

Large Scale Legal Text Classification Using Transformer Models

Zein Shaheen
ITMO University
St. Petersburg, Russia
shaheen@itmo.ru

Gerhard Wohlgenannt
ITMO University
St. Petersburg, Russia
gwohlg@corp.ifmo.ru

Erwin Filtz
Vienna University of Economics and Business (WU)
Vienna, Austria
erwin.filtz@wu.ac.at

Abstract—Large multi-label text classification is a challenging Natural Language Processing (NLP) problem that is concerned with text classification for datasets with thousands of labels. We tackle this problem in the legal domain, where datasets, such as JRC-Acquis and EURLEX57K labeled with the EuroVoc vocabulary were created within the legal information systems of the European Union. The EuroVoc taxonomy includes around 7000 concepts. In this work, we study the performance of various recent transformer-based models in combination with strategies such as generative pretraining, gradual unfreezing and discriminative learning rates in order to reach competitive classification performance, and present new state-of-the-art results of 0.661 (F1) for JRC-Acquis and 0.754 for EURLEX57K. Furthermore, we quantify the impact of individual steps, such as language model fine-tuning or gradual unfreezing in an ablation study, and provide reference dataset splits created with an iterative stratification algorithm.

Keywords—multi-label text classification; legal document datasets; transformer models; EuroVoc.

I. INTRODUCTION

Text classification, i.e., the process of assigning one or multiple categories from a set of options to a document [1], is a prominent and well-researched task in Natural Language Processing (NLP) and text mining. Text classification variants include simple binary classification (for example, decide if a document is spam or not spam), multi-class classification (selection of one from a number of classes), and multi-label classification. In the latter, multiple labels can be assigned to a single document. In *Large Multi-Label Text Classification (LMTC)*, the label space is typically comprised of thousands of labels, which obviously raises task complexity. The work presented here tackles an LMTC problem in the legal domain.

LMTC tasks often occur when large taxonomies or formal ontologies are used as document labels, for example in the medical domain [2] [3], or when using large open domain taxonomies for labelling, such as annotating Wikipedia with labels [4]. A common feature of many LMTC tasks is that some labels are used frequently, while others are used very rarely (few-shot learning) or are never used (zero-shot learning). This situation is also referred to by *power-law* or *long-tail* frequency distribution of labels, which also characterizes our datasets and which is a setting that is largely unexplored for text classification [3]. Another difficulty often faced in LMTC datasets [3] are long documents, where finding the relevant areas to correctly classify documents is a needle in a haystack situation.

In this work, we focus on LMTC in the legal domain, based on two datasets, the well-known JRC-Acquis dataset [5] and the new EURLEX57K dataset [6]. Both datasets contain

legal documents from Eur-Lex [7], the legal database of the European Union (EU). The usage of language in the given documents is highly domain specific, and includes many legal text artifacts such as case numbers. Modern neural NLP algorithms often tackle domain specific text by fine-tuning pretrained language models on the type of text at hand [8]. Both datasets are labelled with terms from the the European Union’s multilingual and multidisciplinary thesaurus *EuroVoc* [9].

The goal of this work is to advance the state-of-the-art in LMTC based on these two datasets which exhibit many of the characteristics often found in LMTC datasets: power-law label distribution, highly domain specific language and a large and hierarchically organized set of labels. We apply current NLP transformer models, namely BERT [10], RoBERTa [11], DistilBERT [12], XLNet [13] and M-BERT [10], and combine them with a number of training strategies such as gradual unfreezing, slanted triangular learning rates and language model fine-tuning. In the process, we create new standard dataset splits for JRC-Acquis and EURLEX57 using an iterative stratification approach [14]. Providing a high-quality standardized dataset split is very important, as previous work was typically done on different random splits, which makes results hard to compare [15]. Further, we make use of the semantic relations inside the EuroVoc taxonomy to infer reduced label sets for the datasets. Some of our main evaluation results are the Micro-F1 score of 0.661 for JRC-Acquis and 0.754 for EURLEX57K, which sets new states-of-the-art to the best of our knowledge.

The main findings and contributions of this work are: (i) the experiments with BERT, RoBERTa, DistilBERT, XLNet, M-BERT (trained on three languages), and AWD-LSTM in combination with the training tricks to evaluate and compare the performance of the models, (ii) providing new standardized datasets for further investigation, (iii) ablation studies to measure the impact and benefits of various training strategies, and (iv) leveraging the EuroVoc term hierarchy to generate variants of the datasets for which higher classification performance can be achieved.

The remainder of the paper is organized as follows: After a discussion of related work in Section II, we introduce the EuroVoc vocabulary and the two datasets (Section III), and then present the main methods (AWD-LSTM, BERT, RoBERTa, DistilBERT, XLNet) in Section IV. Section V contains extensive evaluations of the methods on both datasets as well as ablation studies, and after a discussion of results (Section VI) we conclude the paper in Section VII.

II. RELATED WORK

In connection with the *JRC-Acquis* dataset, Steinberger et al. [16] present the “JRC EuroVoc Indexer JEX”, by the Joint Research Centre (JRC) of the European Commission. The tool categorizes documents using the EuroVoc taxonomy by employing a profile-based ranking task; the authors report an F-score between 0.44 and 0.54 depending on the document language. Boella et al. [17] manage to apply a support vector machine approach to the problem by transforming the multi-label classification problem into a single-label problem. Liu et al. [18] present a new family of Convolutional Neural Network (CNN) models tailored for multi-label text classification. They compare their method to a large number of existing approaches on various datasets; for the EurLex/JRC dataset however, another method (SLEEC), provided the best results. SLEEC (Sparse Local Embeddings for Extreme Classification) [19], creates local distance preserving embeddings which are able to accurately predict infrequently occurring (tail) labels. The results on precision for SLEEC applied in Liu et al. [18] are P@1: 0.78, P@3: 0.64 and P@5: 0.52 – however, they use a previous version of the JRC-Acquis dataset with only 15.4K documents.

Chalkidis et al. [6] recently published their work on the new EURLEX57K dataset. The dataset will be described in more detail (incl. dataset statistics) in the next sections. Chalkidis et al. also provide a strong baseline for LMTC on this dataset. Among the tested neural architectures operating on the full documents, they have best results with BIGRUs with label-wise attention. As input representation they use either GloVe [20] embeddings trained on domain text, or ELMO embeddings [21]. The authors investigated using only the first zones of the (long) documents for classification, and show that the title and recitals part of each document leads to almost the same performance as considering the full document [6]. This helps to alleviate BERT’s limitation of having a maximum of 512 tokens as input. Using only the first 512 tokens of each document as input, BERT [10] archives the best performance overall. The work of Chalkidis et al. is inspired by You et al. [22] who experimented with RNN-based methods with self attention on five LMTC datasets (RCV1, Amazon-13K, Wiki-30K, Wiki-500K, and EUR-Lex-4K). Similar work has been done in the medical domain, Mullenbach et al. [2] investigate label-wise attention in LMTC for medical code prediction (on the MIMIC-II and MIMIC-III datasets).

