KEVLAR: the Complete Resource for EuroVoc Classification of Legal Documents

Lorenzo Bocchi[1,†], Camilla Casula[2,†] and Alessio Palmero Aprosio[1,*,†]

[1]University of Trento, Italy
[2]Fondazione Bruno Kessler, Trento, Italy

Abstract

The use of Machine Learning and Artificial Intelligence in the Public Administration (PA) has increased in the last years. In particular, recent guidelines proposed by various governments for the classification of documents released by the PA suggest to use the EuroVoc thesaurus. In this paper, we present KEVLAR, an all-in-one solution for performing the above-mentioned task on acts belonging to the Public Administration. First, we create a collection of 8 million documents in 24 languages, tagged with EuroVoc labels, taken from EUR-Lex, the web portal of the European Union legislation. Then, we train different pre-trained BERT-based models, comparing the performance of base models with domain-specific and multilingual ones. We release the corpus, the best-performing models, and a Docker image containing the source code of the trainer, the REST API, and the web interface. This image can be employed out-of-the-box for document classification.

Keywords

EuroVoc taxonomy, multilingual text classification, BERT, web interface

1. Introduction

EuroVoc is a multilingual and multidisciplinary thesaurus that has seen a significant rise in its use and importance in recent years. In particular, the taxonomy used in this thesaurus has become crucial for a number of activities of European Public Administrations, shaping the way information is organized, disseminated, and accessed. Containing over 7,000 concepts, EuroVoc acts as a reliable and efficient indexing system for a vast range of documents, legislative texts, and reports. Due to this, a growing number of governmental institutions around Europe has begun to use it internally for document categorization.

The Spanish government, for instance, has suggested the adoption of EuroVoc since 2014 [1], and has more recently started using it regularly in its official open data portal,[1] and in the *Portal de la Administración Electrónica* website.[2] Similarly, German and French public administrations are following the same strategy, in the `DCAT-AP.de`[3] and `data.gouv.fr`[4] portals respectively. Furthermore, Rovera et al. [2] presented a preliminary study that explores the migration of the *Gazzetta Ufficiale*, the official journal of records of the Italian government, towards the adoption of the EuroVoc taxonomy. Similar initiatives have also grown in other European countries [3, 4].

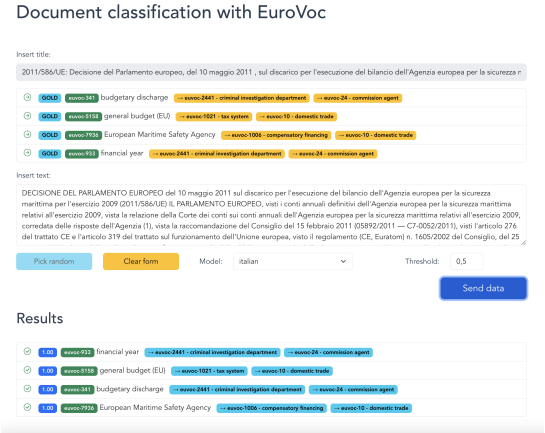

Figure 1: Screenshot of the web interface.

In this paper we present KEVLAR, Kessler EuroVoc Laws and Acts Repository, which aims at fulfilling a number of purposes.

1. First, we release a collection of more than 8 million documents from EUR-Lex, the European Union's official web portal, which gives comprehensive access to EU legal documents, spanning more than 70 years of EU legislation (1948-2022), and covering 24 languages. Over half of these

CLiC-it 2024: Tenth Italian Conference on Computational Linguistics, Dec 04 — 06, 2024, Pisa, Italy
*Corresponding author.
†These authors contributed equally.
lorenzo.bocchi@unitn.it (L. Bocchi); ccasula@fbk.eu (C. Casula); a.palmeroaprosio@unitn.it (A. Palmero Aprosio)
0000-0003-3360-5586 (C. Casula); 0000-0002-1484-0882 (A. Palmero Aprosio)

[1]https://bit.ly/eurovoc-es
[2]https://bit.ly/eurovoc-es-ae
[3]https://bit.ly/eurovoc-de
[4]https://www.data.gouv.fr/fr/

texts are already tagged with the corresponding EuroVoc concepts.
2. Secondly, we perform a series of experiments for automatic tagging of the documents using the EuroVoc taxonomy, comparing different approaches and language models.
3. Finally, we develop a web interface (see Figure 1) and a REST API that anyone (citizen or public administration) could use both to easily try automatic classification of documents and to integrate such categorization in any systems that might need it.

