On the use of Pretrained Language Models for Legal Italian Document Classification

Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 225 (2023) 2244–2253

Procedia Computer Science

www.elsevier.com/locate/procedia

27th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES 2023)

On the use of Pretrained Language Models for Legal Italian Document Classification

Irene Benedetto[a,b], Gianpiero Sportelli[b], Sara Bertoldo[b], Francesco Tarasconi[b], Luca Cagliero[b], Giuseppe Giacalone[c]

[a]*name.surname@polito.it, Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129 Torino TO, Italy*
[b]*name.surname@h-farm.com, H-Farm Innovation, Via San Quintino, 31, 10121 Torino, Italy*
[c]*Giuffrè Francis Lefebvre, Via Busto Arsizio, 40, 20151, Milano, Italy*

Abstract

Document classification is helpful for law professionals to improve content browsing and retrieval. Pretrained Language Models, such as BERT, have become established for legal document classification. However, legal content is quite diversified. For example, documents vary in length from very short maxims to relatively long judgements; certain document types are rich of domain-specific expressions and can be annotated with multiple labels from domain-specific taxonomies. This paper studies to what extent existing pretrained models are suited to the legal domain. Specifically, we examine a real business case focused on Italian legal document classification. On a proprietary dataset with thousands of diversified categories (e.g., legal judgements, maxims, and legal news) we explore the use of Pretrained Language Models adapted to handle various content types. We collect both quantitative and qualitative results, highlighting best and worst cases, anomalous categories, and limitations of currently available models.

Peer-review under responsibility of the scientific committee of the 27th International Conference on Knowledge Based and Intelligent Information and Engineering Systems

Keywords: Multi-Class Legal Document Classification; Pretrained Language Models; Natural Language Processing;

1. Introduction

Legal databases collect highly diversified document types such as laws, regulations, court rulings, and principles [5]. To retrieve and browse content in large databases law professionals such as lawyers and judges make use of human-generated annotations. However, with the rapid increase of the number of legal documents available in elec-

* Corresponding author. Tel.: +39-011-090-7084 ; fax: +39-011-090-7099.
* Corresponding author Irene Benedetto
E-mail address: irene.benedetto@{h-farm.com, polito.it}

1877-0509
Peer-review under responsibility of the scientific committee of the 27th International Conference on Knowledge Based and Intelligent Information and Engineering Systems
10.1016/j.procs.2023.10.215

tronic form manual content annotation has become extremely labor-intensive [33]. The constant progress in Natural Language Processing techniques have found confirmation in the legal domain [37]. In particular, Pretrained Language Models (PLMs) have been successfully applied to address tasks such as legal question answering [14], legal document summarization [17], and legal entity recognition [1]. Classification techniques automatically assign to a given document one label from a predefined set of classes. Classifiers are of great interest for legal applications because they relieve domain experts of annotation duties. In particular, classifiers based on PLMs (e.g., [4, 5]) have shown to be particularly efficient and effective thanks to use of attention mechanism [35]. PLMs used for classification need to face the inherent complexity of legal documents. Firstly, documents can be either relatively short (e.g., a maxim) or quite long (e.g., contracts or judgements). However, most PLMs are neither capable of handling very long pieces of text nor directly portable from one document type to another [13]. Secondly, documents are likely to be associated with multiple labels at the same time. Candidate labels have arbitrary semantic relationships among each other, often based on hierarchical models. Thirdly, examples of human-annotated data are typically unevenly distributed across the law areas and document types. Lastly, legal documents written in Romance languages such as Italian can be even more complex to handle because of the richness of the legal vocabulary that is used [33]. Therefore, training and fine-tuning PLMs for the legal domain can be challenging. This paper examines the generalization capability of established PLMs across different types of legal documents. It explores the use of BERT-based PLMs [11] in a real business case focused on Italian legal document classification. We carry out a wide range of experiments on a proprietary dataset with thousands of documents diversified in data source (e.g., legal judgements, maxims and legal news) and law area (e.g., public administration, family, civil and telematic process, tax, housing, corporate, criminal, labour and bankruptcy law). Documents are annotated with more than 20,000 distinct labels. Starting from a standard classification pipeline, we extend the capability of the business solution to address the multi-class problem and handle long documents based on hierarchical modeling and multi-label attention [32, 34]. We assess the quality of the PLM outcomes from both a quantitative and qualitative viewpoint. The results highlight promising performance on maxims, worse results on longer documents, and contrasting outcomes on documents belonging to the *Tax* law area. The paper is organized as follows. Section 2 compares the current study with the prior works. Section 3 describes the business case and the pipeline extension. Section 4 describes the data and summarizes the main results. Finally, Section 6 draws the conclusions and discusses the future research agenda.

