 / 11	29 / 14 / 18
ro					21 / 21 / 21	127 / 126 / 127	481 / 454 / 465	11 / 11 / 11	25 / 16 / 18
sk					21 / 21 / 21	127 / 126 / 127	485 / 454 / 465	11 / 11 / 11	25 / 16 / 18
sl					21 / 21 / 21	127 / 126 / 127	486 / 454 / 465		
sv					21 / 21 / 21	127 / 126 / 127	493 / 454 / 465	11 / 11 / 10	23 / 15 / 15

Table 8: Overview of the number of labels for each language-specific subset of multilingual tasks. The order of the values is train / validation / test.

C Detailed Multilingual Results

	Mean	BCD-J	BCD-U	GAM	GLC-V	GLC-C	GLC-S	SJP	OTS-UL	OTS-CT	C19	MEU-1	GLN	LNR	LNB	MAP-C	MAP-F
Model																	
MiniLM	55.3	52.8 (±6.7)	55.1 (±6.6)	72.1 (±0.9)	82.0 (±1.0)	39.4 (±1.0)	**5.1 (±1.6)**	68.9 (±0.7)	71.0 (±5.0)	15.3 (±3.4)	5.9 (±1.5)	64.8 (±0.3)	41.5 (±3.1)	**63.5 (±5.3)**	86.0 (±0.4)	80.1 (±0.2)	62.8 (±2.3)
DistilBERT	61.7	52.1 (±4.5)	60.0 (±9.8)	70.6 (±1.7)	84.9 (±0.5)	68.0 (±0.6)	33.9 (±2.0)	68.7 (±0.7)	66.9 (±3.4)	49.6 (±9.1)	41.4 (±5.6)	68.2 (±0.1)	38.9 (±2.9)	**63.5 (±3.7)**	70.3 (±1.6)	78.7 (±0.2)	58.8 (±1.8)
mDeBERTa v3	64.7	**68.2 (±3.9)**	69.9 (±5.6)	69.5 (±2.0)	85.0 (±0.8)	58.2 (±7.5)	12.3 (±2.5)	**71.2 (±0.7)**	**85.2 (±2.9)**	52.1 (±4.6)	43.4 (±4.3)	68.4 (±0.6)	44.6 (±1.8)	62.3 (±3.1)	88.5 (±2.2)	**81.1 (±0.9)**	**67.6 (±0.9)**
XLM-R base	64.2	67.5 (±2.2)	63.4 (±12.3)	72.5 (±1.9)	**85.4 (±0.2)**	68.1 (±1.6)	15.7 (±12.7)	69.6 (±0.9)	72.6 (±4.2)	52.4 (±6.0)	44.1 (±7.9)	69.2 (±0.1)	45.9 (±1.8)	63.1 (±2.8)	85.3 (±1.5)	80.1 (±1.0)	63.0 (±0.7)
XLM-R large	**66.4**	58.1 (±9.3)	**70.4 (±3.7)**	**73.0 (±1.4)**	58.2 (±50.2)	**73.0 (±0.9)**	38.9 (±33.7)	70.0 (±1.8)	84.9 (±2.7)	**62.9 (±6.1)**	**53.8 (±10.5)**	**71.2 (±1.4)**	**47.5 (±3.7)**	54.9 (±3.7)	**88.7 (±1.1)**	**81.1 (±0.9)**	65.9 (±1.7)

Table 9: Macro-F1 and standard deviation for multilingual models from the validation set. The best scores are in bold.

