Processing Long Legal Documents with Pre-Trained Transformers: Modding LegalBERT and Longformer

Dimitris Mamakas, Petros Tsotsi, Ion Androutsopoulos, Ilias Chalkidis

ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

KØBENHAVNS UNIVERSITET

Motivation

- **Pre-trained Transformers** are currently the **state-of-the-art** in NLP.
- The **quadratic complexity** of their attention mechanism **restricts** the **maximum input length** of text they can process.
- **Legal domain** datasets often contain texts **far longer** than those limits.
- Even **sparse attention** models (e.g., Longformer) especially designed for long texts, still **cannot cope** with **long legal documents**.
- **BoW** models can process **texts of any length**, but **ignore word order**.

LexGLUE Benchmark

Dataset	Source	Text length (words)		Instances (training/dev/test)	Classes
		Average	Maximum		
ECtHR Task A	Chalkidis et al. (2019)	1.6k	35.4k	9,000 / 1,000 / 1,000	10+1◇
ECtHR Task B	Chalkidis et al. (2021a)	1.6k	35.4k	9,000 / 1,000 / 1,000	10+1◇
SCOTUS	Spaeth et al. (2020)	6.0k	88.6k	5,000 / 1,400 / 1,400	14
EUR-LEX	Chalkidis et al. (2021b)	1.1k	140.1k	55,000 / 5,000 / 5,000	100
LEDGAR	Tuggener et al. (2020)	113	1.2k	60,000 / 10,000 / 10,000	100
UNFAIR-ToS	Lippi et al. (2019)	33	441	5,532 / 2,275 / 1,1607	8+1◇

◇ +1 means that some documents aren't relevant to any class.

LexGLUE example

◇ Example retrieved from ECtHR dataset

1. At the **beginning of the events** relevant to the application, K. had a daughter, P., and a son, M., born in 1986 and 1988 respectively. P.'s father is X and M.'s father is...and **M.'s foster mother died in May 2001**.

[...]

53. On 29 April 1962 **the applicant married Mr A. Gigliozz**i in a religious ceremony which was also valid in the eyes of the law (matrimonio concordatario).", "12. On 23 February 1987...she also withdrew another set of proceedings that she had instituted in the Viterbo Court claiming joint title to property).

Large texts containing more than **500** words on average

Multi-Label classification task

European Court of Human Rights

- A2: Right to life
- A3: Prohibition of torture
- A5: Right to liberty and security
- A6: Right to a fair trial
- A8: Right to respect for private and family life
- A9: Freedom of thought, conscience and religion
- A10: Freedom of expression
- A11: Freedom of assembly and association
- P1-1: Protection of property
- A0: No violation

Prior work

Sparse attention variants

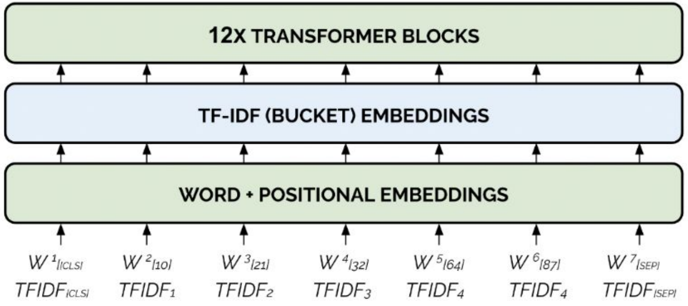
- These models combine a local windowed attention with a global attention and achieve **linear complexity**.
- Longformer (Beltagy et al. 2020), BigBird (Zaheer et al., 2020), ETC (Ainslie et al., 2020).

Hierarchical Transformers

- Use models like BERT to separately encode each paragraph of the input.
- Then additional layers to make the paragraph embeddings aware of surrounding paragraphs.
- Hierarchical LegalBERT *(Chalkidis et al. 2020)*, SMITH (Yang et al., 2020).