In this work, we experiment with BERT, RoBERTa, DistilBERT, XLNet, M-BERT and AWD-LSTM. We provide ablation studies to measure the impact of various training strategies and heuristics. Moreover, we provide new standardized datasets for further investigation by the research community, and leverage the EuroVoc term hierarchy to generate variants of the datasets.

III. DATASETS AND EUROVOC VOCABULARY

In this section, we first introduce the multilingual EuroVoc thesaurus which is used to classify legal documents published by the institutions of the European Union. The EuroVoc thesaurus is also used as a classification schema for the documents contained in the two legal datasets we use for our experiments, the *JRC-Acquis V3* and *EURLEX57K* datasets which are described in this section.

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> .
@prefix skos: <http://www.w3.org/2004/02/skos/core#> .
@prefix dcterms: <http://purl.org/dc/terms/> .
@prefix ev: <http://eurovoc.europa.eu/> .
@prefix evs: <http://eurovoc.europa.eu/schema#> .
<http://eurovoc.europa.eu/100142>
  rdf:type evs:Domain ;
  skos:prefLabel "04 POLITICS"@en .
<http://eurovoc.europa.eu/100166>
  rdf:type evs:MicroThesaurus ;
  skos:prefLabel "0421 parliament"@en ;
  dcterms:subject ev:100142 ;
  skos:hasTopConcept ev:41 .
<http://eurovoc.europa.eu/41>
  rdf:type evs:ThesaurusConcept ;
  skos:prefLabel "powers of parliament"@en ;
  skos:inScheme ev:100166 .
<http://eurovoc.europa.eu/1599>
  rdf:type evs:ThesaurusConcept ;
  skos:prefLabel "legislative period"@en ;
  skos:inScheme ev:100166
  skos:broader ev:41 .
```

Figure 1. EuroVoc example

A. EuroVoc

The datasets we use for our experiments contain legal documents from the legal information system of the European Union (Eur-Lex) and are classified into a common classification schema, the EuroVoc [9] thesaurus published and maintained by the Publications Office of the European Union since 1982. The EuroVoc thesaurus has been introduced to harmonize the classification of documents in the communications across EU institutions and to enable a multilingual search as the thesaurus provides all its terms in the official language of the EU member states. It is organized based on the *Simple Knowledge Organization System (SKOS)* [23], which encodes data using the *Resource Description Format (RDF)* [24] and is well-suited to represent hierarchical relations between terms in a thesaurus like EuroVoc. EuroVoc uses SKOS to hierarchically organize its concepts into 21 domains, for instance *Law*, *Trade* or *Politics*, to name a few. Each domain contains multiple microthesauri (127 in total), which in turn have in total around 600 top terms. About 7K terms (also called *descriptors*, *concepts* or *labels*) are assigned to one or multiple microthesauri and connected to top terms using the predicate `skos:broader`.

All concepts in EuroVoc have a *preferred* (`skos:prefLabel`) label and *non-preferred* (`skos:altLabel`) label for each language; the label language is indicated with language tags. Figure 1 illustrates with an example serialized in Turtle (TTL) [25] format how the terms are organized in the EuroVoc thesaurus. Our example is from the domain *04 POLITICS* and we show only the English labels of the concepts. The domain *04 POLITICS* has the EuroVoc ID `ev:100142` and is of `rdf:type evs:Domain`. Each domain has microthesauri as the next lower level in the hierarchy. In this example, we can see that a `evs:MicroThesaurus` named *0421 parliament* is assigned to the *04 POLITICS* domain using (`dcterms:subject ev:100142`) and is also connected to the next lower level of top terms. The top term *powers of parliament* (`ev:41`) is linked to the microthesaurus using `skos:inScheme`. Finally, the lowest level in this example is the concept *legislative period* (`ev:1599`) which is linked to its

(skos:broader) top term *powers of parliament* (ev:41), and is also directly linked to the microthesaurus *0421 parliament* to which it belongs to using *skos:inScheme*.

The legal documents are annotated with multiple EuroVoc classes typically on the lowest level which results in a huge amount of available classes a document can be potentially classified in. In addition, this also comes with the disadvantage of the power-law distribution of labels such that some labels are assigned to many documents whereas others are only assigned to a few documents or to no documents at all. The advantages of using a multilingual and multi-domain thesaurus for document classification are manifold. Most importantly, it allows us to reduce the numbers of potential classes by going up the hierarchy, which does not make classification incorrect but only more general. Reducing the number of labels allows to compare the efficiency of the model for different label sets, which vary in size and sparsity. In this line, we use a class reduction method to generate datasets with a reduced number of classes by replacing the original labels with the *top terms*, *microthesauri* or *domains* they belong to. For the top terms dataset, we leverage the *skos:broader* relations of the original descriptors, for the microthesauri dataset we follow *skos:inScheme* links to the microthesauri, and the domains dataset is inferred via the *dcterms:subject* links of the microthesauri. This process creates three additional datasets (*top terms*, *microthesauri*, *domains*) [26]. Furthermore, such a thesaurus would also allow to incorporate potentially more fine-grained national thesauri of member states which could be aligned with EuroVoc and therefore enable multilingual search in an extended thesaurus.

B. Legal Text Datasets

In this work we focus on legal documents collected from the Eur-Lex [7] database serving as the official site for retrieving European Union law, such as *Treaties*, *International agreements* and *Legislation*, and case law of the European Union (EU). Eur-Lex provides the documents in the official languages of the EU member states. As discussed in previous work [26] the documents are well structured and written in domain specific language. Furthermore, legal documents are typically longer compared to texts often taken for text classification task such as the Reuters-21578 dataset containing news articles.

In this paper, we use the English versions of the two legal datasets *JRC-AcquisV3* [27] and *EURLEX57K* [28]. The *JRC-Acquis V3* dataset has been compiled by the Joint Research Centre (JRC) of the European Union with the *Acquis Communautaire* being the applicable EU law and contains documents in XML format. Each JRC document is divided into body, signature, annex and descriptors. The *EURLEX57K* dataset has been prepared by academia [6] and is provided in JSON format structured into several parts, namely the header including title and legal body, recitals (legal background references), the main body (organized in articles) and the attachments (appendices, annexes). Furthermore and in contrast to JRC-Acquis, the *EURLEX57K* dataset is already provided with a split into train and test sets.

