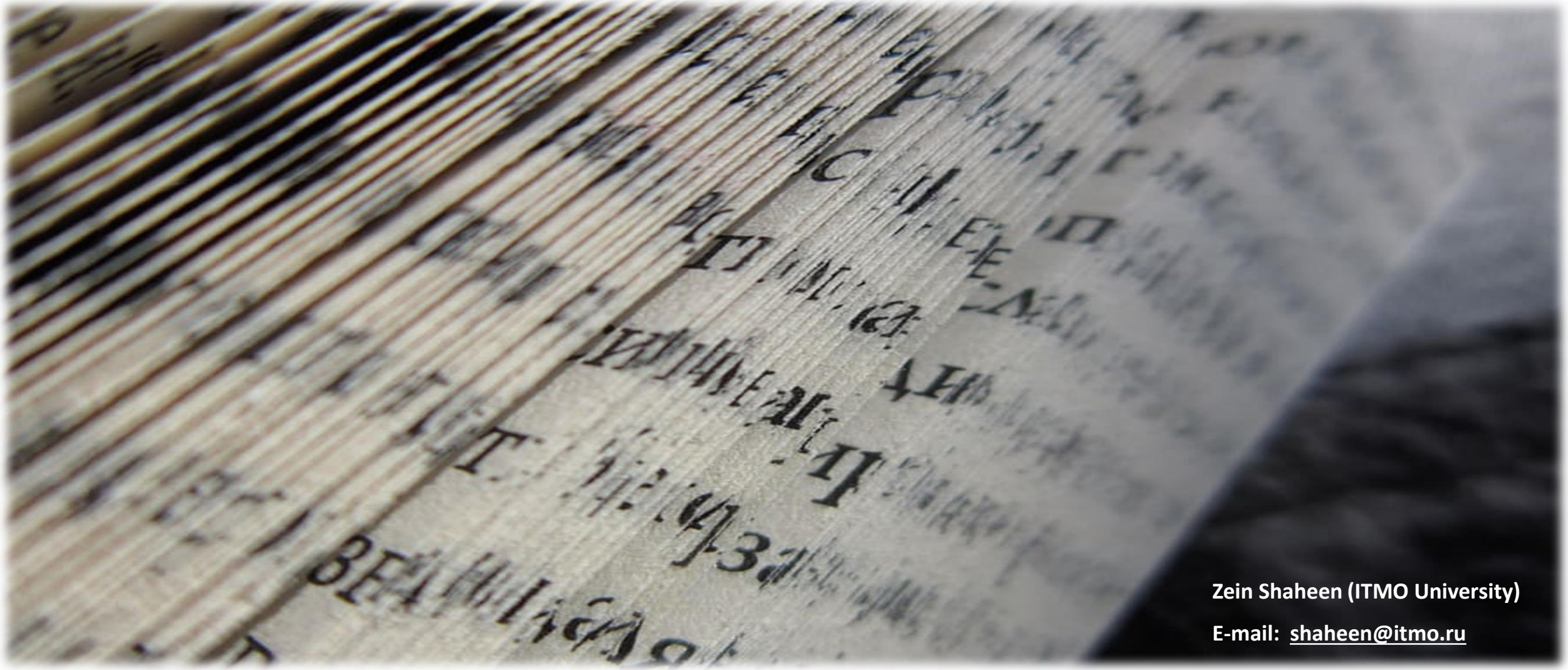
Large Scale Legal Text Classification Using Transformer Models

Zein Shaheen, Gerhard Wohlgenannt, Erwin Flitz



Zein Shaheen (ITMO University)

E-mail: shaheen@itmo.ru

ABOUT ME

- My name is Zein Shaheen.
- 2014: BCs in software engineering (Tishreen University, Syria).
- 2019: Master in applied computer science (SPbPU, Russia).
- 2019 till now: Ph.D Student at ITMO university, Russia. Department of Informatics and Applied Mathematics.
- Machine learning engineer in Huawei R&D center in Saint Petersburg.
- Research engineer in SPbPU.