Our contribution (1): BOW BERT variants

(a) **TFIDF-SRT-EMB-Legal-BERT**



Deduplicate + Sort by TFIDF

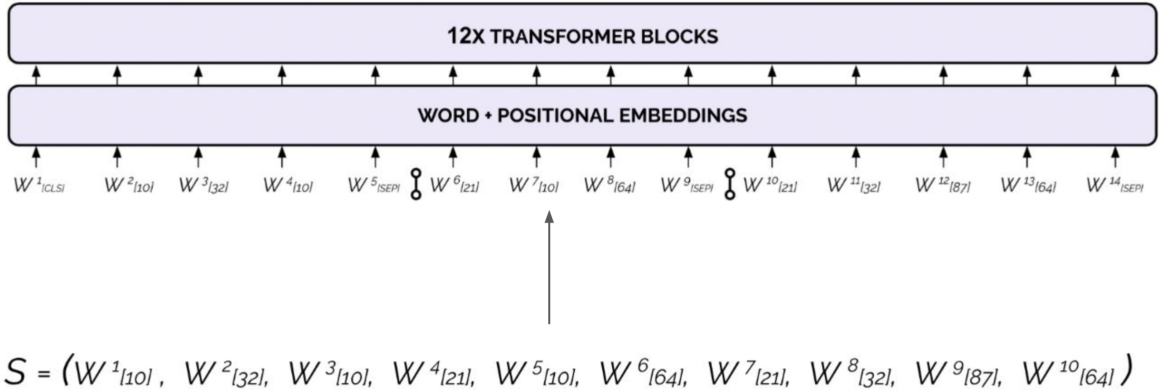
$$S = (W^1_{[10]},\ W^2_{[32]},\ W^3_{[10]},\ W^4_{[21]},\ W^5_{[10]},\ W^6_{[64]},\ W^7_{[21]},\ W^8_{[32]},\ W^9_{[87]},\ W^{10}_{[64]})$$

Our contribution (2): Longformer extensions

(b) **Longformer-8192-PAR**

Can process up to **8,192** tokens whereas the original version can handle only up to **4,096**

Split in chunks



Experimental results (BoW models)

- Results on test data.

Model	ECtHR (Task A)		ECtHR (Task B)		SCOTUS		EUR-LEX		LEDGAR		UNFAIR-ToS	
	μ-F1	m-F1	μ-F1	m-F1	μ-F1	m-F1	μ-F1	m-F1	μ-F1	m-F1	μ-F1	m-F1
TFIDF+SVM	62.6	48.9	73.0	63.8	74.0	64.4	63.4	47.9	87.0	81.4	94.7	75.0
TFIDF-SRT-LegalBERT	69.8	62.8	78.5	71.9	73.4	61.8	69.6	53.7	86.9	80.8	95.3	80.6
TFIDF-SRT-EMB-LegalBERT	68.7	63.1	79.0	72.5	73.9	63.6	69.7	53.9	86.5	80.3	95.8	78.7

Experimental results (BoW models)

- Results on test data.

Model	ECtHR (Task A) *		ECtHR (Task B) *		SCOTUS *		EUR-LEX		LEDGAR		UNFAIR-ToS	
	μ-F1	m-F1	μ-F1	m-F1	μ-F1	m-F1	μ-F1	m-F1	μ-F1	m-F1	μ-F1	m-F1
TFIDF+SVM	62.6	48.9	73.0	63.8	74.0	64.4	63.4	47.9	87.0	81.4	94.7	75.0
TFIDF-SRT-LegalBERT	69.8	62.8	78.5	71.9	73.4	61.8	69.6	53.7	86.9	80.8	95.3	80.6
TFIDF-SRT-EMB-LegalBERT	68.7	63.1	79.0	72.5	73.9	63.6	69.7	53.9	86.5	80.3	95.8	78.7
LegalBERT variants that retain word order												
LegalBERT	70.0	64.0	80.4	74.7	76.4	66.5	72.1	57.4	88.2	83.0	**96.0**	**83.0**
TFIDF-EMB-LegalBERT	70.0	61.9	79.4	73.5	74.9	64.7	71.6	56.9	88.7	83.4	95.9	82.1

* The results were obtained using the hierarchical version of the corresponding model.