Table I shows a comparison of the dataset characteristics. *EURLEX57K* contains almost three times as many documents

TABLE I. DATASET STATISTICS FOR JRC-ACQUIS AND EURLEX57K.

| | 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 | 651.0 | 544 |
| Mode #Tokens/Doc | 275 | 275 |

as the *JRC-Acquis V3* dataset, but the documents are comparable in their minimum number of tokens, median and mode of tokens per document. The large difference in the maximum number of tokens per document impacts the standard deviation and the mean number of tokens. The reason for this difference is that JRC-Acquis also includes documents dealing with the budget of the European Union, comprised of many tables. As both datasets originate from the same source, but with different providers, we analyzed the number of documents contained in both datasets and found an overlap of approx. 12%.

Table II provides an overview of label statistics for both datasets. We created different versions based on the original descriptors (DE), top terms (TT), microthesauri (MT) and domains (DO) and present the numbers for all versions. The maximum number of labels assigned to a single document is similar for both datasets. The average number of labels per document in the original (DE) version is 5.46 (JRC-Acquis) and 5.07 (EURLEX57). Due to the polyhierarchy in the geography domain a label may be assigned to multiple *Top Terms*, therefore the number of *Top Term* labels is higher than that of the original descriptors.

Figure 2 visualizes the power-law (long tail) label distribution, where a large portion of EuroVoc descriptors is used rarely (or never) as document annotations. In the JRC-Acquis dataset only 50% of the labels available in EuroVoc are used to classify documents. Only 417 labels are used frequently (used on more than 50 documents) and 3,3147 labels have a frequency between 1–50 (few-shot). The numbers for the EURLEX57K dataset are similar [6], with 59.31% of all EuroVoc labels being actually present in EURLEX57K. From those labels, 746 are frequent, 3,362 have a frequency between 1–50, and 163 are only in the testing, but not in the training, dataset split (zero-shot). The high number of infrequent labels obviously is a challenge when using supervised learning approaches.

IV. METHODS

In this section we describe the methods used in the LMTC experiments presented in the evaluation section, and the general training process. Furthermore, we discuss important related points such as language model pretraining and fine-tuning, and discriminative learning rates, and other important foundations for the evaluation section like dataset splitting and multilingual training.

A. General Training Strategy and Implementation

In accordance with common NLP practice, as first introduced by Howard and Ruder for text classification [29], we

TABLE II. DATASET STATISTICS – NUMBER OF LABELS PER DOCUMENT.

| | JRC-Acquis | | | | EURLEX57K | | | |
|--------|------------|------|------|------|-----------|------|------|------|
| Label | DE | TT | MT | DO | DE | TT | MT | DO |
| Max | 24 | 30 | 14 | 10 | 26 | 30 | 15 | 9 |
| 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 |

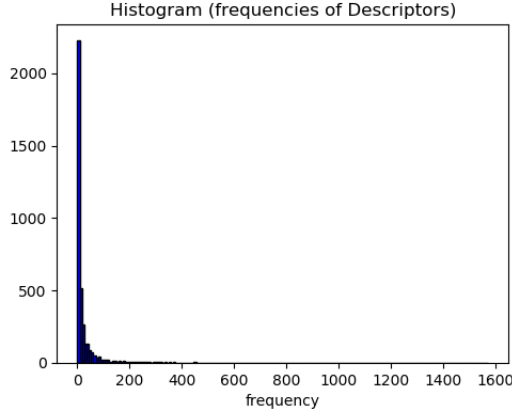


Figure 2. Power-law distribution of descriptors in the JRC-Acquis dataset.

train our models in two steps: first we fine-tune the language modeling part of the model to the target corpus (JRC-Acquis or EURLEX57K), and then we train the classifier on the training-split of the dataset.

The baseline model (AWD-LSTM) and the transformer models are available with pretrained weights, trained with language modelling objectives on large corpora such as Wikitext or Webtext – a process that is computationally very expensive. Fine-tuning allows to transfer the language modeling capabilities to a new domain [29].

Our implementation makes use of the FastAI library [30], which includes the basic infrastructure to apply training strategies like gradual unfreezing or slanted triangular learning rates (see below). Moreover, for the transformer models, we integrate the Hugging Face transformers package [31] with FastAI.

Our implementation including the evaluation results, is available on GitHub [32]. The repository also includes the reference datasets created with iterative splitting, which can be used by other researchers as reference datasets – in order to have a fair comparison of different approaches in the future.

B. Tricks for Performance Improvement (within FastAI)

In their Universal Language Model Fine-tuning for Text Classification (ULMFiT) approach, Howard and Ruder [29] propose a number of training strategies and tricks to improve model performance, which are available within the FastAI library. Firstly, based on the idea that early layers in a deep neural network capture more general and basic features of

language, which need little domain adaption, *discriminative fine-tuning* applies different learning rates depending on the layer; earlier layers use smaller learning rates compared to later layers. Secondly, *slanted triangular learning rates* quickly increase the learning rate at the beginning of a training epoch up to the maximal learning rate in order to find a suitable region of the parameter space, and then slowly reduce the learning rate to refine the parameters. And finally, in *gradual unfreezing* the training process is divided into multiple cycles, where each cycle consists of several training epochs. Training starts after freezing all layers except for the last few layers in cycle one, during later cycles more layers are unfrozen gradually (from last to first layers). The intuition is that, in fine-tuning a deep learning model (similar to discriminative fine-tuning), that later layers are more task and domain specific and need more fine-tuning. In the evaluation section, we provide details about our unfreezing strategy (Table IV).

C. Baseline Model

We use **AWD-LSTM** [33] as a baseline model. Merity et al. [33] investigate different strategies for regularizing word-level LSTM language models, including the *weight-dropped LSTM* with its recurrent regularization, and they introduce NT-ASGD as a new version of average stochastic gradient descent in AWD-LSTM.

In the ULMFiT approach [29] of FastAI, AWD-LSTM is used as encoder, with extra layers added on top for the classification task.

For any of the models (AWD-LSTM and transformers) we apply the basic method discussed above: a) fine-tune the language model on all documents (ignoring the labels) of the dataset (JRC-Acquis or EURLEX57K), and then b) fine-tune the classifier using the training-split of the dataset.