AGENDA

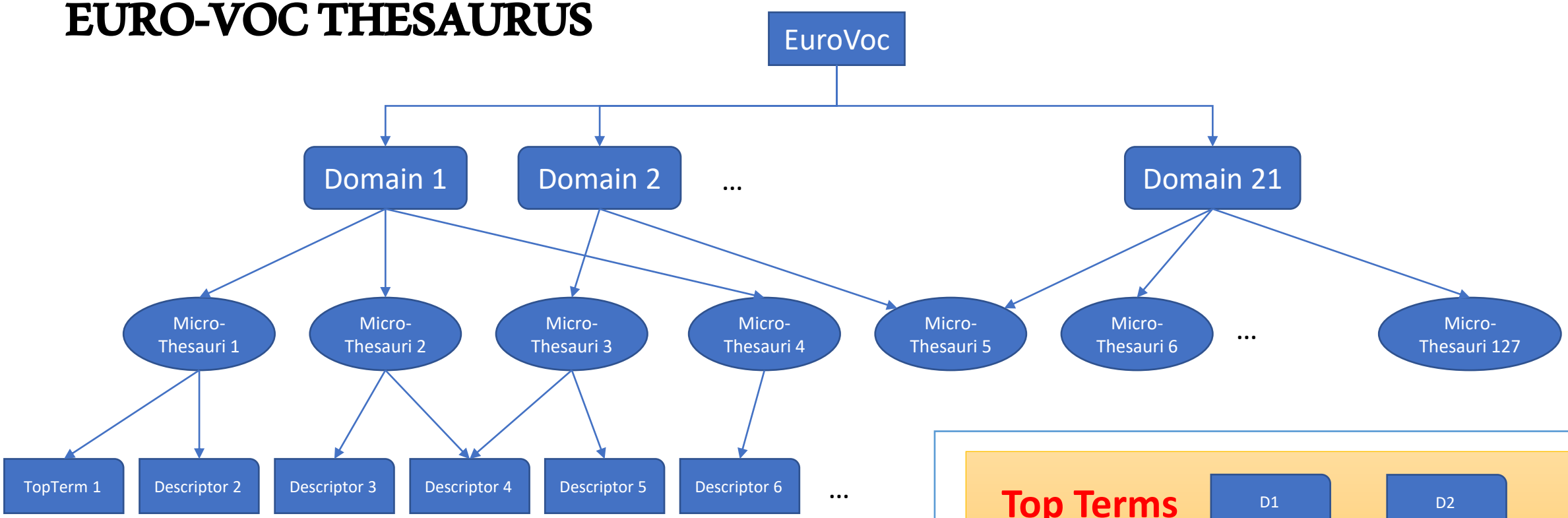
- Motivation
- EuroVoc thesaurus
- Datasets
- Hierarchical reduction method
- Models
- Training strategies
- Results
- Ablation studies
- Summary