2. Related work

Extreme Multi-label Classification (XMC) focuses on assigning the most pertinent labels to a given test document based on the multi-class model trained on data with highly imbalanced label distribution. The main challenge is to properly handle a very large number of labels, which is hard to achieve using traditional classification models. Pretrained Language Models (PLMs) have achieved very promising XMC performance, mainly due to the use of attention mechanism in the transformer architecture [35]. In most real-world applications trasformer-based models (e.g., [8, 30, 9]) have shown to outperform traditional machine learning approaches (e.g., [19, 34]). In the context of legal AI, XMC has been mostly addressed on a monolingual dataset (e.g., [23, 3, 6, 27, 5, 16, 36]), whereas only few works focus on legal data written in languages other than English [5]. To the best of our knowledge, multi-EURLEX [5] is the most recent multi-lingual legal dataset for XMC. It consists of 65k European Union laws, officially translated into 23 languages, including also the Italian language, and annotated with the EUROVOC taxonomy. The main differences between the latter and the present work are: (1) *The domain adaptation*, i.e., just the European regulation for multi-EURLEX vs. legal judgements, maxims, and legal news. (2) *The analysis of the legal area*: we also compare PLMs performance on different areas of law, studying also the effect of level of detail in the label taxonomy.

3. Business case

A Italian publisher of legal texts and related products needs to organize legal news, court judgment's contracts, and maxims. To support efficient retrieval and browsing of the database content, documents need to be annotated with labels stored in a proprietary taxonomy (i.e., a is-a hierarchy). However, prior approaches to Italian legal document classification [32] have been tested on a single law area. Therefore, the level of portability of PLMs towards different

law areas and legal document types is still unknown. In this paper we address the aforesaid limitations by exploring a wider range of document types and law areas. We adopt the following the best practices [12]: (1) *Data benchmarking*, to allow detailed comparisons on a sufficiently large number of cases, and (2) *Error analysis*, to educate the Business on the limits of Artificial Intelligence and understand where the human must intervene.

We address the following directions of extension of the BERT model [11]:

- **Type-specific PLMs**: To examine the impact of the document type on classification performance, we fine-tune a separate BERT multi-label classification model [11] per area of law.
- **Multi-class data**: We tailor the classification process to a multi-class setting by first enforcing a minimum confidence threshold to each class probability and then assigning to the test document all the classes exceeding the minimum confidence level. The model is trained to classify all labels at different hierarchical levels jointly. In such a way, we encourage the model to learn at least one first- and second-level labels. Although not as detailed as the third-level labels, they ensure the identification of the correct sub-area of law.
- **Long documents**: We adapt PLMs to handle documents longer than the standard token size (512). To this end, we adopt a hierarchical model [24] that first splits each piece of text into paragraphs and then generates an intermediate paragraph-level representation from the hidden representation of the Begin-of-Sequence token of each paragraph, i.e., the *paragraph attention vector*. All the paragraph attention vectors for a given document are then aggregated and passed to a classification layer. By leveraging the multi-label attention mechanism [34], the extended model can assign different weights to different parts of the input based on their relevance to the label being predicted, thus improving the accuracy of the model.