	Mean	BCD-J	BCD-U	GAM	GLC-V	GLC-C	GLC-S	SJP	OTS-UL	OTS-CT	C19	MEU-1	GLN	LNR	LNB	MAP-C	MAP-F
Model																	
MiniLM	51.7	49.4 (±7.4)	56.7 (±7.9)	**73.3 (±0.9)**	81.7 (±0.5)	39.4 (±1.4)	**5.2 (±1.6)**	67.6 (±1.2)	74.6 (±1.1)	14.1 (±3.1)	5.6 (±2.2)	62.0 (±0.4)	40.5 (±4.0)	46.8 (±1.9)	86.0 (±0.2)	63.0 (±2.1)	40.4 (±2.1)
DistilBERT	58.0	50.3 (±2.9)	58.8 (±8.7)	69.5 (±0.9)	85.2 (±0.8)	70.0 (±0.3)	33.2 (±1.9)	66.7 (±1.1)	67.2 (±4.1)	46.2 (±8.9)	39.5 (±6.3)	63.6 (±0.1)	38.1 (±2.0)	48.4 (±5.2)	78.7 (±1.1)	61.3 (±2.8)	40.6 (±0.7)
mDeBERTa v3	59.4	**65.8 (±3.5)**	49.3 (±0.1)	70.9 (±0.9)	85.6 (±1.0)	58.6 (±7.8)	12.4 (±2.8)	**69.0 (±0.8)**	79.7 (±3.8)	53.8 (±3.0)	40.7 (±5.0)	65.0 (±0.4)	42.2 (±1.6)	46.6 (±1.1)	87.8 (±0.7)	**65.3 (±3.1)**	**46.3 (±0.6)**
XLM-R base	61.2	65.4 (±3.6)	61.6 (±11.2)	72.0 (±2.4)	**85.9 (±0.1)**	69.3 (±1.6)	15.4 (±12.3)	68.3 (±1.0)	80.8 (±1.9)	55.9 (±2.6)	45.9 (±11.0)	65.6 (±0.1)	47.0 (±2.2)	47.7 (±2.9)	86.0 (±1.9)	61.4 (±2.8)	42.2 (±0.4)
XLM-R large	**63.2**	55.1 (±7.6)	**62.3 (±3.6)**	73.1 (±1.5)	58.3 (±50.3)	**74.7 (±0.9)**	39.1 (±33.9)	68.3 (±1.8)	**83.6 (±4.8)**	**66.9 (±0.5)**	**54.2 (±7.2)**	**68.1 (±1.2)**	**48.0 (±4.2)**	**49.5 (±11.3)**	**88.2 (±0.7)**	65.0 (±5.7)	46.2 (±2.1)

Table 10: Macro-F1 and standard deviation for multilingual models from the test set. The best scores are in bold.

D Histograms

In the following, we provide the histograms for the distribution of the sequence length of the input (sentence or entire document) from each dataset. The length is measured by counting the tokens using the tokenizers of the multilingual models, i.e., DistilBERT, MiniLM, mDeBERTa v3, XLM-R base, XLM-R large. We only display the distribution within the 99th percentile; the rest is grouped together at the end.