Best results

Experimental results (Longformer variants)

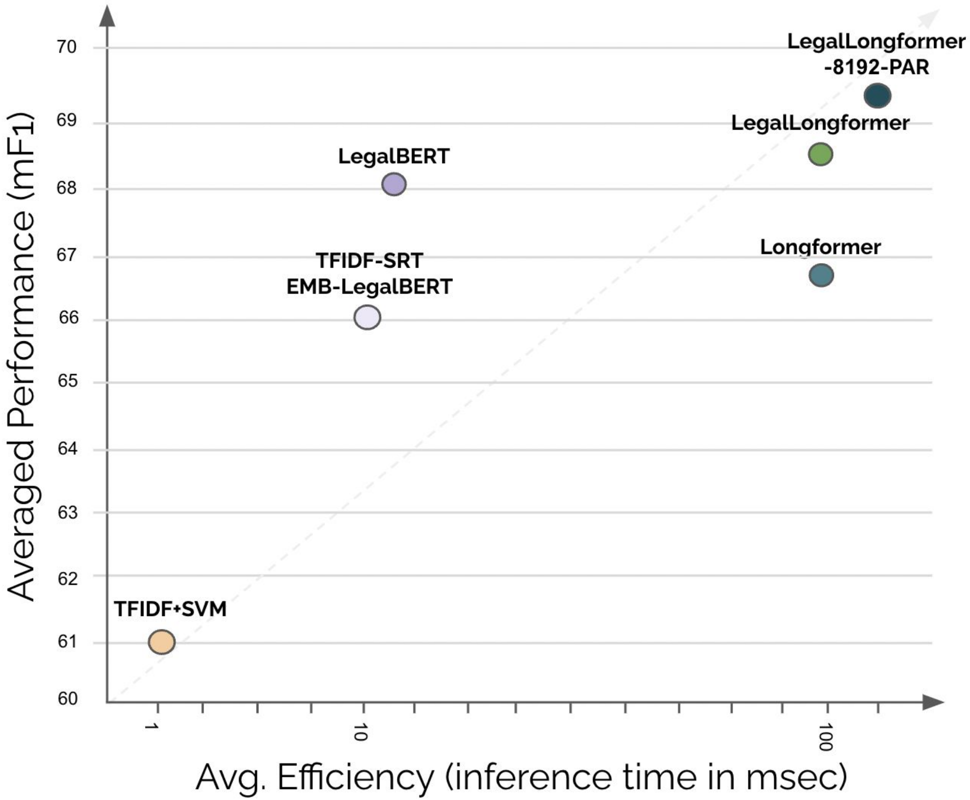
◇ Results on test data.

Method	ECtHR (Task A) *		ECtHR (Task B) *		SCOTUS *		EUR-LEX		LEDGAR		UNFAIR-ToS	
	μ-F1	m-F1	μ-F1	m-F1	μ-F1	m-F1	μ-F1	m-F1	μ-F1	m-F1	μ-F1	m-F1
Longformer	69.9	64.7	79.4	71.7	72.9	64.0	71.6	**57.7**	88.2	83.0	95.5	80.9
Longformer-8192	70.9	62.1	79.2	73.9	73.7	63.6	(Not considered for short-document tasks.)					
Longformer-8192-PAR	70.8	62.3	79.0	73.1	73.9	66.0						
LegalLongformer	**71.7**	63.6	80.5	**76.4**	76.6	66.9	**72.2**	56.6	**88.8**	**83.5**	95.7	80.6
LegalLongformer-8192	71.2	64.3	**81.4**	74.2	**77.5**	**67.3**	(Not considered for short-document tasks.)					
LegalLongformer-8192-PAR	71.4	**68.4**	79.6	73.9	76.2	66.3						

* The results were obtained using the hierarchical version of the corresponding model.

