MultiLegalPile: A 689GB Multilingual Legal Corpus

Joel Niklaus, Veton Matoshi, Matthias Stürmer, Ilias Chalkidis, Daniel E. Ho

Universität Bern

Berner Fachhochschule
Haute école spécialisée bernoise
Bern University of Applied Sciences

Stanford University

Dataset & Models Available

Why create MultiLegalPile?

LLMs need huge amounts of pretraining data

Most corpora are **English-Only** and based on **Web Crawls**

High-Quality **Multilingual Domain-Specific** Corpora are rare

⇒ MultiLegalPile is a **Multilingual Legal** Corpus

4 sources, 5 text types, 24 languages

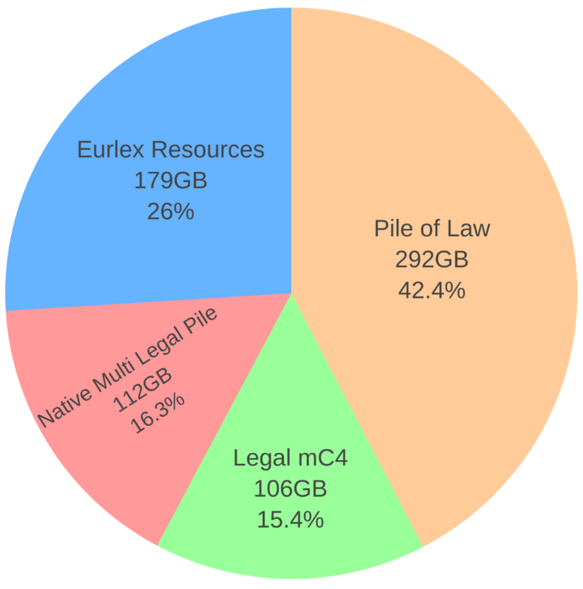

Figure 1. MULTILEGALPILE Source Distribution

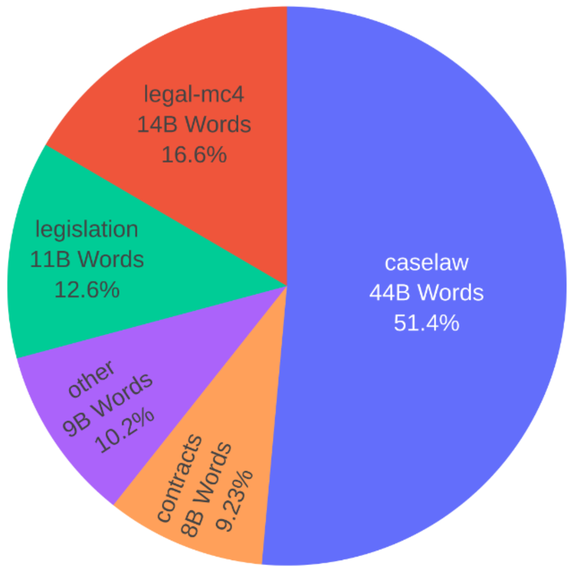

Figure 2. MULTILEGALPILE Text Type Distribution

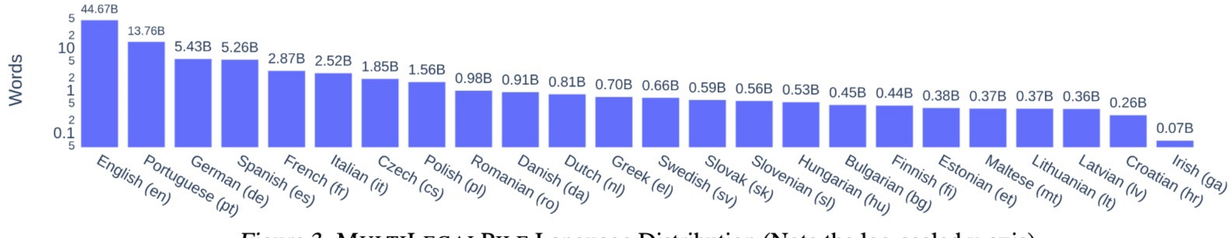

Figure 3. MULTILEGALPILE Language Distribution (Note the log-scaled y-axis)