During validation and testing, we post-process the models' predictions by eliminating redundant intermediate-level labels and prioritizing the most granular predictions. For instance, if the model generates both the labels *Associations and foundations* and *Associations and foundations - Committees*, we consider only the latter, as provides more detailed information.

4. Methods

In this section, we overview the main dataset characteristics (see Section 4.1) and report the main experimental settings (see Section 4.2) and model performance, highlighting the main empirical findings (see Section 5).

4.1. Dataset

The proprietary dataset consists of Italian legal judgments, maxims and legal news covering 10 different law areas:

- *Public administration*: documents that pertain to matters of public interest and concern the organization and functions of the public administration, as well as its relationships with private individuals.
- *Lease and housing*: documents related to rulings that pertain to contracts between private individuals, which include a real estate asset;
- *Family*: documents that deal with the legal relationships between people who make up a family, as defined by law;
- *Civil Liability*: documents related to offenses that violate the provisions of the civil code;
- *Labour*: documents related to the rules governing relations between workers and employers;
- *Civil and Telematic process*: documents with reference to norms ruling mechanisms, jurisdiction and laws that guarantee fair justice within a judicial process, in the civil sphere;
- *Criminal*: documents concerning the imposition of penalties or sanctions according to the severity of the crime;
- *Corporate*: documents in which there are references to regulations concerning the establishment, governance, control, dissolution and liquidation of companies, corporate responsibility, property relations between shareholders, extraordinary corporate transactions, and management of the company crisis;
- *Tax*: documents including procedures related to taxation;
- *Bankruptcy*: documents regarding the regulation of the phenomenon of business crisis.

Fig. 1. Extract of the label taxonomy related to the *family* law area

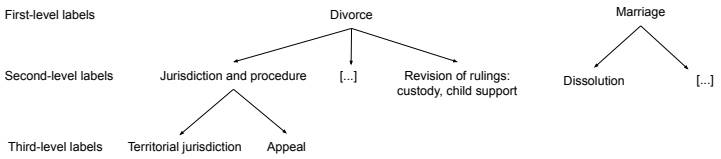
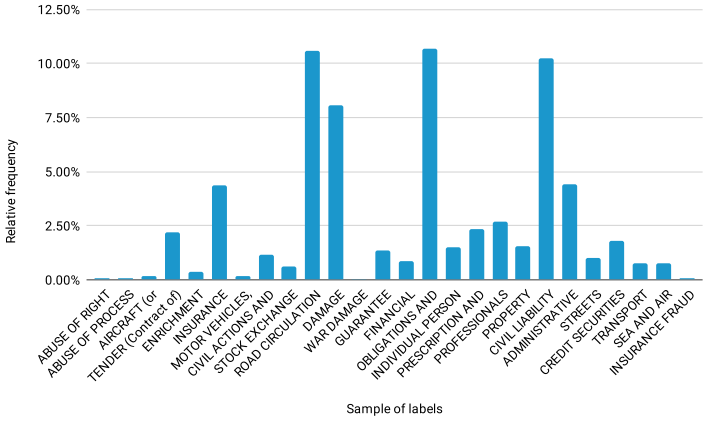


Fig. 2. Distribution of the most frequent first-level labels over the documents (legal judgments and maxims) related to civil liability law area.



Each law area corresponds to a portion of a proprietary taxonomy (i.e., a is-a hierarchy), that organizes the set of labels. Each label denotes a relevant legal principle. The taxonomy has a hierarchical structure. The labels hierarchies consist of three layers of concepts, each of them providing a more detailed characterization (see Figure 1). Notice that only part of the second-level labels are associated with third-level labels, whereas all the top-level labels are specialized by a second-level label. Part of the label hierarchies are in common across multiple areas of law. Legal judgments and maxims are, on average, annotated with five labels per document. Conversely, legal news are not annotated yet. The legal judgments are pre-processed as follows: we consider all paragraphs within the *FactLaw* corpus (the portion of the legal judgment those contained the fact description and the applied rulings), whereas preambles and conclusions are disregarded. Table 1 summarizes the dataset statistics, including the label distribution across different aggregation levels in the taxonomy. In addition, we report the mean document lengths, expressed as the average number of characters per document source. The document length in this dataset exceeds that of well-established text classification datasets [21, 15, 2, 25]. Notice that the number of labels per area of law is quite variable, yielding an imbalanced class label distribution (see Figure 2).

