Extreme Classification of European Union Law Documents driven by Entity Embeddings

Extreme Classification of European Union Law Documents driven by Entity Embeddings

Irene Benedetto[1,2,*], Luca Cagliero[1] and Francesco Tarasconi[2]

[1]Politecnico di Torino, Dipartimento di Automatica e Informatica, Corso Duca degli Abruzzi 24, 10129 Torino, Italy
[2]MAIZE, Via San Quintino 31, 10121 Torino, Italy

Abstract

Extreme Multi-label Classification (XMC) is the task of labeling documents with one or more labels from a large set of classes. In the context of Legal Artificial Intelligence, XMC is relevant to the automatic categorization of documents as they commonly address several orthogonal categorization schemes. Since retrieving a sufficient number of training document examples per class is challenging, XMC models are expected to be particularly effective in zero-shot learning scenarios. Existing approaches rely on transformer-based classification models, which leverage the attention mechanism to attend to specific textual units. However, classical attention scores are not able to differentiate between domain-specific and generic textual units. In this paper, we propose to use a legal entity-aware approach to zero-shot XMC of European Union law documents. By integrating information about domain-specific legal entities we ease the detection of label-sensitive information and prevent XMC models from attending to irrelevant or wrong text spans. The results achieved on the law documents available in the EURLex benchmark show that our approach is superior to both previous transformer-based approaches and opensource Large Language Models.

Keywords

Legal Artificial Intelligence, Extreme Multi-label Classification, Language Models, Law Documents

1. Introduction

The task of eXtreme Multi-label Classification (XMC) aims at assigning to a given text one or more pertinent labels shortlisted from a very large set of classes. Since some of the target classes are likely to be underrepresented or even absent in the training data, classifiers used for XMC are expected to be particularly effective in zero-shot learning scenarios [1, 2].

Transformer-based architectures have shown to be particularly effective in tackling XMC [1] in various application domains such as e-commerce [3], medical diagnosis [4] and legal AI [5]. This paper focuses on solving the XMC task in a particular legal sub-domain, i.e., the automatic classification of law documents.

Legal documents such as laws have peculiar characteristics that make the classification task inherently complex. Firstly, the vocabulary used is very technical and rich of domain-specific expressions and entities [6]. Secondly, legal documents likely have a peculiar structure making content retrieval and ranking particularly challenging [7]. Lastly, the contained text is often verbose as usually contains a lot of preliminaries or repetitions [8].

Benchmark datasets for law classification such as EURLex [5, 9] contain acts and proposals of the European legislation. To support their retrieval and exploration law documents are often annotated by Publication Offices with a very large number of labels (e.g., 4,271 labels in EURLex), which encompass frequent labels as well as few- and zero-shot ones. Therefore, automating the process of law document classification requires the use of accurate XMC models.

In this paper we aim at overcoming the main limitations of existing transformer-based approaches to law document classification (e.g., [10]), which leverage the attention mechanism to attend to the most salient textual units. Since attention scores do not differentiate between legal and general-purpose textual units, the capabilities of transformers to correctly assign law document categories can be limited, particularly in zero-shot learning contexts. To overcome this issue, we propose to adopt an entity-aware attention mechanism based on the LUKE transformer [11], which exploits the semantic characteristics of the domain by the means of entity embeddings, to enhance zero-shot classification. The key idea is to mainly consider the textual dependencies with the tokens associated with entities as they are most likely to be discriminating in law document classification.

The experiments carried out on the EURLex benchmark dataset [9] confirm the effectiveness of entity embeddings in enhancing zero-shot XMC performance. Notably, the proposed approach not only performs better than existing transformer-based methods but also turns out to be more effective than an opensource Large Language Model with a larger number of parameters, i.e., Llama 2 7B [12].

The remainder of this work is organized as follows. Section 2 reviews the existing literature, Section 3 describes the methodology, Section 4 presents the main experimental results whereas Section 5 draws the conclusions of this work.