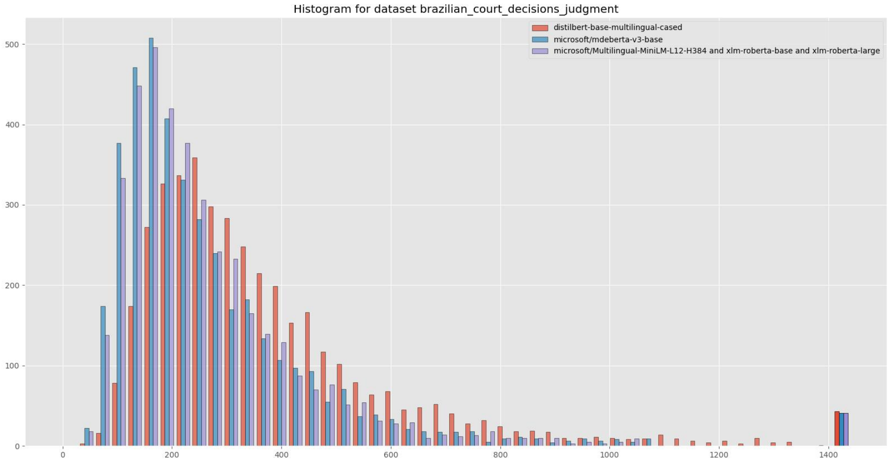

Figure 2: Histogram for dataset BCD-J

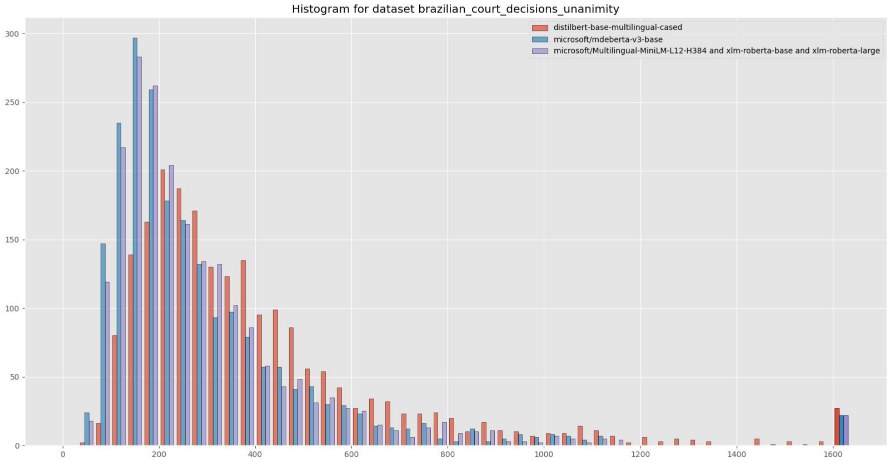

Figure 3: Histogram for dataset BCD-U

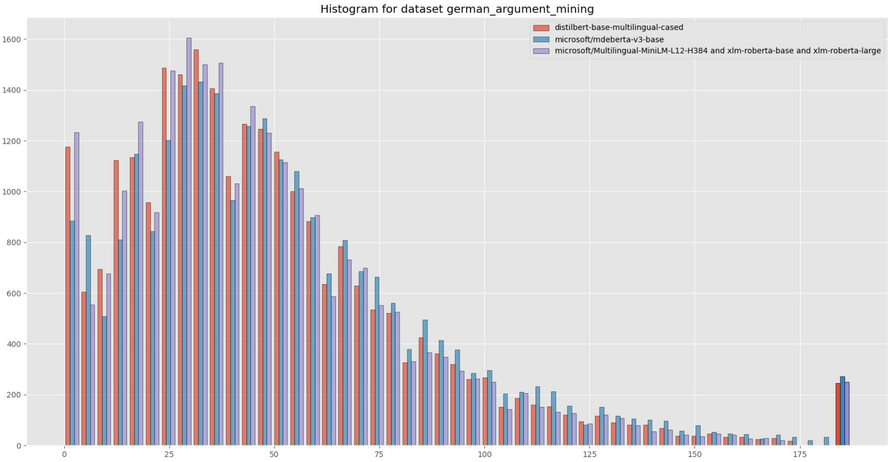

Figure 4: Histogram for dataset GAM