Experimental 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	5.0	29.7	74.0	84.5	93.6	57.8	56.8
DistilBERT	54.5	69.5	62.8	66.8	56.1	25.9	36.4	71.0	85.3	89.6	60.8	61.7
mDeBERTa-v3	60.2	71.3	52.2	69.1	66.5	29.7	37.4	73.3	85.1	94.8	67.2	64.3
XLM-R-base	63.5	72.0	57.4	69.3	67.8	26.4	33.3	**74.6**	**85.8**	94.1	62.0	64.2
XLM-R-large	58.7	73.1	57.4	69.0	**75.0**	29.0	**42.2**	74.1	85.0	**95.3**	68.0	66.1
Legal-XLM-R-base	62.5	72.4	68.9	70.2	70.8	30.7	38.6	73.6	84.1	94.1	**69.2**	66.8
Legal-XLM-R-large	63.3	73.9	59.3	70.1	74.9	**34.6**	39.7	73.1	83.9	94.6	67.3	66.8
Legal-XLM-LF-base	**72.4**	**74.6**	**70.2**	**72.9**	69.8	26.3	33.1	72.1	84.7	93.3	66.2	**66.9**

Table 4. Dataset aggregate scores for multilingual models on LEXTREME. We report macro-F1 and the best scores in bold.

➔ New SotA on LEXTREME, a multilingual legal benchmark

Model	ECtHR-A	ECtHR-B	SCOTUS	EUR-LEX	LEDGAR	UNFAIR-ToS	CaseHOLD	Agg.
TFIDF+SVM *	48.9	63.8	64.4	47.9	81.4	75.0	22.4	49.0
BERT *	63.6	73.4	58.3	57.2	81.8	81.3	70.8	68.2
DeBERTa *	60.8	71.0	62.7	57.4	83.1	80.3	72.6	68.5
RoBERTa-base *	59.0	68.9	62.0	57.9	82.3	79.2	71.4	67.5
RoBERTa-large *	67.6	71.6	66.3	58.1	**83.6**	81.6	74.4	70.9
Longformer *	64.7	71.7	64.0	57.7	83.0	80.9	71.9	69.5
BigBird *	62.9	70.9	62.0	56.8	82.6	81.3	70.8	68.4
Legal-BERT *	64.0	74.7	66.5	57.4	83.0	**83.0**	75.3	70.8
CaseLaw-BERT *	62.9	70.3	65.9	56.6	83.0	82.3	75.4	69.7
Legal-en-R-base (ours)	65.2	73.7	66.4	**59.2**	82.7	78.7	73.3	70.5
Legal-en-R-large (ours)	**70.3**	**77.0**	**67.7**	58.4	82.5	82.4	**77.0**	**72.7**
Legal-XLM-R-base (ours)	64.8	73.9	63.9	58.2	82.8	79.6	71.7	69.7
Legal-XLM-R-large (ours)	68.2	74.2	67.5	58.4	82.7	79.9	75.1	71.4
Legal-XLM-LF-base (ours)	67.9	76.2	61.6	59.1	82.1	78.9	72.0	70.2

Table 6. Results on LexGLUE. We report macro-F1 and best scores in bold. Results from models marked with * are from Chalkidis et al. (2021d). Similar to LEXTREME, we computed the aggregate score as the harmonic mean of individual dataset results.

➔ New SotA on LexGLUE, an English legal benchmark

Future Work

Extend the corpus to more languages and jurisdictions

Train a large generative model

Resources

Dataset on 🤗: https://huggingface.co/datasets/joelniklaus/Multi_Legal_Pile

Models on 🤗: https://huggingface.co/joelniklaus/legal-xlm-roberta-base