4.2. Experimental setup

We run a train-validation-test simulation separately for each area of law. Similar to [5] we keep the number of test documents per area fixed to 15,000 and the validation set proportion to 5%. We employed the *bert_multi_cased*[1] version of BERT, a multi-lingual extension of BERT which includes the Italian language in the pretraining and uses 12 transformer blocks, a hidden size of 768, and 12 attention heads. All the models were trained to maximize the cross entropy loss through the use of the sequence cross entropy loss and the AdamW optimizer [22] with a weight decay of 0.01. Models were trained for a small number of epochs (3), with a learning rate of $5e-5$. During inference, we consider a confidence threshold of 0.30. To avoid overfitting, we drop-out connections in the last classification layer with a drop-out probability of $p = 0.1$ and early stopping during training to further prevent overfitting.

[1] Available at the Tensorflow Hub: https://tfhub.dev/

Table 1. Statistics of the dataset: number of documents and labels and average document length across the law areas

	Dataset cardinality		
Law area	**Number of Legal Judgments**	**Number of Maxims**	**Number of Legal news**
Public administration	555749	280471	392
Lease and Housing	22192	35662	150
Family	34454	37800	502
Civil liability	111751	71842	470
Civil and telematic process	581480	421202	171
Criminal	377397	300252	431
Labour	279397	136639	244
Corporate	28703	9474	314
Tax	706343	459822	305
Bankruptcy	405553	290714	337
All	**3103019**	**2043878**	**3316**
	Labels cardinality		
Law area	**First level classes**	**Second level classes**	**Third level classes**
Public administration	51	832	1641
Lease and Housing	13	163	259
Family	35	230	243
Civil liability	26	400	470
Civil and telematic process	126	1650	2569
Criminal	147	1301	1698
Labour	28	512	784
Corporate	24	129	233
Tax	106	1631	2608
Bankruptcy	66	1028	1762
All	**622**	**7876**	**12267**
	Documents size		
Law area	**Legal judgements lengths**	**Maxims' lengths**	**Legal news lengths**
Public administration	15174.47	646.11	1100.28
Lease and Housing	15572.34	598.09	5967.18
Family	9862.95	640.55	9044.72
Civil liability	15174.47	646.11	12084.64
Civil and telematic process	11606.35	631.87	9258.03
Criminal	11850.67	606.92	3994.46
Labour	13730.64	632.02	1547.79
Corporate	18115.98	557.3	13981.53
Tax	12266.11	594.09	11646.29
Bankruptcy	12289.62	630.08	10149.7
All	**13564.36**	**618.31**	**7877.46**

Hardware and execution times. The experiments were conducted on a single NVidia® Tesla® A100 GPU with 85 GB of memory. The execution time for training and inference took from 1 hour to one day depending on the law area considered.

Evaluation metrics. We assess PLM performance from both quantitative and qualitative perspectives. Quantitative analyses require labeled data (i.e., the ground truth), available for legal judgments and maxims only. They quantify

the overlap between predicted and expected labels per documents in terms of precision and recall measures [31]. Specifically, precision measures the proportion of correctly classified samples over all samples classified as positive, recall measures the proportion of correctly classified positive samples over all positive samples. We also compute the weighted f1-score as the mean of the f1-score values (i.e., the harmonic mean of precision and recall) over all the input labels. The qualitative assessment aims at verifying the compliance of PLMs outcomes with the human expert's expectation. To double-check the quantitative outcomes, we run the qualitative evaluation on both a set of an unlabeled legal news dataset (neglected by the quantitative assessment procedure) and on the labeled legal judgments and maxims. In the validation process we involve a group of 10 expert evaluators asking them to annotate the document labels as *correct*, *partially correct*, or *incorrect*. For each area of law we count the percentage of legal documents belonging to any of the following categories:

- The number of documents where all the predicted labels are correct, and no labels are missing;
- The number of documents whose predictions are partially correct: in this case the evaluator considers the classification incomplete or identifies that part of the labels returned by the model are wrong;
- The number of documents where all the predictions are wrong: none of the labels returned by the model is acceptable.

5. Results

5.1. Comparison between data sources

Table 2 reports the weighted f1-score (as well as the corresponding confidence intervals) separately for each law area. The table columns indicate either the separate data sources (legal judgments or maxims) or the combination of the above (*all*). The f1-score values ranging from 0.48 to 0.76 indicate that the type of legal source relevantly affects PLM performance. The variability in the results is roughly comparable over all law areas. Specifically, maxims are less numerous but easier to classify. They consist of concise statements expressing general truths about the law, often used in legal reasoning and decision-making. Their inherent simplicity of the linguistic expressions likely helps the PLMs to achieve high-quality results. Conversely, the classification performance on judgments is on average worse across all the law areas. Judgments are detailed written decisions that include a summary of the facts of a specific case, legal arguments, and a ruling. They may include additional information that is not relevant to the classification task. Italian judgements may also include domain-specific terms that are underused in the other legal document types. The law areas with relatively higher standard deviations include *Family, Civil Liability, Criminal, Labour, Corporate*. This variability may suggest that the PLM's performance in these areas is more dependent on specific instances or variations within the legal texts. In contrast, law areas like *Public Administration* and *Lease and Housing* have relatively lower standard deviations, indicating more consistent performance across different instances or samples. This consistency may imply that the PLM exhibits a stable and reliable performance in these specific domains. The higher standard deviation observed for *Maxims* compared to Legal Judgments could be attributed to the fact that maxims may vary more widely in their wording, structure, and complexity compared to legal judgments, which are more structured and standardized. This variability in the language and content of Maxims could contribute to a higher standard deviation in performance scores.

5.2. Comparison between areas of law

PLM performance, expressed in terms of f1-score, does not exhibit a correlation with the number of labels in each area of law according to the Pearson correlation coefficient. Hence, generally speaking, considering the area of law while disregarding the document types is commonly not enough to achieve satisfactory classification results.

5.3. Comparing between levels of label granularity

Table 3 details the PLM performance separately for each set of labels belonging to a different level of granularity in the expert-provided taxonomy. PLMs exhibit a conservative tendency, resulting in a higher proportion of first- and

Table 2. PLM performance on different law areas in terms of weighted f1-score

Law area	Legal Judgments	Maxims	All
Public administration	0.53 (± 0.16)	0.65 (± 0.31)	0.61 (± 0.28)
Lease and Housing	0.59 (± 0.21)	0.68 (± 0.31)	0.64 (± 0.28)
Family	0.76 (± 0.25)	0.68 (± 0.32)	0.72 (± 0.32)
Civil liability	0.55 (± 0.20)	0.68 (± 0.36)	0.63 (± 0.33)
Civil and telematic process	0.53 (± 0.15)	0.67 (± 0.32)	0.61 (± 0.28)
Criminal	0.60 (± 0.20)	0.64 (± 0.32)	0.63 (± 0.32)
Labour	0.60 (± 0.18)	0.64 (± 0.30)	0.63 (± 0.28)
Corporate	0.48 (± 0.14)	0.65 (± 0.32)	0.62 (± 0.30)
Tax	0.51 (± 0.16)	0.62 (± 0.30)	0.59 (± 0.28)
Bankruptcy	0.56 (± 0.16)	0.63 (± 0.31)	0.61 (± 0.27)
All	**0.57 (± 0.19)**	**0.654 (± 0.31)**	