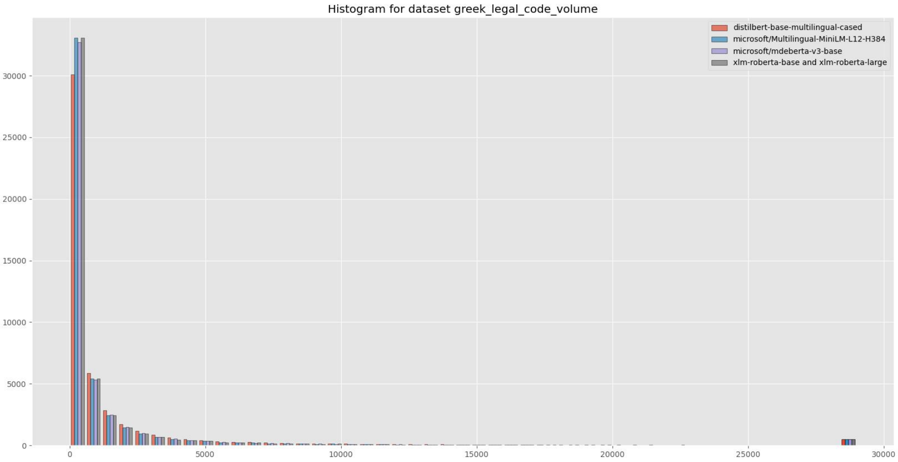

Figure 5: Histogram for dataset GLC-V

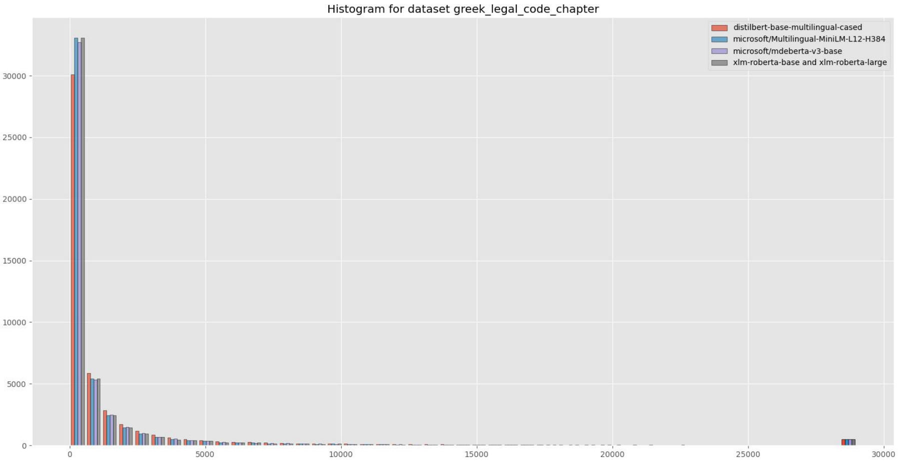

Figure 6: Histogram for dataset GLC-C

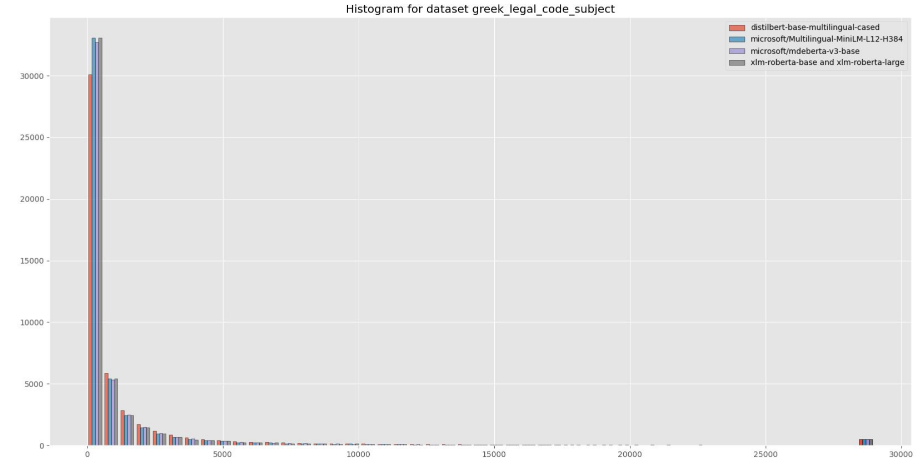

Figure 7: Histogram for dataset GLC-S

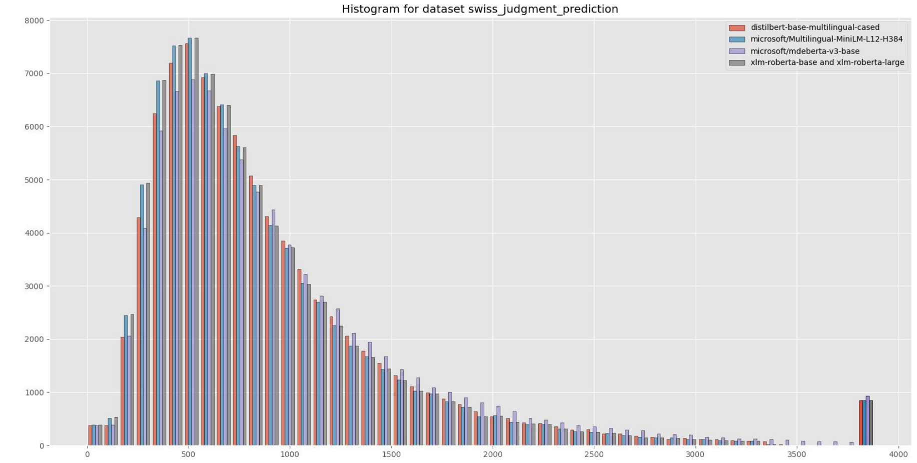