2. Related work

Legal document classification. The most common case of document classification in the legal domain is the automatic categorization of court cases, where the goal is to predict the law area of the given case. Existing related works mainly focused on employing machine learning and deep learning solutions [13, 14, 15, 16]. Parallel studies have delved into the automatic text classification of legislation to discern the law topic, with a particular emphasis on monolingual datasets [10, 17, 18, 19, 20, 21, 22]. A more limited body of work has explored multi-lingual datasets of legislations [9]. Specifically, the work presented in [21] investigates the semantic relationship between each document and labels. However, their performance on English documents is limited. Conversely, the transformer-based approaches pro-

Published in the Proceedings of the Workshops of the EDBT/ICDT 2024 Joint Conference (March 25-28, 2024), Paestum, Italy
*Corresponding author.
irene.benedetto@{polito.it,maize.io} (I. Benedetto); luca.cagliero@polito.it (L. Cagliero); francesco.tarasconi@maize.io (F. Tarasconi)
0000-0001-7086-7898 (I. Benedetto); 0000-0002-7185-5247 (L. Cagliero)

posed in [9, 10, 18] are, to the best of our knowledge, state-of-the-art on English-written law documents. Unlike [9, 10, 18], our work focuses on leveraging entity information in law classification. To the best of our knowledge, the idea to boost the performance of transformer-based approaches to law document classification using entity embeddings has not been addressed in literature so far.

Transformers in Legal Artificial Intelligence. Transformer-based models have demonstrated promising results in several areas of legal AI. Specifically, pre-trained language models have proved to be effective in tackling various downstream tasks [18, 23, 24]. Specifically, they encompass legal entity recognition [6], legal question answering [7], and legal document summarization [8].

Language Models have been designed and fine-tuned for the legal domain as well, mainly on Chinese documents. For example, LaWGPT [25] is pre-trained using a large-scale Chinese legal text database. Lawyer LLaMA [26] is a Chinese Legal Large Language Model (LLM) that undergoes training on a substantial legal dataset. This model is capable of offering legal advice, analyzing legal cases, and generating legal articles. ChatLaw [27] comprises a collection of open-source legal LLMs in Chinese, including models like ChatLaw-13B and ChatLaw-33B. These models are trained on a vast dataset encompassing legal news, forums, and judicial interpretations. Existing legal LLMs are suited to Chinese documents only and are not specifically designed to tackle the eXtreme Multi-label Classification task.

3. Methodology

In this section, we describe the proposed methodology for eXtreme Multi-label Classification (XMC) of law documents. Our purpose is to tackle XMC in a zero-shot setting, i..e., in the absence of ad hoc training examples. To address this issue, we propose to recognize and use entity embeddings in the document text. Specifically, we leverage the pre-trained LUKE model [11] for the classification task by replacing the original classification layer with one trained from scratch on the benchmark dataset. LUKE is a pre-trained contextualized representation of words and entities based on transformer architecture. It produces the contextualized representations of both words and entities thanks to the *entity-aware self-attention mechanism*, an extension of the self-attention mechanism when computing attention scores.

Given a sequence of input vectors $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_k$, where $\mathbf{x}_i \in \mathbb{R}^D$, the attention score e_{ij} is computed as follows:

$$e_{ij} = \begin{cases} \mathbf{K}\mathbf{x}_j^\top \mathbf{Q}\mathbf{x}_i, & \text{if both } \mathbf{x}_i \text{ and } \mathbf{x}_j \text{ are words} \\ \mathbf{K}\mathbf{x}_j^\top \mathbf{Q}_{w2e}\mathbf{x}_i, & \text{if } \mathbf{x}_i \text{ is word and } \mathbf{x}_j \text{ is entity} \\ \mathbf{K}\mathbf{x}_j^\top \mathbf{Q}_{e2w}\mathbf{x}_i, & \text{if } \mathbf{x}_i \text{ is entity and } \mathbf{x}_j \text{ is word} \\ \mathbf{K}\mathbf{x}_j^\top \mathbf{Q}_{e2e}\mathbf{x}_i, & \text{if both } \mathbf{x}_i \text{ and } \mathbf{x}_j \text{ are entities} \end{cases}$$

where $\mathbf{Q}_{w2e}$, $\mathbf{Q}_{e2w}$, $\mathbf{Q}_{e2e} \in \mathbb{R}^{L\times D}$ are query matrices, $\mathbf{K} \in \mathbb{R}^{L\times D}$ is key matrix.

4. Experiments

Dataset. In our experiments, we consider the English portion of EURLEX dataset [9], a multi-label legal document classification dataset. It consists of 65k European Union (EU) laws annotated with the EUROVOC taxonomy labels.