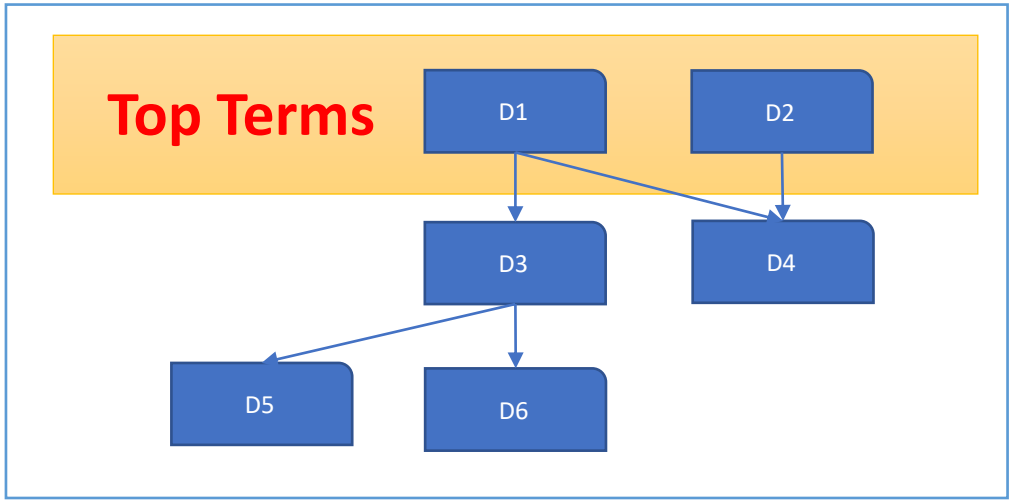
MOTIVATION

- Large Multi-Label text classification.
  - Long texts.
  - Thousands of labels
  - Power-law distribution for the labels.
  - Hierarchy of labels.
- Legal domain.
  - EuroVoc thesaurus.
- Transformer-based models.
- Training strategies:
  - Generative pre-training.
  - Graual unfreezing.
  - Discriminative learning rate.

EuroVoc

EURO-VOC THESAURUS





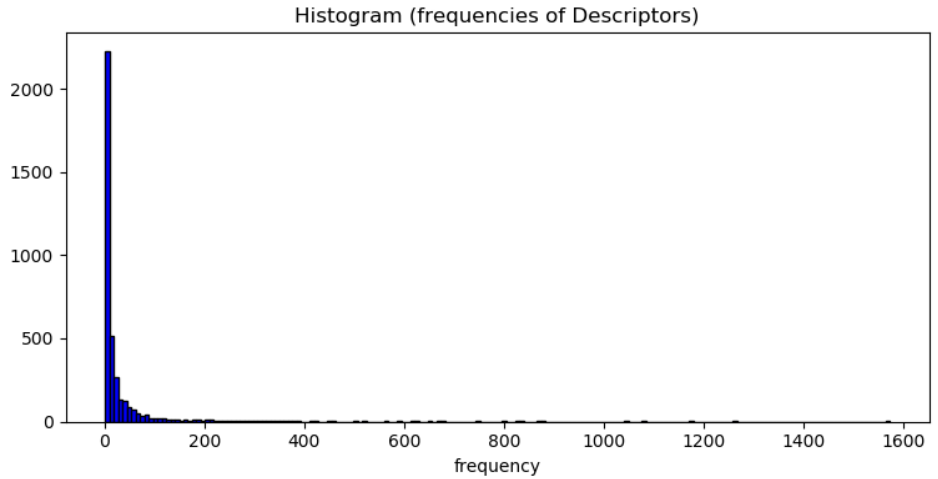
Descriptor-Hierarchy

DATASETS

- Eur-Lex database.
  - Official European Union Law.
  - 20 languages.
- Long documents.
- Power-law distribution for the labels.