D. Transformer Models

In the experiments we study the performance of BERT, RoBERTa, DistilBERT and XLNet on the given text classification tasks. BERT is an early, and very popular, transformer model, RoBERTa is a modified version of BERT trained on a larger corpus, DistilBERT is a distilled version of BERT and thereby with lower computational cost, and finally, XLNet can be fed with larger input token sequences.

BERT: BERT [10] is a bidirectional language model which aims to learn contextual relations between words using the transformer architecture [34]. We use an official release of the pre-trained models, details about the specific hyperparameters are found in Section V-A.

The input to BERT is either a single text (a sentence or document), or a text pair. The first token of each sequence is the special classification token [CLS], followed by WordPiece tokens of the first text A , then a separator token [SEP], and (optionally) after that WordPiece tokens for the second text B .

In addition to token embeddings, BERT uses positional embeddings to represent the position of tokens in the sequence. For training, BERT applies Masked Language Modeling (MLM) and Next Sentence Prediction (NSP) objectives. In MLM, BERT randomly masks 15% of all WordPiece tokens in each sequence and learns to predict these masked tokens. For NSP, BERT is fed in 50% of cases with the actual next sentence B , in the other cases with a random sentence B from the corpus.

RoBERTa: RoBERTa, introduced by Liu et al. [11], re-trains BERT with an improved methodology, much more data, larger batch size and longer training times. In RoBERTa the training strategy of BERT is modified by removing the NSP objective. Further, RoBERTa uses byte pair encoding (BPE) as a tokenization algorithm instead of WordPiece tokenization in BERT.

DistilBERT: We use a distilled version of BERT released by Sanh et al. [12]. DistilBERT provides a lighter and faster version of BERT, reducing the size of the model by 40% while retaining 97% of its capabilities on language understanding tasks [12]. The distillation process includes training a complete BERT model (the teacher) using the improved methodology proposed by Liu et al. [11], then DistilBERT (the student) is trained to reproduce the behaviour of the teacher by using cosine embedding loss.

XLNet: The previously discussed transformer-based models are limited to a fixed context length (such as 512 tokens), while legal documents are often long and exceed this context length limit. XLNet [13] includes segments recurrence, introduced in Transformer-XL [35], allowing it to digest longer documents. XLNet follows RoBERTa in removing the NSP objective, while introducing a novel permutation language model objective. In our work with XLNet, we fine-tune the classifier directly without LM fine-tuning (as LM fine-tuning of XLNet was computationally not possible on the hardware available for our experiments).

E. Dataset Splitting

Stratification of classification data aims at splitting the data in a way that in all dataset splits (training, validation, test) the target classes appear in similar proportions. In multi-label text classification *stratification* becomes harder, because the target is a combination of multiple labels. In *random splitting*, it is possible that most instances of a specific class end up either in the training or test split (esp. for low frequency classes), and therefore the split can be unrepresentative with respect to the original data set. Moreover, random splitting and different train/validation/test ratios create the problem that results from different approaches are hard to compare [15].

Depending on the dataset, other criteria can be used for dataset splitting, for example Azarbonyad et al. [36] split JRC-Acquis documents according to document’s year, where older documents could be used in training, and newer in testing.

For splitting both JRC-Acquis and EURLEX57K, we use the iterative stratification algorithm proposed by Sechidis et al. [14], ie. its implementation provided by the scikit-multilearn library [37]. Applying this algorithm leads to a better document split with respect to the target labels, and in turn, helps with generalization of the results and allows for a fair comparison of different approaches. The reference splits of the dataset are available online [32].

In the experiments in Section V we use these dataset splits, but in addition for EURLEX57K also the dataset split of the dataset creators [6], in order to compare to their evaluation results.

F. Multilingual Training

JRC-Acquis is a collection of parallel texts in 22 languages – we make use of this property to train multilingual BERT [38] on an extended version of JRC-Acquis in 3 languages. Multilingual BERT provides support for 104 languages and it is useful for zero-shot learning tasks in which a model is trained using data from one language and then used to make inference on data in other languages.

We extend the English JRC-Acquis dataset with parallel data in German and French. The additional data has the same dataset split as in the English version, ie. if an English document is in the training set then the German and French versions will be in the same split as well.

V. EVALUATION

This section first discusses evaluation setup (for example model hyperparameters) and then evaluation results for JRC-Acquis and EURLEX57K.

A. Evaluation Setup

Evaluation setup includes important aspects such as dataset splits, preprocessing, the specific model architectures and variants, and major hyperparameters used in training.

a) Dataset Splits:: The official JRC-Acquis dataset does not include a standard train-validation-test split, and as discussed in Section IV-E a random split exhibits unfavorable characteristics. We apply iterative splitting [14] to ensure that each split has the same label distribution as the original data. We split with an 80%/10%/10% ratio for training/validation/test sets. For the EURLEX57K the dataset creators already provide a split and a strong baseline evaluation. We run our models on the given split in order to compare results, and also create our own split with iterative splitting (dataset available in the mentioned GitHub repository [32]).

b) Text Preprocessing:: All described models have their own preprocessing included (e.g. WordPiece tokenization in BERT), we do not apply extra preprocessing to the text.

c) Neural Network Architectures:: For **AWD-LSTM**, we use the standard setup of the pretrained model included in FastAI, which has an input embedding layer with embedding size of 400, followed by three LSTM layers with hidden sizes of 1152 and weight dropout probability of 0.1.