The models used for the web demo and the release are the best-performing ones we found, as described in Section 5. All the data and tools (the set of documents labeled with EuroVoc labels, the models, and the demo code) are freely available for download.

2. Related work

Several investigations have delved into the categorization of European legislation using EuroVoc labels. Notably, the task can be regarded as Extreme Multilabel Classification, as recognized in Liu et al. [5].

The JRC EuroVoc Indexer, detailed in Steinberger et al. [6], stands as a tool facilitating document categorization through EuroVoc classifiers across 22 languages. However, the dataset used for this tool [7] is limited to documents up to 2006. Their method entails the creation of lemma frequencies and associated weights, linked to specific descriptors referred to as *associates* or *topic signatures* in the research. When classifying a new document, the algorithm selects descriptors from the *topic signatures* exhibiting the highest resemblance to the lemma frequency list of the new document.

Later, You et al. [8] explored the application of Recurrent Neural Networks (RNNs) to extreme multi-label classification datasets, encompassing RCV1 [9], Amazon-13K [10], Wiki-30K, Wiki-500K [11], and an older EUR-Lex dataset from 2007 [12]. Attention-based RNNs proved to be particularly effective, outperforming other methods in 4 out of 5 datasets.

Chalkidis et al. [13] explored diverse deep learning architectures for this task. Among these, a fine-tuned `BERT-base` model [14] showed the highest performance, achieving a micro-averaged F1 score of 0.732 (considering all labels). Furthermore, they released a dataset consisting of 57,000 tagged documents from EUR-Lex.[5]

One of the most complete contributions to document classification using EuroVoc is PyEuroVoc, outlined in Avram et al. [15]. This study employs various pre-trained BERT models in 22 different languages, which were fine-tuned for the task. The source code in Python is publicly released, but cannot be used out-of-the-box and a known bug[6] may have led to unreliable results.

Some similar recent works on multi-language classification are described in Chalkidis et al. [16], Shaheen et al. [17], and Wang et al. [18]. Outside of the EuroVoc ecosystem, two large-sized legal datasets were released by Niklaus et al. [19, 20] for language model creation.

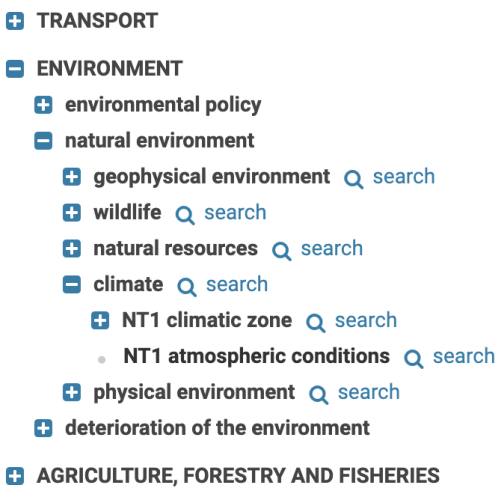

Figure 2: Example of EuroVoc taxonomy.

3. Dataset description

3.1. EUR-Lex

The reference for European legislation is EUR-Lex[7], a web portal that grants users comprehensive access to EU legal documents. It is available in all of the European Union's 24 official languages and is updated daily by its Publications Office. Most of the documents present in EUR-Lex are manually categorized using EuroVoc concepts.

3.2. EuroVoc

EuroVoc's hierarchical structure is organized into three different layers: Thesaurus Concept (TC), Micro Thesaurus (MT, previously referred to as "sub-sector" level), and Domain (DO, previously referred to as "main sector" level). The TC level is the base level, where all the key concepts are found. The documents on EUR-Lex are tagged with labels from this level. Every TC is assigned