The EUROVOC taxonomy is a multilingual classification and thesaurus system used by the European Union. This tool is designed to organize and categorize concepts and terms used in official EU documents, facilitating research and access to information. Each european act in the EURLEX dataset is associated to one or more EUROVOC concept.

Similar to [9] we focused on third level labels. For training and test our models we follow the dataset split provided by the respective authors.

Competitors. We compare our methodology with:

- **Logistic Regression**: A baseline consisting of a Term Frequency-Inverse Document Frequency (TF-IDF) encoder, counting both local and global frequencies of occurrence of the input tokens, and a logistic regression model trained on top of the encoded text.
- **RoBERTa** [9]: builds on BERT [28] removing the next-sentence pre-training objective and training with much larger mini-batches and learning rates;
- **LLama 2 7B** [12]: a pre-trained Large Language Model with approximately 7 billion parameters that showcases remarkable performance in both few-shot and zero-shot scenarios. Analogously to [29], to compare with LLMs we treated the XMC task as a generative problem.

Experimental setting. We finetuned the base version of LUKE model (*studio-ousia/luke-base*), for 10 epochs. This model was trained with AdamW optimizer [30] with a weight decay of 0.01 and a learning rate of 1e-5. During training, we applied a 0.1 probability of dropout on classification layer.

For the sake of fairness, LLama 2 7B has been trained with Parameter-efficient fine-tuning (PEFT) [31], LoRA [32] that freezes pre-trained model weights and introduces trainable rank decomposition matrices into each layer of the models architecture.

We trained the 8-bit quantized version of this model for a maximum of 3 epochs, with a learning rate of 1.4e-5, LORA $\alpha = 16$ and $r = 64$.

Metrics. Here we describe the various metrics used to evaluate the performance of the models in our study.

- *R*@5 and *P*@5: precision and recall at k predictions where k is equal to 5 in our dataset. It corresponds to the mean number of labels in the training set.

$$\text{Precision@k} = \frac{\text{TP}_k}{\text{TP}_k + \text{FP}_k}$$

$$\text{Recall@k} = \frac{\text{TP}_k}{\text{TP}_k + \text{FN}_k}$$

- *mRP*: for each document, the metric ranks the labels selected by the model by decreasing confidence, computes Precision@k, where k is the document's number of gold labels, and then averages the results over documents.

Hardware. We conducted all the experiments on a single NVidia® Tesla® V100 GPU with 16 GB of memory, running on Ubuntu 22.04 LTS.

4.1. Results

Performance comparison with different training strategies. We conducted experiments with different training procedures in order to test the performance of the proposed methodology and to compare it with that of different architectures. To this end, we first freezed the 9 attention blocks and fine-tune the classification layer to test the goodness of the hidden representation of our model. Secondly, we perform an end-to-end evaluation of the proposed model to fully assess its potential.

Table 1
Models comparison

Models	mRP
Logistic Regression	0.21
State-of-the-art [9] (first 9 blocks frozen)	0.27
Our approach (first 9 blocks frozen)	**0.33**
State-of-the-art [9] (end-to-end training)	0.67
Our approach (end-to-end training)	**0.68**
LLama 2 7B [12]	0.65

Table 1 reports the overall performance of our model with different training strategies. Our results show that the proposed method performs better than both the state-of-the-art model and the Large Language Model Llama 2. Notably, the model attains superior performance compared to the state-of-the-art competitor even when the first 9 attention blocks are kept fixed. This suggests the efficacy of our model in generating highly informative hidden representations that enhance the classification task.

Zero-shot performance comparison. We conducted a comparative analysis of the performance of our model and competitors on zero-shot labels (i.e. labels not present in the training set). In this case, we trained all models without employing any freezing of model layers.

We report the results in Table 2 in terms of Precision@5 and Recall@5. Our evaluation focuses on evaluate the model's ability to retrieve all relevant results without any knowledge about labels. The number of predictions considered is always five, in compliance with [20].

Our results indicate that the baseline model performs poorly in this zero-shot learning context, with very low scores for both Precision@5 and Recall@5. The state-of-the-art model exhibits slightly higher scores, but still performs worse than the model proposed in this work. The proposed method achieves significantly higher Precision@5 and Recall@5 scores, indicating its superiority over the other two models in this zero-shot learning context. These results demonstrate the accuracy of our proposed model and the completeness of the model's predictions. Interestingly, LLMs demonstrate superior Recall@5 performance, even though their overall results are worse.