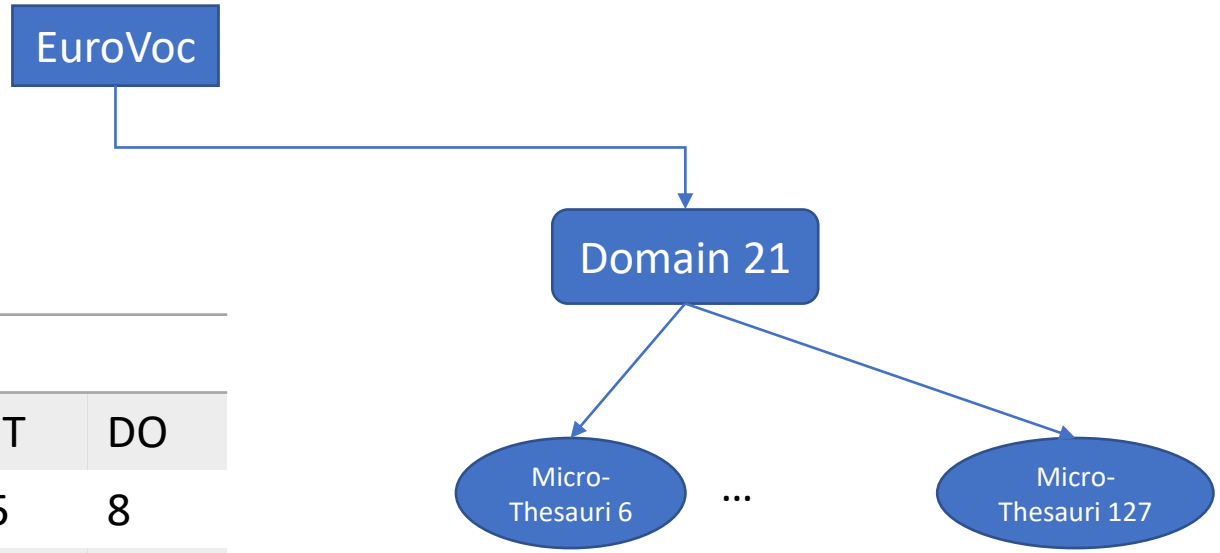
	JRC-Acquis	EURLEX57K
#Documents	20382	57000
Max #Tokens/Doc	469820	3934
Min #Tokens/Doc	21	119
Mean #Tokens/Doc	2243.43	758.46
StdDev #Tokens/Doc	7075.94	542.86
Median #Tokens/Doc	651	544

DATASET STATISTICS FOR JRC-AQUIS AND EURLEX57K.



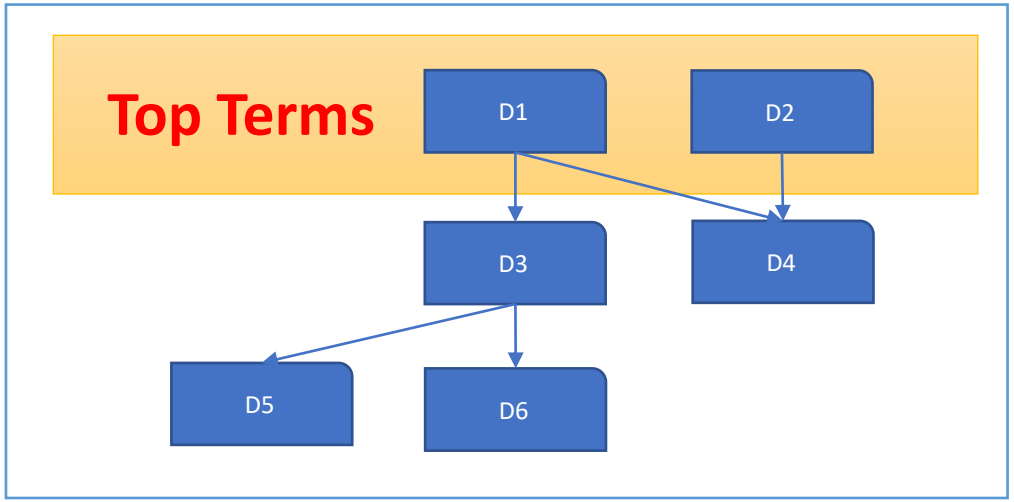
POWER-LAW DISTRIBUTION OF DESCRIPTORS IN THEJRC-ACQUIS DATASET.

HIERARICHAR REDUCTION



	JRC-Acquis				EURLEX57K			
Label	DE	TT	MT	DO	DE	TT	MT	DO
Max	24	30	14	10	26	30	15	8
Min	1	1	1	1	1	1	1	1
Mean	5.46	6.04	4.74	3.39	5.07	5.94	4.55	3.24
StdDev	1.73	3.14	1.92	1.17	1.7	3.06	1.82	1.04
Median	6	5	5	3	5	5	4	3
Mode	6	4	4	3	6	4	4	3

DATASET STATISTICS–NUMBER OF LABELS PER DOCUMENT



Descriptor-Hierarchy

MODELS

- AWD-LSTM base-line.
- We train MLTC classifier by finetuning pre-trained models from famous transformer-based architectures.
  - We extract a representation and build a classifier upon it.
  - Different transformer-based models, vary in size and training objectives.
- Additionally, We finetuned multi-lingual BERT on JRC-Aquis (French , English, German).