Figure 8: Histogram for dataset SJP

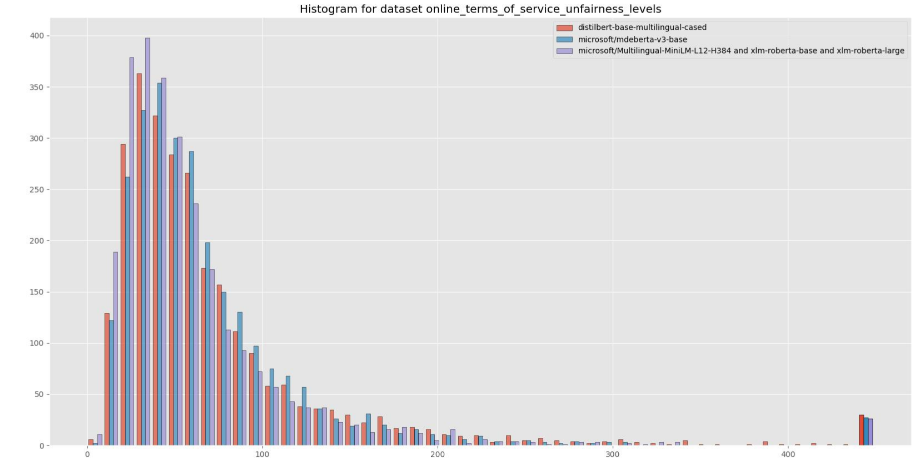

Figure 9: Histogram for dataset OTS-UL

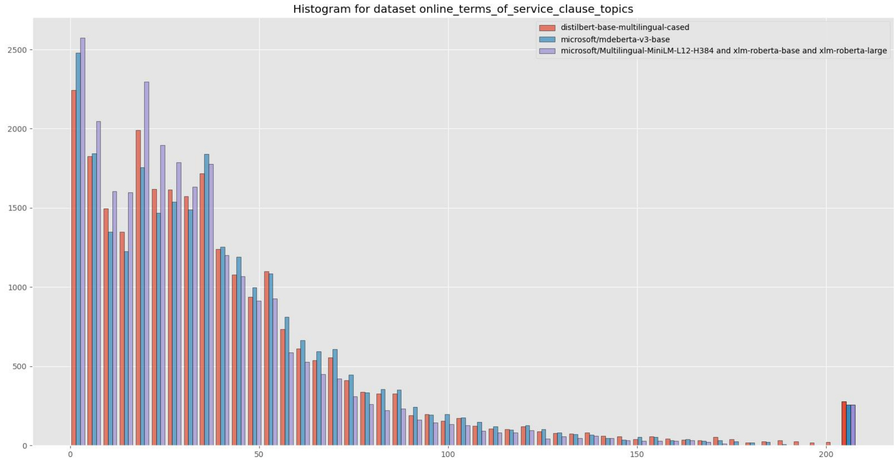

Figure 10: Histogram for dataset OTS-CT

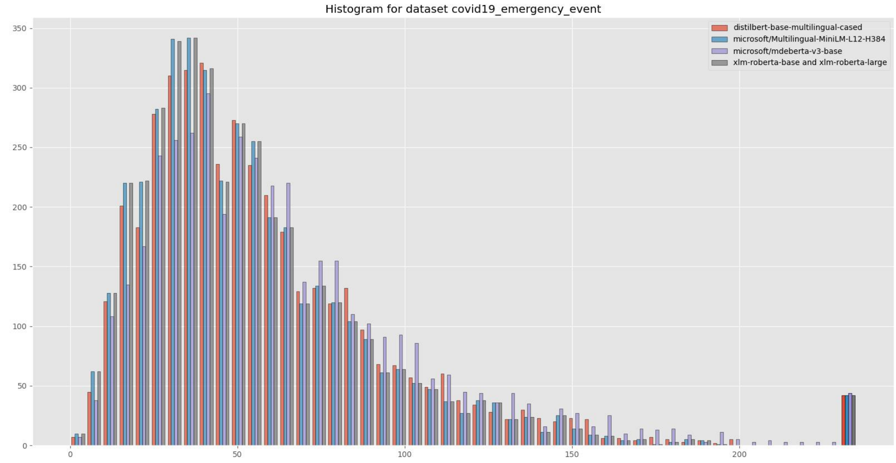