TABLE III. ARCHITECTURE HYPERPARAMETERS OF TRANSFORMER MODELS

| 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 |

For the transformer models, we start from pretrained models, the uncased BERT model [39], the RoBERTa model [40], DistilBERT [41], and the XLNET model [42].

In Table III, we see that many architectural details are similar for the different model types. The transformer models all have 12 network layers, except DistilBERT with 6 layers, and 12 attention heads. XLNet allows for longer input contexts, but for performance reasons we limited the context to 1024 tokens, and it was necessary to reduce the batch size to 2 to fit the model into GPU memory, and also we could not unfreeze the whole pretrained model (see below).

To create the text classifiers, we take the representation of the text generated by the transformer model or AWD-LSTM, and add two fully connected layers of size 1200 and 50, respectively, with a dropout probability of 0.2, and an output layer. We apply batch normalization on the fully connected layers.

d) Gradual Unfreezing:: Gradual unfreezing is one of the ULMFiT strategies discussed in Section IV-B, where the neural network layers are grouped, and trained starting with the last group, then incrementally unfrozen and trained further.

TABLE IV. GRADUAL UNFREEZING DETAILS: LEARNING RATES (LR), NUMBER OF EPOCHS (ITERS), AND LAYER GROUPS THAT ARE UNFROZEN.

| Cycle | Max LR | # Iters | # 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 | 36 | 12 | 6 | 8 | |
| 5 | 5e-5 | 36 | 12 | 6 | 8 | |

Except for DistilBERT, which has only 2 layers per layer group, all transformer models have 3 groups of 4 layers used in the unfreezing process. Table IV gives an overview of the training setup for the transformer models. We trained the classifier for 5 cycles, starting in cycle 1 with 4 layers and a $LR = 2e - 4$, and 12 training epochs (Iters). The setup of the other cycles is shown in the table. Overall, we used the same setup for all transformer models with a goal of better comparison between models. (Remark: hand-picking LRs and training epochs might lead to slightly better results.)

Table V shows the main hyperparameters of AWD-LSTM training, we trained the model in 6 cycles, with LRs, epochs

TABLE V. GRADUAL UNFREEZING SETTINGS FOR AWD-LSTM

| Cycle | # Max LR | # Unfrozen Layers | # Iterations |
|-------|----------|-------------------|--------------|
| 1 | 2e-1 | 1 | 2 |
| 2 | 1e-2 | 2 | 5 |
| 3 | 1e-3 | 3 | 5 |
| 4 | 5e-3 | all | 20 |
| 5 | 1e-4 | all | 32 |
| 6 | 1e-4 | all | 32 |

per cycle, and unfrozen layers as shown in the table.

e) LM Fine-tuning:: For the transformer models we do LM fine-tuning for 5 iterations, with a batch size of 4 and LR of $5e - 5$. Transformer fine-tuning is done with a script¹ provided by Hugging Face. For the AWD-LSTM model we first fine-tune the frozen LM for 2 epochs, and then in cycle two fine-tune the unfrozen model for another 5 epochs.

f) Hardware specifications: We trained the models on a single GPU device (NVIDIA GeForce GTX 1080 with 11 GB of GDDR5X memory). For inference, we use an Intel i7-8700K CPU @ 3.70GHz and 16GB RAM.

B. Evaluation Metrics

In the evaluations, in line with Chalkidis et al. [6], we apply the following evaluation metrics: *micro-averaged F1*, *R-Precision@K (RP@K)*, and *Normalized Discounted Cumulative Gain (nDCG@K)*. *Precision@K (P@K)* and *Recall@K (R@K)* are popular measures in LTMC, too, but they unfairly penalize in situations where the number of gold labels is unequal to K , which is the typical situation in our datasets. This problem led to the introduction of more suitable metrics like $RP@K$ and $nDCG@K$. In the following, we briefly discuss the metrics.

The $F1$ -score is a common metric in information retrieval systems, and it is calculated as the harmonic mean between precision and recall. If we have a label L , Precision, Recall, and $F1$ -score with respect to L are calculated as follows:

$$Precision_L = \frac{TruePositives_L}{TruePositives_L + FalsePositives_L}$$

$$Recall_L = \frac{TruePositives_L}{TruePositives_L + FalseNegatives_L}$$

$$F1_L = 2 * \frac{Precision * Recall}{Precision + Recall}$$

Micro-F1 is an extension of the $F1$ -score for multi-label classification tasks, and it treats the entire set of predictions as one vector and then calculates the $F1$. We use grid search to pick the *threshold* on the output probabilities of the models that gives the best Micro-F1 score on the validation set. The threshold determines which labels we assign to the documents.

Propensity scores prioritize predicting a few relevant labels over the large number of irrelevant ones [43]. R -Precision@K ($RP@K$) calculates precision for the top K ranked labels, if the number of ground truth labels for a document is less than K , K is set to this number for this document.

$$RP@K = \frac{1}{N} \sum_{n=1}^N \sum_{k=1}^K \frac{Rel(n,k)}{\min(K, R_n)}$$

Where N is the number of documents, $Rel(n, k)$ is set to 1 if the k -th retrieved label in the top- K labels of the n -th

¹https://github.com/huggingface/transformers/blob/master/examples/language-modeling/run_language_modeling.py

document is correct, otherwise it is set to 0. R_n is the number of ground truth labels for the n -th document.

Normalized Discounted Cumulative Gain $nDCG@k$ for the list of top K ranked labels measures ranking quality. It is based on the assumption that highly relevant documents are more useful than moderately relevant documents.

$$nDCG@K = \frac{1}{N} \sum_{n=1}^N Z_{k_n} \sum_{k=1}^K \frac{2^{Rel(n,k)} - 1}{\log_2(1+k)}$$

N is the number of documents, $Rel(n, k)$ is set to 1 if the k -th retrieved label in the top- K labels of the n -th document is correct, otherwise it is set to 0. Z_{k_n} is a normalization factor to ensure $nDCG@K = 1$ for a perfect ranking.

C. Evaluation Results

The evaluation results are organized into three subsections, results for the JRC-Acquis dataset, results for the EURLEX57K dataset, and finally results from ablation studies.

1) *JRC-Acquis*: Table VI presents an overview of the results on the JRC-Acquis dataset for the transformer models and the AWD-LSTM baseline, and initial results from the multilingual model.

The observations here are as follows: Firstly, transformer-based models outperform the LSTM baseline by a large margin. Further, within the transformer models RoBERTa and BERT yield best results, the scores are almost the same. As expected, the distilled version of BERT is a bit lower in most metrics like Micro-F1, but the difference is small.