Comparison between models' attention. To further support the efficacy of the entity-aware self-attention mechanism for the given task, we examine the attention scores obtained by the best overall models according to the results in Table 1. For each class we compute the mean tokens attention score assigned by the state-of-the-art and LUKE

Table 2
Comparison in zero-shot learning context

	R@5	P@5
Logistic Regression	0.001	0.001
State-of-the-art [9]	0.028	0.006
Our approach	0.087	**0.164**
LLama 2 7B [12]	**0.253**	0.056

models, considering the last attention layer[1]. We sorted the results in decreasing order, ranking the tokens according to the attention given by the model. Then, separately for each class $c \in C$, the Mean Reciprocal Rank (MRR) of model m_i with the most frequent k tokens of class c was computed, i.e. :

$$\text{MRR}_{m_i,c,k} = \text{MRR}(\text{R}_{a(m_i),k_c}) \quad (1)$$

where $\text{R}_{a(m_i)}$ is the model m_i attention ranking position of k most frequent tokens of class $c \in C$.

We then compute the MRR difference between our model and the state-of-the-art model for different values of k:

$$\text{MRR}_k = \frac{1}{|C|} \sum_{c \in C} (\text{MRR}_{\text{LUKE},c,k} - \text{MRR}_{\text{SOTA},c,k}) \quad (2)$$

where

- $\text{MRR}_{\text{LUKE},c,k}$ is the Mean Reciprocal Rank computed with the LUKE model ranking, for class $c \in C$ considering the K most frequent term.
- $\text{MRR}_{\text{SOTA},c,k}$ is the Mean Reciprocal Rank computed with the state-of-the-art model ranking, for class $c \in C$ considering the K most frequent term.

These values are reported in Figures 1 and 2 which consider the frequent and zero-shot labels, respectively.

Scores above zero indicate that, on average, our model is giving more attention to the most frequent terms of the classes. These results reveal that our model is giving more attention to terms more frequently appear in each class, especially in correspondence of zero-shot labels, although differences decreases while k increases.

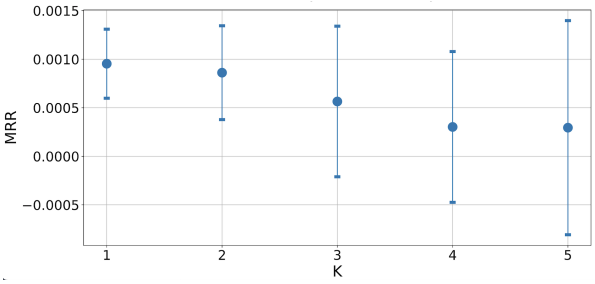

Figure 1: Comparison of Mean Reciprocal Rank (MRR) Differences in Token Attention Scores between our proposed model and the state-of-the-art model for various values of k computed considering **frequent labels**. Positive scores indicate that our model assigns higher attention to the most frequent terms of each class.

[1]We consider the last attention head because is the closest to the classification layer.

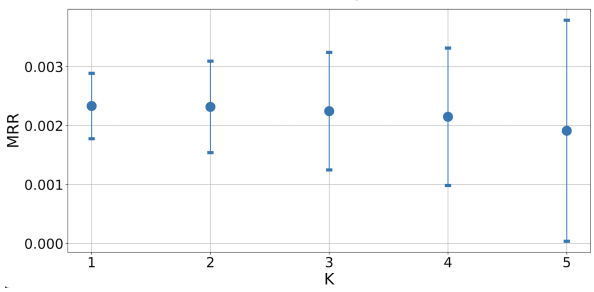

Figure 2: Comparison of Mean Reciprocal Rank (MRR) Differences in Token Attention Scores between our proposed model and the state-of-the-art model for various values of k computed considering only **zero-shot labels**. Positive scores indicate that our model assigns higher attention to the most frequent terms of each class.

5. Conclusion and future work

In this paper we explored the use of an entity-aware attention-based method to eXtreme Multi-label Classification of law documents. We show that attending to entity-related tokens enhances the capability of the transformer to attend to class-related pieces of text. The proposed method shows performance superior to both state-of-the-art transformers and Large Language Models, achieving higher precision and recall scores, especially in the most challenging zero-shot learning context. The experiments also highlight the impact of different training strategies and the effectiveness of the proposed model in generating informative hidden representations.