Model Name	# Layers	# Heads	Context Length	Is Cased	batch-size
BERT	12	12	512	False	4
Roberta	12	12	512	False	4
DistilBERT	6	12	512	False	4
XLNet	12	12	1024	True	2

ARCHITECTURE HYPERPARAMETERS OF TRANSFORMERMODELS

TRAINING STRATEGIES

- Discriminative fine-tuning: applies different learning rates depending on the layer; earlier layers use smaller learning rates compared to later layers.
- Slanted triangular learning rates.
- Gradual unfreezing: the training process is divided into multiple cycles, where each cycle consists of several training epochs.

Cycle	Max LR	# epochs	# Unfrozen Layers		
			BERT RoBERTa	DistilBERT	XLNet
1	2e-4	12	4	2	4
2	5e-5	12	8	4	6
3	5e-5	12	12	6	8
4	5e-5	12	12	6	8
5	5e-5	12	12	6	8

GRADUAL UNFREEZING DETAILS: LEARNINGRATES(LR),NUMBER OF EPOCHS(ITERS),AND LAYER GROUPS THAT ARE UNFROZEN.

Cycle	Mar LR	# Unfrozen Layers	# epochs
1	2e-1	1	2
2	1e-2	2	5
3	1e-3	3	5
4	5e-3	All	20
5	1e4	All	32
6	1e-4	All	32

GRADUAL UNFREEZING SETTINGS FORAWD-LSTM

RESULTS

	BERT	RoBERTa	XLNet	DistilBERT	AWD-LSTM	Multilingual BERT
Micro-F1	0.661	0.659	0.605	0.652	0.493	**0.663**
RP@1	0.867	0.873	0.845	**0.884**	0.762	0.873
RP@3	0.784	**0.788**	0.736	0.78	0.619	0.783
RP@5	0.715	0.716	0.661	0.711	0.548	**0.717**
RP@10	0.775	**0.778**	0.733	0.775	0.627	0.777
nDCG@1	0.867	0.873	0.845	**0.884**	0.762	0.873
nDCG@3	0.803	**0.807**	0.762	0.805	0.651	0.804
nDCG@5	0.750	**0.753**	0.703	0.75	0.594	0.752
nDCG@10	0.778	**0.781**	0.746	0.779	0.630	0.780

COMPARISON BETWEEN DIFFERENT TRANSFORMER MODELS,FINE-TUNED USING THE SAME NUMBER OF ITERATIONS ON JRC-ACQUIS.

RESULTS

	DE	TT	MT	DO
Micro-F1	0.661	0.745	0.778	0.839
RP@1	0.867	0.922	0.943	0.967
RP@3	0.784	0.838	0.871	0.905
RP@5	0.715	0.804	0.844	0.928
RP@10	0.775	0.857	0.908	0.974
nDCG@1	0.867	0.922	0.943	0.967
nDCG@3	0.803	0.858	0.888	0.919
nDCG@5	0.750	0.829	0.864	0.929
nDCG@10	0.778	0.852	0.896	0.952

BERTRESULTS FORJRC-ACQUIS WITHclass reductionMETHODS APPLIED,WHICH LEAD TO4DATASETS: DE (DESCRIPTORS), TT(TOP-TERMS), MT (MICROTHESAURI, DO (DOMAINS)