In this set of experiments, XLNet is behind DistilBERT, which we attribute to two main causes: (i) for computational reasons (given the available GPU hardware), we could *not* fine-tune the LM on XLNet, and in classifier training we could *not* unfreeze the full model. (ii) We used the same LR on all models; the choice of LR was influenced by a recommendation on BERT learning rates in Devlin et al. [10], and may not be optimal for XLNet. Overall, we could not properly test XLNet due to its high computational requirements, and did therefore not include it in the set of experiments on the EURLEX57K dataset.

The initial set of experiments with multilingual BERT (M-BERT) provides very promising results, on par with RoBERT and BERT. This is remarkable given the fact that we use the same amount of global training steps – which means, because our multilingual dataset is 3 times larger, that on individual documents we train only a 1/3 of the time. We expect even better results with more training epochs. LM fine-tuning of the M-BERT model was done on the text from all three languages (en, de, fr).

Regarding comparisons to existing baseline results, firstly because of the problem of different dataset splits (see Section IV-E) results are hard to compare. However, Steinberger et al. [16] report an F1-score of 0.48, Esuli et al. [44] report an F1 of 0.589 and Chang et al. [15] do not provide F1, but only P@5 (62.64) and R@5 (61.59).

For Table VII, we picked one transformer-based method, namely BERT, and analyzed its performance on the various JRC datasets resulting from *class reduction* described in Section III-A. By using inference on the EuroVoc hierarchy, we

created, additionally to the default descriptors dataset, datasets for EuroVoc Top Terms (TT), Micro-Thesauri (MT), and EuroVoc Domains (DO). With the reduced number of classes, classification performance is clearly rising, for example from a Micro-F1 of 0.661 (descriptors) to 0.839 (EuroVoc domains). We argue that the results with the inferred labels show that our approach might be well-suitable for real-world applications in scenarios like automatic legal document classification or keyword/label suggestion – for example the RP@5 for domains (DO) is at 0.928, so the classification performance (depending on the use case requirements) may be sufficient.

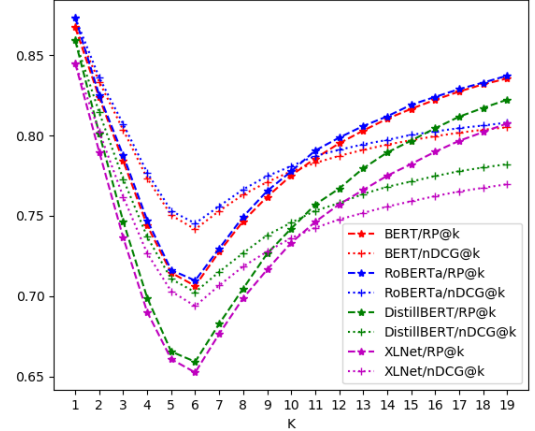


Figure 3. A visualization of RP@K and nDCG@K for all transformer models for JRC-Acquis.

Figure 3 contains a visual representation of RP@K and nDCG@K for the transformer models applied to the JRC-Acquis dataset. We can see how similar the performance of BERT and RoBERTa is for different values of K , and RoBERTa scores are consistently marginally better.

2) *EURLEX57K*: In this subsection we report the evaluation results on the new EURLEX57K dataset by Chalkidis et al. [6]. In order to compare to the results of the dataset creators, we ran the experiments on the dataset and dataset split (45K training, 6K validation, 6K testing) provided by Chalkidis et al. [6]. Below, we also show evaluation results on our dataset split (created with the iterative stratification approach). Table VIII gives an overview of results for our transformer models, and compares them to the strong baselines in existing work. Chalkidis et al. [6] evaluate various architectures, the results of the three best models presented here: BERT-BASE, BIGRU-LWAN-ELMO and BIGRU-LWAN-L2V. BERT-BASE is a BERT model with an extra classification layer on top, BIGRU-LWAN combines a BIGRU encoder with Label-Wise Attention Networks (LWAN), and uses either Elmo (ELMO) or word2vec (L2V) embeddings as inputs. Table VIII shows that our models outperform the previous baseline, the best results are delivered by RoBERTa and DistilBERT. The good performance of DistilBERT in these experiments is surprising (We need further future experiments to explain the results sufficiently. One intuition might be that the random weight initialization of the added layers was very suitable.).

Overall, the results are much better than for the smaller

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

| | 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 |

TABLE VII. BERT RESULTS FOR JRC-ACQUIS WITH *class reduction* METHODS APPLIED, WHICH LEAD TO 4 DATASETS: DE (DESCRIPTORS), TT (TOP-TERMS), MT (MICROTHESAURI, DO (DOMAINS))

| | 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 |

JRC dataset, with the best Micro-F1 for JRC being 0.661 (BERT), while for EURLEX57K we reach 0.758 (RoBERTa).

Table IX presents the results for BERT on the additional datasets with Top Terms (TT), Micro-Thesauri (MT) and Domains (DO) labels inferred from the EuroVoc taxonomy (similar to Table VII, which presents the scores of JRC-Acquis). As expected from the general results on the EURLEX57 dataset, the values on the derived datasets are better than for JRC-Acquis, for example RP@5 is now at 0.956 for the domains (DO).

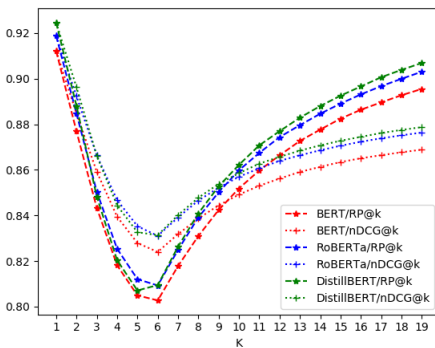


FIGURE 4. RP@K AND NDCG@K FOR THE TRANSFORMER MODELS TRAINED ON EURLEX57K.

Similar to Figure 3, Figure 4 shows RP@K and nDCG@K for BERT, RoBERTa and DistilBERT depending on the value of K . RoBERTa and DistilBERT are almost identical in their performance, BERT lags behind a little in this set of

experiments.

Finally, in Table X, we trained a BERT model on our iterative split of the EURLEX57K dataset in order to provide a strong baseline for future work on a standardized and arguably improved version of the EURLEX57K dataset.

3) *Ablation Studies*: In this section, we want to study the contributions of various training process components – by excluding some of those components individually (or reducing the number of training epochs). We focus on three important aspects: (i) the use of Language Model (LM) fine-tuning, (ii) gradual unfreezing, (iii) and a reduction of the number of training cycles.

In Table XI, we compare the evaluation metrics when removing the LM fine-tuning (on the legal target corpus) step before classification model training to the original version including LM fine-tuning (in parenthesis). For all examined models, we can see a small but consistent improvement of results when using LM fine-tuning. The relative improvement in the metrics is in the range of 1%–3%. In conclusion, LM fine-tuning to the legal text corpus is a crucial step for reaching a high classification performance.

In Table XII, we examine the effect of two factors, the training epochs (Iter.) hyperparameter, and of the use of the gradual unfreezing technique. Regarding number of epochs, both models benefit from longer training, for BERT the difference is large (about 4% relative improvement in F1-score), while for the simpler DistilBERT model less training appears to be required, after 36 epochs it even provides better accuracy than BERT at this point, and finally only gains a 1.2% improvement from more training epochs. Secondly, we study the effect of Gradual Unfreezing (GU), which for BERT has a large impact, with a relative improvement in F1 of about 6%. In summary, longer training times benefit esp. more complex models like BERT, and gradual unfreezing is a very helpful strategy for optimizing performance.