Based on the preliminary results, we envision the following future research directions:

- **Cross-lingual Transfer**: We plan to study the models' performance in the zero-shot cross-lingual transfer scenario for legal text classification in languages other than English.
- **LLMs Fine-tuning Strategies**: Another line of research will be the exploration of additional LLM fine-tuning strategies that incorporate hierarchical clustering [29].

Acknowledgments

The research leading to these results has been partially supported by the SmartData@PoliTO Center for Big Data Technologies. This study was partially carried out within the the MICS (Made in Italy – Circular and Sustainable) Extended Partnership and received funding from Next-GenerationEU (Italian PNRR – M4 C2, Invest 1.3 – D.D. 1551.11-10-2022, PE00000004) and within the FAIR - Future Artificial Intelligence Research - and received funding from the European Union Next-GenerationEU (PNRR MISSIONE 4 COMPONENTE 2, INVESTIMENTO 1.3 D.D. 1555 11/10/2022, PE00000013). This paper reflects only the authors' views and opinions, neither the European Union nor the European Commission can be considered responsible for them.

References

[1] H.-F. Yu, K. Zhong, I. S. Dhillon, W.-C. Wang, Y. Yang, X-bert: extreme multi-label text classification using bidirectional encoder representations from transformers, in: NeurIPS 2019 Workshop on Science Meets Engineering of Deep Learning, 2019.

[2] W. Chang, H. Yu, K. Zhong, Y. Yang, I. S. Dhillon, A modular deep learning approach for extreme multi-label text classification, CoRR abs/1905.02331 (2019). URL: http://arxiv.org/abs/1905.02331. arXiv:1905.02331.

[3] R. Agrawal, A. Gupta, Y. Prabhu, M. Varma, Multi-label learning with millions of labels: Recommending advertiser bid phrases for web pages, in: Proceedings of the 22nd International Conference on World Wide Web, WWW '13, Association for Computing Machinery, New York, NY, USA, 2013, p. 13–24. URL: https://doi.org/10.1145/2488388.2488391. doi:10.1145/2488388.2488391.

[4] A. Johnson, T. Pollard, L. Shen, L.-w. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Celi, R. Mark, Mimic-iii, a freely accessible critical care database, Scientific Data 3 (2016) 160035. doi:10.1038/sdata.2016.35.

[5] I. Chalkidis, E. Fergadiotis, P. Malakasiotis, I. Androutsopoulos, Large-scale multi-label text classification on EU legislation, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 6314–6322. URL: https://aclanthology.org/P19-1636. doi:10.18653/v1/P19-1636.

[6] I. Angelidis, I. Chalkidis, M. Koubarakis, Named entity recognition, linking and generation for greek legislation, in: JURIX, 2018.

[7] D. Hendrycks, C. Burns, A. Chen, S. Ball, CUAD: an expert-annotated NLP dataset for legal contract review, CoRR abs/2103.06268 (2021). URL: https://arxiv.org/abs/2103.06268. arXiv:2103.06268.

[8] D. Jain, M. D. Borah, A. Biswas, Summarization of legal documents: Where are we now and the way forward, Computer Science Review 40 (2021) 100388. URL: https://www.sciencedirect.com/science/article/pii/S1574013721000289. doi:https://doi.org/10.1016/j.cosrev.2021.100388.

[9] I. Chalkidis, M. Fergadiotis, I. Androutsopoulos, Multieurlex – a multi-lingual and multi-label legal document classification dataset for zero-shot cross-lingual transfer, 2021. URL: https://arxiv.org/abs/2109.00904. doi:10.48550/ARXIV.2109.00904.

[10] 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.

[11] I. Yamada, A. Asai, H. Shindo, H. Takeda, Y. Matsumoto, Luke: Deep contextualized entity representations with entity-aware self-attention, in: EMNLP, 2020.

[12] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. C. Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez,

M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M.-A. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, T. Scialom, Llama 2: Open foundation and fine-tuned chat models, 2023. arXiv:2307.09288.

[13] O. Sulea, M. Zampieri, S. Malmasi, M. Vela, L. P. Dinu, J. van Genabith, Exploring the use of text classification in the legal domain, CoRR abs/1710.09306 (2017). URL: http://arxiv.org/abs/1710.09306. arXiv:1710.09306.