[5] https://bit.ly/eurlex57k

[6] https://bit.ly/pyeurovoc-bug

[7] https://eur-lex.europa.eu/

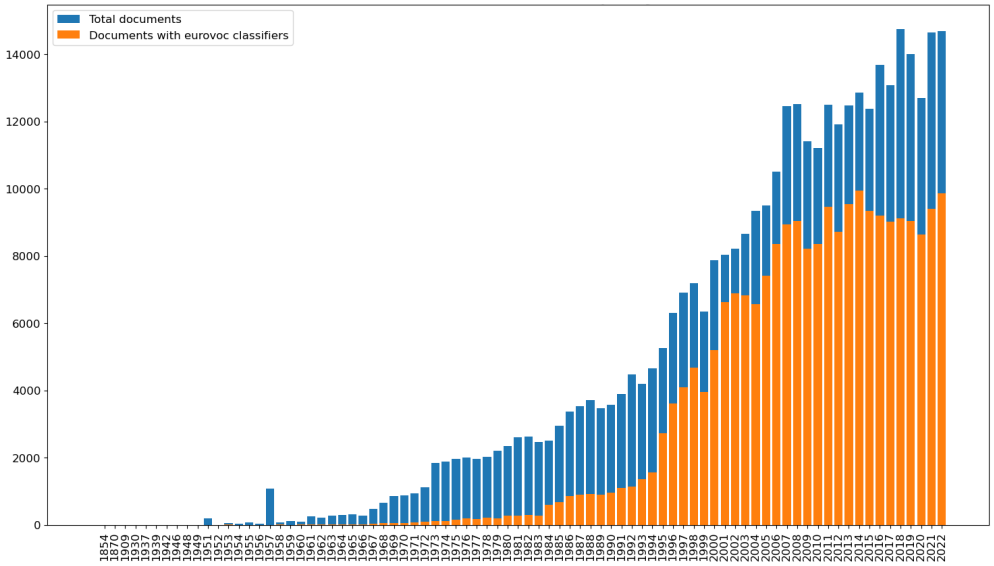

Figure 3: Number of documents per year (with the percentage already tagged with EuroVoc labels highlighted).

to an MT, which in turn is part of a specific DO. For example, the label "Confidentiality"[8] is assigned to the MT "Information and information processing", which belongs to the DO concept "Education and communication". Figure 2 shows a small subset of the EuroVoc taxonomy.

The experiments of this work have been launched on version 4.17 of EuroVoc. It contains 7,382 TCs, 127 MTs, and 21 DOs.

3.3. Dataset collection

KEVLAR was collected by downloading the documents from EUR-Lex. We built a set of tools written in Python that can be customized to obtain different subsets of the data (year, language, etc.).

In total, 8,368,328 documents were collected in 24 languages, 5,158,438 of which are annotated with EuroVoc descriptors, for a total of 32,021,783 tags. On average, 6.2 tags are associated with each document.

After filtering out these documents,[9] around 1.1 million texts with EuroVoc labels are collected.

Figure 3 shows the number of documents per year in English. The blue bars show the total number of documents retrieved for the year, while the orange bars show the number of documents that were labelled and have full text. The reduction is quite significant, especially before the year 2000.

[8] http://eurovoc.europa.eu/92

[9] Laws without any EuroVoc concept associated are not useful for our study. Regarding documents available in PDF format only, one could extract the text from them using OCR: this could be done in future work.

4. Experiments

In this section we provide a detailed account of the experiments conducted on document classification with respect to the EuroVoc taxonomy.

4.1. Deprecated labels and labels frequency

The EuroVoc thesaurus was initially developed in the 1980s and has constantly been updated and revised. Some labels started being used much earlier than others, and some are even deprecated for modern use but are still present in older documents.[10] This means that certain topics could stop being used in the future, potentially resulting in concept being replaced or merged with other existing concepts in future releases of EuroVoc.

Figure 4 shows the total occurrences of deprecated labels on a yearly basis. The result shows that from 2010 the usage of these labels decreased dramatically compared to the previous decade.

In addition to this, in EuroVoc labels assignment there is a strong imbalance in the data. For example, the most frequent label in the Italian documents, "economic concentration" with ID 69, is used more than 13,000 times, while the least frequent ones were assigned to just one document.

[10] https://bit.ly/eurovoc-handbook

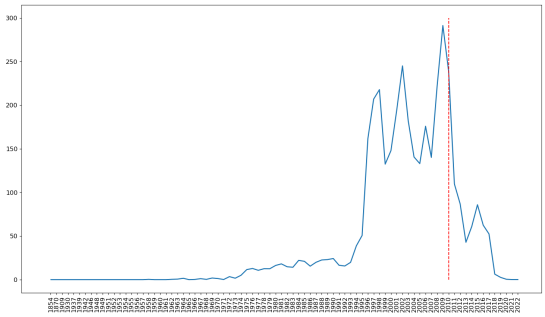

Figure 4: Total occurrence of deprecated labels per year. Marked in red is the year 2010.