Figure 11: Histogram for dataset C19

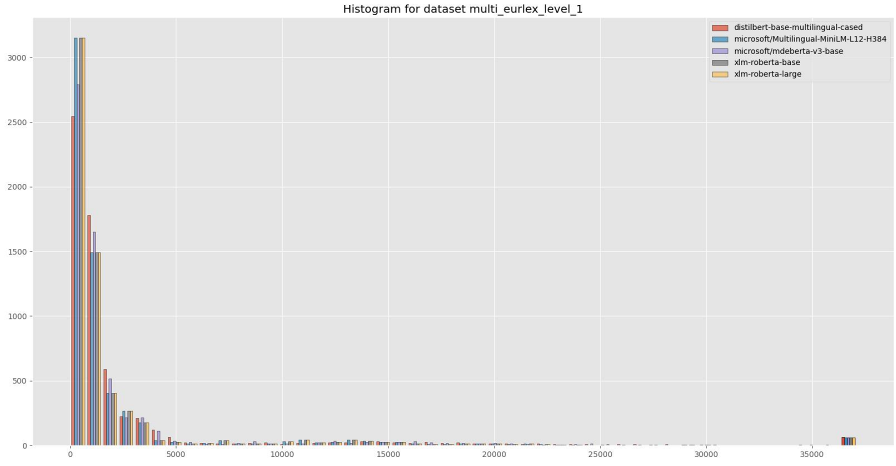

Figure 12: Histogram for dataset MEU-1

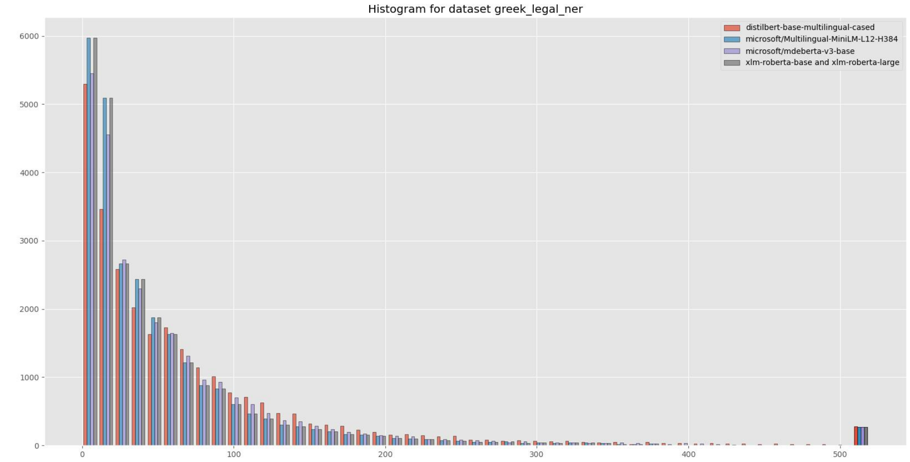

Figure 13: Histogram for dataset GLN

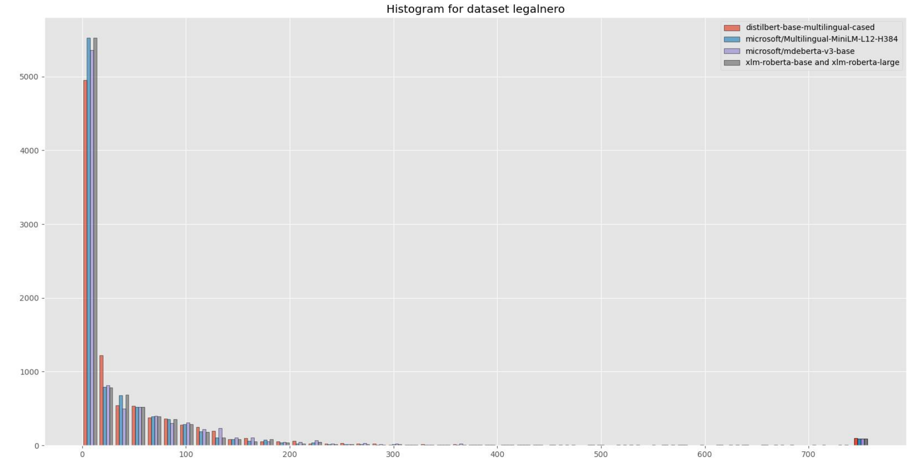

Figure 14: Histogram for dataset LNR