RESULTS

	Ours			**Chalkidis et al.**		
	BERT	RoBERTa	DistilBERT	BERT-BASE	BIGRU-LWAN-ELMO	BIGRU-LWAN-L2V
Micro-F1	0.751	**0.758**	0.754	0.732	0.719	0.709
RP@1	0.912	0.919	**0.925**	0.922	0.921	0.915
RP@3	0.843	0.85	0.848	-	-	-
RP@5	0.805	**0.812**	0.807	0.796	0.781	0.770
RP@10	0.852	0.860	**0.862**	0.856	0.845	0.836
nDCG@1	0.912	0.919	**0.925**	0.922	0.921	0.915
nDCG@3	0.859	**0.866**	**0.866**	-	-	-
nDCG@5	0.828	**0.835**	0.833	0.823	0.811	0.801
nDCG@10	0.849	0.857	**0.858**	0.851	0.841	0.832

RESULTS FOR OUR TRANSFORMER-BASED MODELS ONEURLEX57K,AND STRONG BASELINES FROMCHALKIDIS ET AL.

RESULTS

	DE	TT	MT	DO	DE (baseline)
Micro-F1	0.751	0.825	0.84	0.883	0.732
RP@1	0.912	0.948	0.959	0.978	0.922
RP@3	0.843	0.896	0.915	0.939	-
RP@5	0.805	0.876	0.902	0.956	0.796
RP@10	0.852	0.909	0.943	0.986	0.856
nDCG@1	0.912	0.948	0.959	0.978	0.922
nDCG@3	0.859	0.907	0.924	0.947	-
nDCG@5	0.828	0.891	0.912	0.955	0.823
nDCG@10	0.849	0.904	0.931	0.97	0.851

BERT RESULTS ON EURLEX57K WITH class reduction METHODS APPLIED, PLUS THE BASELINE RESULTS OFBERT-BASE (DE)FROMCHALKIDIS ET AL.

ABLATION STUDIES

	BERT	RoBERTa	DistilBERT
Micro-F1	0.64 (0.66)	0.65 (0.66)	0.61 (0.62)
RP@1	0.86 (0.87)	0.87 (0.87)	0.86 (0.87)
RP@3	0.77 (0.78)	0.77 (0.79)	0.75 (0.76)
RP@5	0.70 (0.72)	0.70 (0.72)	0.67 (0.68)
RP@10	0.76 (0.78)	0.77 (0.78)	0.74 (0.75)
nDCG@1	0.86 (0.87)	0.87 (0.87)	0.86 (0.87)
nDCG@3	0.79 (0.80)	0.79 (0.81)	0.77 (0.78)
nDCG@5	0.74 (0.75)	0.74 (0.75)	0.71 (0.72)
nDCG@10	0.77 (0.72)	0.77 (0.78)	0.75 (0.76)

CLASSIFICATION METRICS FOR THEJRC-ACQUIS DATASET,WHENnotUSINGLMFINE-TUNING—IN PARENTHESES THE RESULTSwithFINE-TUNING(FOR COMPARISON).

ABLATION STUDIES

	Iter.	Use GU	Prec.	Rec.	Mic.-F1
BERT	36	True	0.678	0.601	0.637
	108	False	0.674	0.575	0.621
	108	True	0.695	0.630	**0.661**
DistilBERT	36	True	0.696	0.601	0.645
	108	False	0.663	0.583	0.620
	108	True	0.701	0.611	**0.653**

ABLATION STUDY: BERT AND DISTILBERT PERFORMANCE ON JRC-ACQUIS REGARDING THE NUMBER OF TRAINING EPOCHS(ITER.) AND THE USE OF GRADUAL UNFREEZING(GU).

SUMMARY

- Attention mechanism used in transformers is superior to LSTMs in finding aspects relevant for the classification task in long documents.
- We measured little differences between BERT and RoBERTa.
- DistilBERT delivered surprisingly good results for the EURLEX57K dataset, and had the benefits of lower computational cost.
- XLNet require a lot of computational resources, and we were not able to properly train the model fot that reason.
- The first set of experiments on multilingual training with M-BERT gave promising results, it will be further studied in future work.
- Ablation studies showed the positive effects of the training strategies that we applied, both LM-finetuning on the target domain, as well as the gradual unfreezing proved to be crucial in reaching state-of –the-art classification performance.

THANK YOU