4.2. Data filtering

Given the properties of the dataset described above, especially with regards to class imbalance, some filtering was carried out before proceeding with the experiments. First of all, all labels that have less than 10 samples assigned to them were filtered out. This number was kept low in order not to remove too much data and to preserve as many labels as possible. The threshold of 10 samples per label is a common reference, as stated in Chalkidis et al. [13].

Secondly, we filtered examples based on timespan. The percentage of documents with EuroVoc labels (as compared to the number of documents without them) became consistent starting from 2004 (see Figure 3), while a number deprecated labels are still present in documents, especially prior to 2010 (see Section 4.1). In order to obtain a more balanced dataset, in our experiments we consider only documents published in the interval 2010-2022, consisting of 471,801 documents. On average, each law is labelled with around 6 EuroVoc concepts.

After removing all the labels appearing in less than 10 documents, we removed documents that had 0 labels associated with them. This resulted in only 3 documents for each language being discarded. Conversely, more than 2000 labels out of 6079 were removed using this filter. It is interesting to note that even by using such a small threshold relative to the number of documents, around a third of the labels were discarded, meaning that 1/3 of the labels are barely used by the annotators of EU legislation.

4.3. Data Splits

To keep our experiments consistent with previous similar approaches (e.g. Avram et al. [15]), we split the data into train, dev, and test sets with an approximate ratio of 80/10/10, respectively.

In order to make the training reproducible and to avoid a single random extraction that could be too (un)lucky, we repeat the split using three different seeds and a pseudo-random number generator.

Each partition into train/dev/test is done using Iterative Stratification [27, 28], in order to preserve the concept balance.

Unless differently specified, all the results in the rest of the paper refer to the average of the values obtained by our experiments on the three seeds.

4.4. Training

We carry out our experiments using Transformer-based pre-trained language models. In particular, we use both BERT-based [14] and RoBERTa-based [29] models.

These families of language models have an intrinsic limit regarding the maximum number of words present in a text (usually 512), therefore each record of our data is created by concatenating the title and the text and then truncating at 512 tokens. While this might appear to entail a loss of information, Chalkidis et al. [30] have shown that the utilization of sparse-attention mechanisms, as exemplified by models like Longformer [31] and BigBird [32], to extend Transformer-based models for accommodating longer sequences, does not result in performance improvements in EuroVoc document classification.

Chalkidis et al. [33] found that classification tasks over the legal domain obtain better performance when pre-trained on domain-specific corpora. For our experiments, we focus on five major European languages, for which legal language models are available: English, Spanish, French, Italian, and German. For each of them, we test our dataset using: (i) the best-known base model; (ii) a monolingual legal model; (iii) the multilingual legal model proposed by Niklaus et al. [19].[11]. Table 1 lists the models for each language.

4.5. Hyperparameter Choice

After some preliminary experiments in which we experimented with the learning rate suggested in Avram et al. [15], 6e-5, we settled for a learning rate value of 3e-5, which led to better Macro-F1 results in our preliminary trials. Similarly, we increased the number of epochs from 30 to 100, as we noticed that the F1 score began to plateau at around 80 epochs. In each run, we saved the model with the best validation performance out of all the epochs, which typically fell within the last 10 epochs (although the difference between 80 and 100 epochs is relatively minor).

[11] `joelniklaus/legal-swiss-roberta-large`

	Base model	Legal model
en	bert-base-uncased	nlpaueb/legal-bert-base-uncased
fr	flaubert/flaubert_base_uncased	joelniklaus/legal-french-roberta-base
it	dbmdz/bert-base-italian-cased	dlicari/Italian-Legal-BERT
es	dccuchile/bert-base-spanish-wwm-cased	joelniklaus/legal-spanish-roberta-base
de	bert-base-german-cased	joelniklaus/legal-german-roberta-base

Table 1
Models used for the benchmark languages. Base: [en] Devlin et al. [14], [fr] Le et al. [21], [it] Schweter [22], [es] Cañete et al. [23], [de] Chan et al. [24]. Legal: [it] Licari and Comandè [25], [en] Chalkidis et al. [26], [fr, es, de] Niklaus et al. [20].