second-level labels assigned compared to the third-level label predictions. On average, the models generate two first- and second-level labels for every single third-level label. Level-1 predictions averagely achieve fairly high F1-score values (≥ 80%), with slightly better recall values compared to the precision. Level-2 and -3 classification outcomes show a severe performance drop (e.g., roughly 30% f1-score decrease from level-1 to level-3). The main reason is that levels 2 and 3 have a significantly higher number of labels thus making the problem of multi-class classification much more complex. In addition, as described in Section 3, the label trees are expanded and the model is trained to classify all the labels at the same time. Therefore labels of higher levels will be considered more frequently by the model during the training phase.

Table 3. PLM performance (Precision, Recall and F1-score) across different law area and taxonomy levels of granularity

Law area	Level 1 labels			Level 2 labels			Level 3 labels		
	P	R	F1	P	R	F1	P	R	F1
Public administration	0.82	0.84	0.83	0.60	0.57	0.56	0.48	0.39	0.40
Lease and Housing	0.93	0.92	0.92	0.65	0.52	0.57	0.51	0.33	0.38
Bankruptcy	0.77	0.84	0.80	0.57	0.57	0.55	0.46	0.43	0.42
Family	0.86	0.90	0.88	0.69	0.72	0.70	0.45	0.44	0.43
Labour	0.83	0.86	0.85	0.58	0.59	0.57	0.46	0.42	0.41
Civil and telematic process	0.76	0.83	0.79	0.57	0.57	0.55	0.46	0.43	0.42
Corporate	0.82	0.86	0.84	0.57	0.60	0.57	0.45	0.39	0.39
Tax	0.74	0.82	0.77	0.57	0.55	0.54	0.45	0.40	0.40
Civil liability	0.85	0.81	0.83	0.65	0.52	0.56	0.57	0.41	0.45
Criminal	0.78	0.82	0.80	0.57	0.60	0.57	0.49	0.46	0.45
All	**0.81**	**0.86**	**0.83**	**0.60**	**0.58**	**0.57**	**0.48**	**0.40**	**0.42**

5.4. Qualitatitive evaluation

Table 4 summarizes the results of the qualitative assessment. It reports the accuracy in terms of the percentage of correct, partially correct, and incorrect predictions. The percentage of agreement between annotators was around 70% (i.e. in 70% of cases, annotators agreed that a classification is completely correct, partially correct or completely wrong). On legal judgments and maxims, the outcomes are in line with the quantitative results (i.e., around 80% of correct assignments). Notably, around 10% of the other assignments are partially correct thus PLMs predictions can be deemed as quite reliable. Areas *Corporate* and *Tax* are the only exceptions, with a significant number of incorrect assignments probably due to the very peculiar terminology used in these document types. On legal news, PLMs get

worse performance, especially on the areas *Public administration* and *Criminal*. On the other hand, in 8 law areas out of 10, the number of wrong predictions remains stable, thus in most cases the automatic assignments are either correct or partially correct. Furthermore, on legal news, the number of predictions returned by the model decreases (3, on average). Annotators also reported that the most frequent cause of model errors is when there is a lack of a sufficiently detailed prediction. In conclusion, regardless of the domain PLMs correctly identify the sub-areas of law identified by the first-level labels, but, in order to achieve a higher level of detail, it needs some form of domain adaptation on unseen data sources.

Table 4. Summary of the results achieved in the qualitative validation.