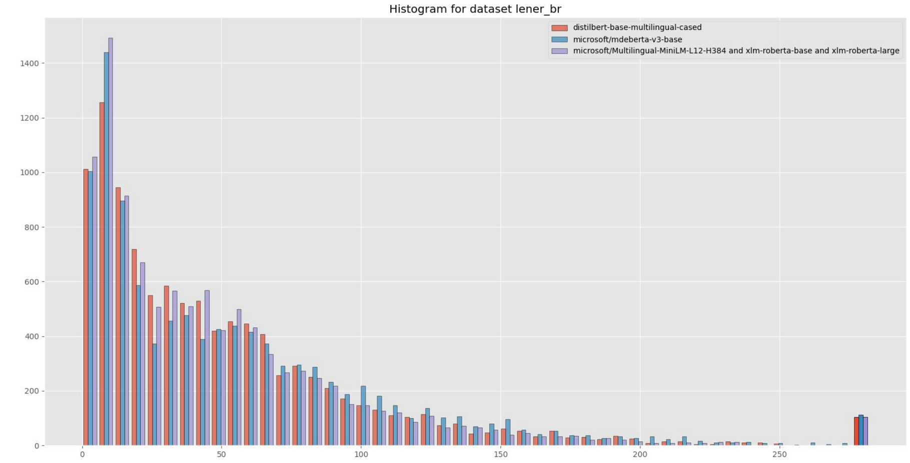

Figure 15: Histogram for dataset LNB

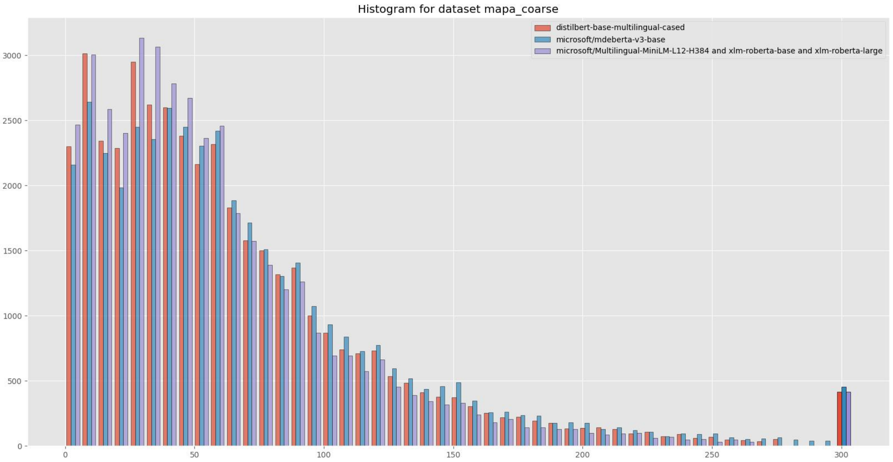

Figure 16: Histogram for dataset MAP-C

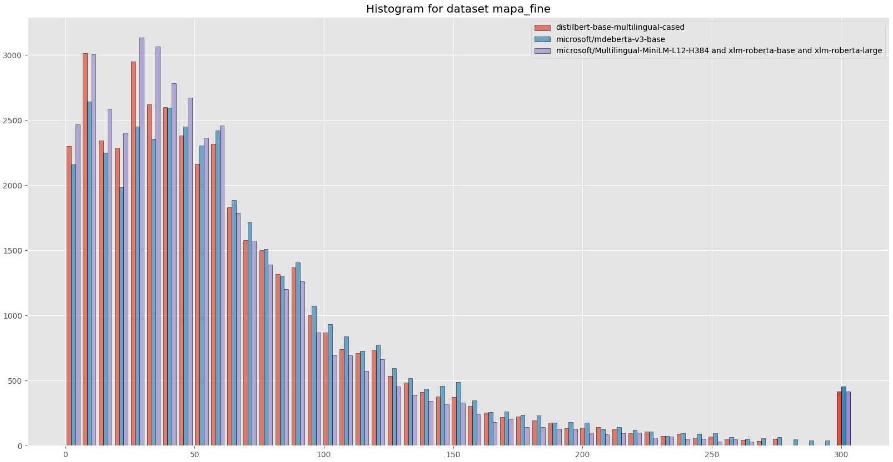

Figure 17: Histogram for dataset MAP-F