5. Discussion

Table 2 shows the classification results in terms of average macro F1 on the test sets of the three splits (see Section 4.3). Columns TC, MT, and DO show the result in terms of Thesaurus Concept (TC), Micro Thesaurus (MT), and Domain (DO), as described in Section 3.2.

In general, the classifiers achieving the best performances are trained on language models based on legal data. With the exception of French, for which the FlauBERT general model yields comparable results to the top legal model, the multilingual model introduced in the work by Niklaus et al. [19] outperforms all other models in the remaining benchmark languages.

Apart from French MT and DO, all the differences between the multilanguage model and the other ones are statistically significant (with a one-tailed t-test at 0.05).

The bottom part of Table 2 reports the performance of the multilingual model on the remaining languages.

6. Release and demo

All the data[12] and models[13] described in this paper are available for download under the CC-BY 4.0.

In addition to the documents, we also release on GitHub the code used to train and evaluate the models.[14]

Given that one of the main objectives of our research is to offer a comprehensive solution for aiding public administrations in document classification, we have also shared the source code for a REST API and a demonstration interface system (see Figure 1), alongside a Docker image for effortless deployment.

While the training phase requires GPUs for optimal performance, the models discussed in this article – accessible through package installation via Docker – can be utilized efficiently with CPU processing. Upon tool installation, users have the flexibility to select the desired languages, allowing only necessary models to be downloaded and loaded into memory.

A running instance of the API and the web demo is available for testing purposes.[15]

7. Conclusions and Future Work

In this paper, we release KEVLAR, an all-in-one solution for performing the document classification task on acts belonging to the Public Administration. We collected more than 8 million documents in 24 languages, compared different BERT and RoBERTa-based models on the classification of documents with respect to the EuroVoc taxonomy, and built an out-of-the-box tool for easily applying the classification to any text.

In the future, we will continue the exploration of novel methods to address this task with potentially better performance, for example using better-performing models or exploiting generation-based solutions.

References

[1] F.-J. Martínez-Méndez, R. López-Carreño, J.-A. Pastor-Sánchez, Open data en las administraciones públicas españolas: categorías temáticas y apps, Profesional de la información 23 (2014) 415–423.

[2] M. Rovera, A. P. Aprosio, F. Greco, M. Lucchese, S. Tonelli, A. Antetomaso, Italian legislative text classification for Gazzetta Ufficiale, AI per la Pubblica Amministrazione, at Ital-IA (2023).

[3] T. D. Prekpalaj, The role of key words and the use of the multilingual eurovoc thesaurus when searching for legal regulations of the republic of croatia - research results, in: 2021 44th International Convention on Information, Communication and Electronic Technology (MIPRO), IEEE, 2021, pp. 1470–1475. doi:10.23919/MIPRO52101.2021.9597043.

[4] D. Caled, M. Won, B. Martins, M. J. Silva, A hierarchical label network for multi-label eurovoc classification of legislative contents, in: Digital Libraries for Open Knowledge: 23rd Interna-

[12] https://bit.ly/kevlar-2024
[13] https://dh.fbk.eu/software/kevlar-models
[14] https://github.com/dhfbk/kevlar
[15] https://dh-server.fbk.eu/kevlar-ui/

	TC	MT	DO
en (base)	0,455	0,714	0,800
en (legal)	0,484	0,729	0,812
en (legal-ml)	**0,544**	**0,769**	**0,842**
it (base)	0,450	0,709	0,798
it (legal)	0,330	0,619	0,736
it (legal-ml)	**0,487**	**0,735**	**0,818**
fr (base)	**0,529**	**0,750**	**0,827**
fr (legal)	0,461	0,719	0,808
fr (legal-ml)	0,495	0,737	0,822
de (base)	0,435	0,689	0,786
de (legal)	0,371	0,656	0,766
de (legal-ml)	**0,514**	**0,738**	**0,823**
es (base)	0,485	0,730	0,812
es (legal)	0,408	0,686	0,783
es (legal-ml)	**0,523**	**0,754**	**0,830**
nl (legal-ml)	0,400	0,669	0,774
cs (legal-ml)	0,406	0,675	0,778
da (legal-ml)	0,359	0,633	0,746
et (legal-ml)	0,413	0,677	0,775
fi (legal-ml)	0,412	0,672	0,772
pt (legal-ml)	0,385	0,662	0,769
hu (legal-ml)	0,438	0,695	0,792
lt (legal-ml)	0,302	0,608	0,732
sv (legal-ml)	0,429	0,684	0,783
bg (legal-ml)	0,399	0,669	0,771
el (legal-ml)	0,414	0,680	0,782
ga (legal-ml)	0,213	0,298	0,494
hr (legal-ml)	0,386	0,660	0,770
lv (legal-ml)	0,299	0,600	0,727
mt (legal-ml)	0,371	0,646	0,756
pl (legal-ml)	0,434	0,688	0,786
ro (legal-ml)	0,417	0,680	0,781
sk (legal-ml)	0,390	0,665	0,770
sl (legal-ml)	0,391	0,663	0,768