Best results

Performance - Efficiency tradeoff

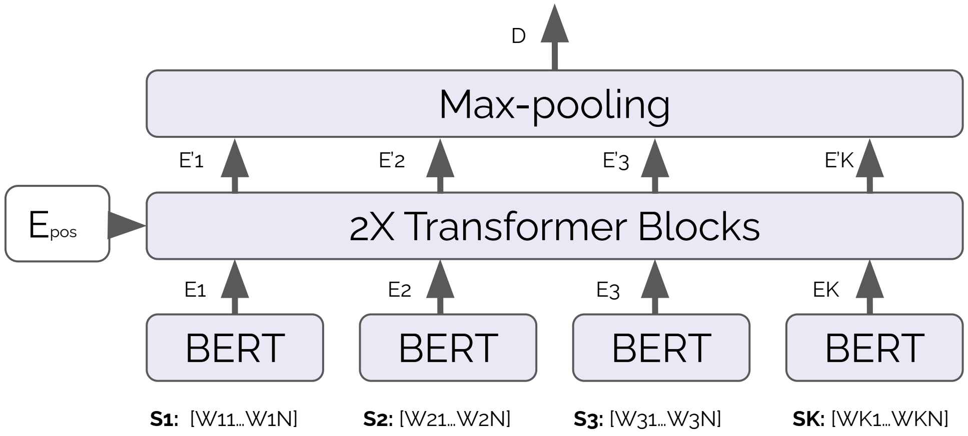


Thanks for your attention!

ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

KØBENHAVNS UNIVERSITET

Prior work



- Hierarchical Transformers
  - Hierarchical LegalBERT *(Chalkidis et al. 2020)*
  - Smith (Yang et al., 2020)
- Sparse-attention variants
  - Longformer *(Beltagy et al. 2020)*
  - BigBird
  - ETC

- Longformer *(Beltagy et al. 2020)*

Model params., memory footprint (GBs/sample), and inference time (sec/sample)

◇ Results on test data.

Method	Params.	ECtHR*		SCOTUS*		EUR-LEX		LEDGAR		UNFAIR-ToS	
		Mem.	Time	Mem.	Time	Mem.	Time	Mem.	Time	Mem.	Time
BoW models (word order lost)											
TFIDF-SVM	0.5M	0.1	.001	0.1	.001	0.1	.001	0.1	.001	0.1	.001
TFIDF-SRT-LegalBert	110M	0.9	.012	0.9	.012	0.9	.012	0.9	.007	0.9	.007
TFIDF-SRT-EMB-LegalBERT	110M	0.9	.012	0.9	.012	0.9	.012	0.9	.007	0.9	.007
LegalBERT variants that retain word order											
LegalBERT	110M	1.3	.014	1.3	.014	1.9	.012	1.9	.007	1.9	.007
TFIDF-EMB-LegalBERT	110M	1.3	.014	1.3	.014	1.9	.012	1.9	.007	1.9	.007

Model params., memory footprint (GBs/sample), and inference time (sec/sample)

◇ Results on test data.

Method	Params.	ECtHR*		SCOTUS*		EUR-LEX		LEDGAR		UNFAIR-ToS	
		Mem.	Time	Mem.	Time	Mem.	Time	Mem.	Time	Mem.	Time
Longformer variants (all retain word order)											
TFIDF-SVM	148M	1.7	.164	1.7	.164	1.3	.033	1.3	0.33	1.3	.033
TFIDF-SRT-LegalBert	151M	2.2	.318	2.2	.318	(Not considered for short-document class.)					
TFIDF-SRT-EMB-LegalBERT	151M	2.2	.331	2.2	.331						

LexGLUE Benchmark

Dataset	Source	Subdomain	Task Type	Instances	Classes
ECtHR Task A	Chalkidis et al. (2019)	ECHR	Multi-label classification	9,000 / 1,000 / 1,000	10+1◇
ECtHR Task B	Chalkidis et al. (2021a)	ECHR	Multi-label classification	9,000 / 1,000 / 1,000	10+1◇
SCOTUS	Spaeth et al. (2020)	US Law	Multi-class classification	5,000 / 1,400 / 1,400	14
EUR-LEX	Chalkidis et al. (2021b)	EU Law	Multi-label classification	55,000 / 5,000 / 5,000	100
LEDGAR	Tuggener et al. (2020)	Contracts	Multi-class classification	60,000 / 10,000 / 10,000	100
UNFAIR-ToS	Lippi et al. (2019)	Contracts	Multi-label classification	5,532 / 2,275 / 1,1607	8+1◇

◇ +1 means that some documents aren't relevant to any class.