[14] J. Gao, H. Ning, Z. Han, L. Kong, H. Qi, Legal text classification model based on text statistical features and deep semantic features, in: P. M. 0001, T. M. 0001, P. Majumder, M. Mitra (Eds.), Working Notes of FIRE 2020 - Forum for Information Retrieval Evaluation, Hyderabad, India, December 16-20, 2020, volume 2826 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2020, pp. 35–41. URL: http://ceur-ws.org/Vol-2826/T1-7.pdf.

[15] H. Chen, L. Wu, J. Chen, W. Lu, J. Ding, A comparative study of automated legal text classification using random forests and deep learning, Information Processing & Management 59 (2022) 102798. URL: https://www.sciencedirect.com/science/article/pii/S0306457321002764. doi:https://doi.org/10.1016/j.ipm.2021.102798.

[16] A. Aguiar, R. Silveira, V. Pinheiro, V. Furtado, J. A. Neto, Text classification in legal documents extracted from lawsuits in brazilian courts, in: A. Britto, K. Valdivia Delgado (Eds.), Intelligent Systems, Springer International Publishing, Cham, 2021, pp. 586–600.

[17] E. Loza Mencía, J. Fürnkranz, Efficient Multilabel Classification Algorithms for Large-Scale Problems in the Legal Domain, Springer-Verlag, Berlin, Heidelberg, 2010, p. 192–215.

[18] 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.

[19] C. Papaloukas, I. Chalkidis, K. Athinaios, D. Pantazi, M. Koubarakis, Multi-granular legal topic classification on greek legislation, CoRR abs/2109.15298 (2021). URL: https://arxiv.org/abs/2109.15298. arXiv:2109.15298.

[20] I. Chalkidis, M. Fergadiotis, I. Androutsopoulos, Multieurlex - A multi-lingual and multi-label legal document classification dataset for zero-shot cross-lingual transfer, CoRR abs/2109.00904 (2021). URL: https://arxiv.org/abs/2109.00904. arXiv:2109.00904.

[21] X. Huang, B. Chen, L. Xiao, L. Jing, Label-aware document representation via hybrid attention for extreme multi-label text classification, CoRR abs/1905.10070 (2019). URL: http://arxiv.org/abs/1905.10070. arXiv:1905.10070.

[22] W. Zhao, H. Peng, S. Eger, E. Cambria, M. Yang, Towards scalable and reliable capsule networks for challenging NLP applications, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 1549–1559. URL: https://aclanthology.org/P19-1150. doi:10.18653/v1/P19-1150.

[23] P. Henderson, M. S. Krass, L. Zheng, N. Guha, C. D. Manning, D. Jurafsky, D. E. Ho, Pile of law: Learning responsible data filtering from the law and a 256GB open-source legal dataset, in: Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track, 2022. URL: https://openreview.net/forum?id=3HCT3xfNm9r.

[24] S. Paul, A. Mandal, P. Goyal, S. Ghosh, Pre-training transformers on indian legal text, arXiv preprint arXiv:2209.06049 (2022). URL: https://arxiv.org/abs/2209.06049.

[25] H. Nguyen, A brief report on lawgpt 1.0: A virtual legal assistant based on gpt-3, arXiv preprint arXiv:2302.05729 (2023).

[26] Q. Huang, M. Tao, Z. An, C. Zhang, C. Jiang, Z. Chen, Z. Wu, Y. Feng, Lawyer llama technical report, arXiv preprint arXiv:2305.15062 (2023).

[27] J. Cui, Z. Li, Y. Yan, B. Chen, L. Yuan, Chatlaw: Open-source legal large language model with integrated external knowledge bases, arXiv preprint arXiv:2306.16092 (2023).

[28] J. Devlin, M. 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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), Association for Computational Linguistics, 2019, pp. 4171–4186. URL: https://doi.org/10.18653/v1/n19-1423. doi:10.18653/v1/n19-1423.

[29] T. Jung, J.-K. Kim, S. Lee, D. Kang, Cluster-guided label generation in extreme multi-label classification, in: EACL 2023, 2023.

[30] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, in: International Conference on Learning Representations, 2017.

[31] H. Liu, D. Tam, M. Muqeeth, J. Mohta, T. Huang, M. Bansal, C. Raffel, Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning, 2022. arXiv:2205.05638.

[32] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, W. Chen, Lora: Low-rank adaptation of large language models, CoRR abs/2106.09685 (2021). URL: https://arxiv.org/abs/2106.09685. arXiv:2106.09685.