Table 2
Results (in terms of macro F1) for all languages.

tional Conference on Theory and Practice of Digital Libraries, TPDL 2019, Oslo, Norway, September 9-12, 2019, Proceedings, Springer-Verlag, Berlin, Heidelberg, 2019, p. 238–252. URL: https://doi.org/10.1007/978-3-030-30760-8_21. doi:10.1007/978-3-030-30760-8_21.

[5] J. Liu, W.-C. Chang, Y. Wu, 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, SIGIR '17, Association for Computing Machinery, New York, NY, USA, 2017, p. 115–124. URL: https://doi.org/10.1145/3077136.3080834. doi:10.1145/3077136.3080834.

[6] R. Steinberger, M. Ebrahim, M. Turchi, Jrc eurovoc indexer jex-a freely available multi-label categorisation tool, arXiv preprint arXiv:1309.5223 (2013).

[7] R. Steinberger, B. Pouliquen, A. Widiger, C. Ignat, T. Erjavec, D. Tufiş, D. Varga, The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages, in: Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06), European Language Resources Association (ELRA), Genoa, Italy, 2006. URL: http://www.lrec-conf.org/proceedings/lrec2006/pdf/340_pdf.pdf.

[8] R. You, Z. Zhang, Z. Wang, S. Dai, H. Mamitsuka, S. Zhu, Attentionxml: Label tree-based attention-aware deep model for high-performance extreme multi-label text classification, Advances in Neural Information Processing Systems 32 (2019).

[9] D. D. Lewis, Y. Yang, T. G. Rose, F. Li, Rcv1: A new benchmark collection for text categorization research, J. Mach. Learn. Res. 5 (2004) 361–397.

[10] J. McAuley, J. Leskovec, Hidden factors and hidden topics: Understanding rating dimensions with review text, in: Proceedings of the 7th ACM Conference on Recommender Systems, RecSys '13, Association for Computing Machinery, New York, NY, USA, 2013, p. 165–172. URL: https://doi.org/10.1145/2507157.2507163. doi:10.1145/2507157.2507163.

[11] A. Zubiaga, Enhancing navigation on wikipedia with social tags, arXiv preprint arXiv:1202.5469 (2012).

[12] E. Loza Mencía, J. Fürnkranz, Efficient multilabel classification algorithms for large-scale problems in the legal domain, 2010. URL: http://dx.doi.org/10.1007/978-3-642-12837-0_11. doi:10.1007/978-3-642-12837-0_11.

[13] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, I. Androutsopoulos, Large-scale multi-label text classification on eu legislation, arXiv preprint arXiv:1906.02192 (2019).

[14] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: J. Burstein, C. Doran, T. Solorio (Eds.), 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), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423. doi:10.18653/v1/N19-1423.

[15] A. Avram, V. F. Pais, D. Tufis, Pyeurovoc: A tool for multilingual legal document classification with eurovoc descriptors, CoRR abs/2108.01139 (2021). URL: https://arxiv.org/abs/2108.01139. arXiv:2108.01139.

[16] I. Chalkidis, M. Fergadiotis, I. Androutsopoulos, 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, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 6974–6996. URL: https://aclanthology.org/2021.emnlp-main.559. doi:10.18653/v1/2021.emnlp-main.559.

[17] Z. Shaheen, G. Wohlgenannt, E. Filtz, Large scale legal text classification using transformer models, 2020. arXiv:2010.12871.

[18] L. Wang, Y. W. Teh, M. A. Al-Garadi, Adopting the multi-answer questioning task with an auxiliary metric for extreme multi-label text classification utilizing the label hierarchy, 2023. arXiv:2303.01064.