VI. DISCUSSION

Much of the detailed discussion is already included in the *Evaluation Results* section (Section V-C), so here we will summarize and extend on some of the key findings.

In comparing model performance, starting with LSTM versus transformer architectures, the results show that the attention mechanism used in transformers is superior to LSTMs in finding aspects relevant for the classification task in long documents. Within the transformer models, firstly we did not

TABLE VIII. RESULTS FOR OUR TRANSFORMER-BASED MODELS ON EURLEX57K, AND STRONG BASELINES FROM CHALKIDIS ET AL.

| | Ours | | | Chalkidis et al. [6] | | |
|----------|-------|--------------|--------------|----------------------|-----------------|----------------|
| | 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 |

TABLE IX. BERT RESULTS ON EURLEX57K WITH *class reduction* METHODS APPLIED, PLUS THE BASELINE RESULTS OF BERT-BASE (DE) FROM CHALKIDIS ET AL. [6].

| | 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 |

TABLE X. BERT RESULTS ON EURLEX57K WITH THE NEW ITERATIVE STRATIFICATION DATASET SPLIT.

| Micro-F1 | RP@1 | RP@5 | nDCG@1 | nDCG@5 |
|----------|-------|-------|--------|--------|
| 0.760 | 0.914 | 0.809 | 0.914 | 0.833 |

TABLE XI. CLASSIFICATION METRICS FOR THE JRC-ACQUIS DATASET, WHEN *not* USING LM FINE-TUNING – IN PARENTHESES THE RESULTS *with* FINE-TUNING (FOR COMPARISON).

| | 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) |

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

| | # 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 |
| Distil-BERT | 36 | True | 0.696 | 0.601 | 0.645 |
| | 108 | False | 0.663 | 0.583 | 0.620 |
| | 108 | True | 0.701 | 0.611 | 0.653 |

notice much difference between BERT and RoBERTa, which is not unexpected, as they are technically very similar. Overall, results were a bit better for RoBERTa. DistilBERT delivered surprisingly good results for the EURLEX57K dataset, and has the benefits of lower computational cost. Both for the JRC-Aquis and the EURLEX57K datasets, the results indicate that DistilBERT is better in retrieving the most probable label compared with RoBERTa and BERT. XLNet on the other hand, requires a lot of computational resources, and we were not able to properly train the model for that reason. Finally, the first set of experiments on multilingual training with M-BERT gave promising results, hence it will be further studied in future work.

The ablation studies showed the positive effects of the training (fine-tuning) strategies that we applied, both LM-finetuning on the target domain, as well as gradual unfreezing of the network layers (in groups) proved to be crucial in reaching state-of-the-art classification performance.

To compare the computational costs, we calculated inference times for each model on an Intel i7-8700K CPU @ 3.70GHz. DistilBERT provides the lowest run time at 12 ms/example. RoBERTa and BERT (which have an identical architecture) have very similar run times with 17.1 ms, and 17.3 ms/example, respectively. XLNet, the heaviest model, requires 77 ms/example.

For a fair comparison, we trained all transformer models with the same set of hyperparameters (such as learning rate and number of training epochs). With customized and hand-picked parameters for each training cycle we expect further improvements of scores, which will be studied in future work together with model ensemble approaches and text data augmentation.

VII. CONCLUSIONS

Natural Language Processing (In) this work we evaluate current transformer models for natural language processing in combination with training strategies like language model (LM) fine-tuning, slanted triangular learning rates and gradual unfreezing in the field of LMTC (large multi-label text classification) on legal text datasets with long-tail label distributions. The datasets contain around 20K documents (JRC-Aquis) and 57K documents (EUROLEX57K) and are labeled with EuroVoc descriptors from the 7K terms in the EuroVoc taxonomy. The use of an iterative stratification algorithm for dataset splitting (into training/validation/testing) allows

to create standardized splits on the two datasets to enable comparison and reproducibility in future experiments. In the experiments, we provide new state-of-the-art results on both datasets, with a micro-F1 of 0.661 for JRC-Acquis and 0.754 for EUROLEX57K, and even higher scores for new datasets with reduced label sets inferred from the EuroVoc hierarchy (*top terms, microthesauri, and domains*).

The main contributions are: (i) new state-of-the-art LMTC classification results on both datasets for a problem type that is still largely unexplored [3], (ii) a comparison and interpretation of the performance of the applied models: AWD-LSTM, BERT, RoBERTa, DistilBERT and XLNet, (iii) the creation and provision (on GitHub) of new standardized versions of the two legal text datasets created with an iterative stratification algorithm, (iv) deriving new datasets with reduced label sets via the semantic structure within EuroVoc, and (v) ablation studies that quantify the contributions of individual training strategies and hyperparameters such as gradual unfreezing, number of training epochs and LM fine-tuning in this complex LMTC setting.

There are multiple angles for *future work*, including potentially deriving higher performance by using hand-picked learning rates and other hyperparameters for each model individually, and further experiments on using models such as multilingual BERT to profit from the availability of parallel corpora. Moreover, experiments with new architectures such as Graph Neural Networks [45] and various data augmentation techniques are candidates to improve classification performance.

ACKNOWLEDGEMENTS

This work was supported by the Government of the Russian Federation (Grant 074-U01) through the ITMO Fellowship and Professorship Program.

REFERENCES

- [1] F. Sebastiani, "Machine learning in automated text categorization," *ACM computing surveys (CSUR)*, vol. 34, no. 1, 2002, pp. 1–47.
- [2] J. Mullenbach, S. Wiegrefe, J. Duke, J. Sun, and J. Eisenstein, "Explainable prediction of medical codes from clinical text," *arXiv preprint arXiv:1802.05695*, 2018.
- [3] A. Rios and R. Kavuluru, "Few-shot and zero-shot multi-label learning for structured label spaces," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, vol. 2018. NIH Public Access, 2018, p. 3132.
- [4] P. Ioannis et al., "Lshtc: A benchmark for large-scale text classification," *arXiv preprint arXiv:1503.08581*, 2015.