[19] J. Niklaus, V. Matoshi, M. Stürmer, I. Chalkidis, D. E. Ho, Multilegalpile: A 689gb multilingual legal corpus, 2023. arXiv:2306.02069.

[20] J. Niklaus, V. Matoshi, P. Rani, A. Galassi, M. Stürmer, I. Chalkidis, Lextreme: A multi-lingual and multi-task benchmark for the legal domain, 2023. arXiv:2301.13126.

[21] H. Le, L. Vial, J. Frej, V. Segonne, M. Coavoux, B. Lecouteux, A. Allauzen, B. Crabbé, L. Besacier, D. Schwab, Flaubert: Unsupervised language model pre-training for french, in: Proceedings of The 12th Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2020, pp. 2479–2490. URL: https://www.aclweb.org/anthology/2020.lrec-1.302.

[22] S. Schweter, Italian bert and electra models, 2020. URL: https://doi.org/10.5281/zenodo.4263142. doi:10.5281/zenodo.4263142.

[23] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, J. Pérez, Spanish pre-trained bert model and evaluation data, in: PML4DC at ICLR 2020, 2020.

[24] B. Chan, S. Schweter, T. Möller, German's next language model, in: Proceedings of the 28th International Conference on Computational Linguistics, International Committee on Computational Linguistics, Barcelona, Spain (Online), 2020, pp. 6788–6796. URL: https://aclanthology.org/2020.coling-main.598. doi:10.18653/v1/2020.coling-main.598.

[25] D. Licari, G. Comandè, ITALIAN-LEGAL-BERT: A Pre-trained Transformer Language Model for Italian Law, in: D. Symeonidou, R. Yu, D. Ceolin, M. Poveda-Villalón, D. Audrito, L. D. Caro, F. Grasso, R. Nai, E. Sulis, F. J. Ekaputra, O. Kutz, N. Troquard (Eds.), Companion Proceedings of the 23rd International Conference on Knowledge Engineering and Knowledge Management, volume 3256 of *CEUR Workshop Proceedings*, CEUR, Bozen-Bolzano, Italy, 2022. URL: https://ceur-ws.org/Vol-3256/#km4law3, iSSN: 1613-0073.

[26] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, I. Androutsopoulos, LEGAL-BERT: The muppets straight out of law school, in: Findings of the Association for Computational Linguistics: EMNLP 2020, Association for Computational Linguistics, Online, 2020, pp. 2898–2904. URL: https://aclanthology.org/2020.findings-emnlp.261. doi:10.18653/v1/2020.findings-emnlp.261.

[27] K. Sechidis, G. Tsoumakas, I. Vlahavas, On the stratification of multi-label data, Machine Learning and Knowledge Discovery in Databases (2011) 145–158.

[28] P. Szymański, T. Kajdanowicz, A network perspective on stratification of multi-label data, in: L. Torgo, B. Krawczyk, P. Branco, N. Moniz (Eds.), Proceedings of the First International Workshop on Learning with Imbalanced Domains: Theory and Applications, volume 74 of *Proceedings of Machine Learning Research*, PMLR, ECML-PKDD, Skopje, Macedonia, 2017, pp. 22–35.

[29] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, arXiv preprint arXiv:1907.11692 (2019).

[30] I. Chalkidis, A. Jana, D. Hartung, M. Bommarito, I. Androutsopoulos, D. Katz, N. Aletras, 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), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 4310–4330. URL: https://aclanthology.org/2022.acl-long.297. doi:10.18653/v1/2022.acl-long.297.

[31] I. Beltagy, M. E. Peters, A. Cohan, Longformer: The long-document transformer, arXiv:2004.05150 (2020).

[32] M. Zaheer, G. Guruganesh, K. A. Dubey, J. Ainslie, C. Alberti, S. Ontanon, P. Pham, A. Ravula, Q. Wang, L. Yang, et al., Big bird: Transformers for longer sequences, Advances in neural information processing systems 33 (2020) 17283–17297.

[33] I. Chalkidis, E. Fergadiotis, P. Malakasiotis, N. Aletras, I. Androutsopoulos, Extreme multi-label legal text classification: A case study in EU legislation, in: Proceedings of the Natural Legal Language Processing Workshop 2019, Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 78–87. URL: https://aclanthology.org/W19-2209. doi:10.18653/v1/W19-2209.