- [5] E. Loza Mencía and J. Fürnkranz, *Efficient Multilabel Classification Algorithms for Large-Scale Problems in the Legal Domain*. Berlin, Heidelberg: Springer, 2010, pp. 192–215, retrieved: 09, 2020. [Online]. Available: https://doi.org/10.1007/978-3-642-12837-0_11
- [6] I. Chalkidis, E. Fergadiotis, P. Malakasiotis, and I. Androutsopoulos, "Large-scale multi-label text classification on EU legislation," in *Proc 57th Annual Meeting of the ACL*. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 6314–6322, retrieved: 09, 2020. [Online]. Available: <https://www.aclweb.org/anthology/P19-1636>
- [7] European Union Law Website. Retrieved: 09,2020. [Online]. Available: <https://eur-lex.europa.eu>
- [8] S. Ruder, M. E. Peters, S. Swayamdipta, and T. Wolf, "Transfer learning in natural language processing," in *2019 NAACL: Tutorials*, 2019, pp. 15–18.
- [9] The European Union's multilingual and multidisciplinary thesaurus. Retrieved: 09,2020. [Online]. Available: <https://eur-lex.europa.eu/browse/eurovoc.html>
- [10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. 2019 NAACL: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: ACL, Jun. 2019, pp. 4171–4186, retrieved: 09, 2020. [Online]. Available: <https://www.aclweb.org/anthology/N19-1423>
- [11] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.
- [12] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, 2019.
- [13] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, "Xlnet: Generalized autoregressive pretraining for language understanding," in *Advances in neural information processing systems*, 2019, pp. 5754–5764.
- [14] K. Sechidis, G. Tsoumakas, and I. Vlahavas, "On the stratification of multi-label data," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2011, pp. 145–158.
- [15] W.-C. Chang, H.-F. Yu, K. Zhong, Y. Yang, and I. Dhillon, "X-bert: extreme multi-label text classification using bidirectional encoder representations from transformers," *arXiv preprint arXiv:1905.02331*, 2019.
- [16] R. Steinberger, M. Ebrahim, and M. Turchi, "Jrc eurovoc indexer jex-a freely available multi-label categorisation tool," *arXiv preprint arXiv:1309.5223*, 2013.
- [17] G. Boella et al., "Linking legal open data: breaking the accessibility and language barrier in european legislation and case law," in *Proceedings of the 15th International Conference on Artificial Intelligence and Law*, 2015, pp. 171–175.
- [18] J. Liu, W.-C. Chang, Y. Wu, and Y. Yang, "Deep learning for extreme multi-label text classification," in *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2017, pp. 115–124.
- [19] K. Bhatia, H. Jain, P. Kar, M. Varma, and P. Jain, "Sparse local embeddings for extreme multi-label classification," in *Advances in neural information processing systems*, 2015, pp. 730–738.
- [20] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1532–1543. [Online]. Available: <http://www.aclweb.org/anthology/D14-1162>
- [21] M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," in *Proc. 2018 NAACL: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 2227–2237, retrieved: 09, 2020. [Online]. Available: <https://www.aclweb.org/anthology/N18-1202>
- [22] R. You, S. Dai, Z. Zhang, H. Mamitsuka, and S. Zhu, "Attentionxml: Extreme multi-label text classification with multi-label attention based recurrent neural networks," *arXiv preprint arXiv:1811.01727*, 2018.
- [23] SKOS Simple Knowledge Organization System. Retrieved: 09,2020. [Online]. Available: <https://www.w3.org/2004/02/skos/>
- [24] Resource Description Framework. Retrieved: 09,2020. [Online]. Available: <https://eur-lex.europa.eu/browse/eurovoc.html>
- [25] RDF 1.1 Turtle. Retrieved: 09,2020. [Online]. Available: <https://www.w3.org/TR/turtle>
- [26] E. Filtz, S. Kirrane, A. Polleres, and G. Wohlgenannt, "Exploiting eurovoc's hierarchical structure for classifying legal documents," in *OTM Confederated International Conferences' On the Move to Meaningful Internet Systems'*. Springer, 2019, pp. 164–181.
- [27] JRC-Acquis. Retrieved: 09,2020. [Online]. Available: <https://ec.europa.eu/jrc/en/language-technologies/jrc-acquis>
- [28] EURLEX57K dataset. Retrieved: 09,2020. [Online]. Available: http://nlp.cs.aueb.gr/software_and_datasets/EURLEX57K/
- [29] J. Howard and S. Ruder, "Universal language model fine-tuning for text classification," *arXiv preprint arXiv:1801.06146*, 2018.

- [30] Fastai documentation. Retrieved: 09,2020. [Online]. Available: <https://docs.fast.ai/>
- [31] Huggingface transformers. Retrieved: 09,2020. [Online]. Available: <https://huggingface.co/transformers>
- [32] Legal Documents, Large Multi-Label Text Classification. Retrieved: 09,2020. [Online]. Available: <https://github.com/zeinsh/Legal-Docs-Large-MLTC>
- [33] S. Merity, N. S. Keskar, and R. Socher, "Regularizing and optimizing lstm language models," arXiv preprint arXiv:1708.02182, 2017.
- [34] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in Advances in neural information processing systems, 2017, pp. 5998–6008.
- [35] Z. Dai, Z. Yang, Y. Yang, J. Carbonell, Q. V. Le, and R. Salakhutdinov, "Transformer-xl: Attentive language models beyond a fixed-length context," arXiv preprint arXiv:1901.02860, 2019.
- [36] H. Azarbonyad and M. Marx, "How many labels? determining the number of labels in multi-label text classification," in International Conference of the Cross-Language Evaluation Forum for European Languages. Springer, 2019, pp. 156–163.
- [37] Multi-label data stratification. Retrieved: 09,2020. [Online]. Available: <http://scikit.ml/stratification.html#Multi-label-data-stratification>
- [38] BERT, Multi-Lingual Model. Retrieved: 09,2020. [Online]. Available: <https://github.com/google-research/bert/blob/master/multilingual.md>
- [39] Huggingface BERT base uncased model. Retrieved: 09,2020. [Online]. Available: <https://huggingface.co/bert-base-uncased>
- [40] Huggingface RoBERTa base model. Retrieved: 09,2020. [Online]. Available: <https://huggingface.co/roberta-base>
- [41] Huggingface DistilBERT cased model. Retrieved: 09,2020. [Online]. Available: <https://huggingface.co/distilbert-base-uncased>
- [42] Huggingface XLNET cased model. Retrieved: 09,2020. [Online]. Available: <https://huggingface.co/xlnet-base-cased>
- [43] H. Jain, Y. Prabhu, and M. Varma, "Extreme multi-label loss functions for recommendation, tagging, ranking & other missing label applications," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 935–944.
- [44] A. Esuli, A. Moreo, and F. Sebastiani, "Funnelling: A new ensemble method for heterogeneous transfer learning and its application to cross-lingual text classification," ACM Transactions on Information Systems (TOIS), vol. 37, no. 3, 2019, pp. 1–30.
- [45] A. Pal, M. Selvakumar, and M. Sankarasubbu, "Magnet: Multi-label text classification using attention-based graph neural network," in Proc. 12th Int. Conf. on Agents and Artificial Intelligence - Volume 2: ICAART, INSTICC. SciTePress, 2020, pp. 494–505.