	Legal judgement/maxims			**Legal news**		
Law area	**Correct**	**Partial correct**	**Incorrect**	**Correct**	**Partial correct**	**Incorrect**
Public administration	80%	14%	6%	49%	44%	7%
Lease and Housing	78%	12%	10%	77%	15%	8%
Family	81%	16%	3%	75%	20%	5%
Civil liability	83%	9%	8%	78%	13%	9%
Civil and Telematic process	79%	14%	7%	80%	18%	2%
Criminal	83%	12%	5%	59%	33%	8%
Labour	82%	13%	5%	77%	16%	7%
Corporate	79%	13%	8%	60%	12%	28%
Tax	82%	12%	6%	67%	18%	15%
Bankruptcy	81%	13%	6%	71%	24%	5%

5.5. Errors and models limitations

In this section we summarize the main models errors and limitations we identified:

- Performance Variation: The performance of the PLMs (Pre-trained Language Models) varies depending on the type of legal source. Maxims, which are concise and express general truths about the law, are easier to classify, resulting in higher performance scores on average with higher variability. On the other hand, judgments, which are detailed written decisions containing additional information, tend to have lower and more stable classification performance across all law areas.
- Law Area vs. Document Types: Simply considering the area of law without taking into account the document types is insufficient for achieving satisfactory classification results. The performance of PLMs in terms of f1-score does not exhibit a correlation with the number of labels in each area of law. Therefore, considering both the area of law and the document types is crucial for improving classification results.
- Label Granularity: PLMs tend to exhibit a conservative tendency in assigning labels, resulting in a higher proportion of first- and second-level label predictions compared to third-level labels. The F1-score decreases significantly from level-1 to level-3, indicating a drop in performance as the level of granularity increases. This is due to the higher complexity of multi-class classification with a larger number of labels at higher levels.
- On legal news, the performance decreases, especially in areas such as Public Administration and Criminal. The number of incorrect predictions remains stable across most law areas, indicating that PLMs are generally reliable but may require domain adaptation for unseen data sources.

6. Conclusions and future work

The paper studied the generalization capabilities of Pretrained Language Models for multi-label classification of Italian legal documents. It reported an extensive evaluation of 10 BERT-based models' performance on different law areas and document typologies. PLM performance was quantitatively and qualitatively evaluated using a proprietary dataset and taxonomy including over 5,000,000 documents and 20,000 labels. The results of the study indicate that

the type of legal information has a great impact on PLM performance, as shown by the performance variations across different types of law. They also indicate that maxims are easier to classify compared to legal judgments and laws. As expected, PLM performance decreases while considering labels at deeper levels of granularity in the taxonomy because the classification problem gets much more complex. As future work, we plan to address the following research lines:

- **Efficient PLMs**: further experimentation with compact language models, similar to DistillBERT [29] and ALBERT [20] but pretrained on legal documents. We seek to train more compact models that are capable of performing fairly good on legal data.
- **Modeling labels hierarchy**: leveraging learn the presence of inter-dependencies between the labels with multi-task approaches [28] or clustered guided approaches [18].
- **Domain adaptation**: we aim at improving classification performances on new data sources by reducing the drift in performance with more robust pre-training [10] tailored to the legal domain or using different data selection strategies [13].
- **Temporal concept drift**: similar to legal systems [26] taxonomies evolve over time according to a *temporal concept drift*, which is typical of legal topic classification [7]. Therefore, existing taxonomy-based document relationships may become unreliable. The updates of the original document collection and the presence of a relevant drift in the covered topics trigger the periodic retraining of the entire classification model. Such an activity can be labor-intensive and time-consuming. As a future research direction, we plan to explore different strategies of *continuous learning* applied to the legal domain.

References

[1] Iosif Angelidis, Ilias Chalkidis, and Manolis Koubarakis. 2018. Named Entity Recognition, Linking and Generation for Greek Legislation. In *JURIX*.

[2] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. DBpedia: A Nucleus for a Web of Open Data. In *The Semantic Web*, Karl Aberer, Key-Sun Choi, Natasha Noy, Dean Allemang, Kyung-Il Lee, Lyndon Nixon, Jennifer Golbeck, Peter Mika, Diana Maynard, Riichiro Mizoguchi, Guus Schreiber, and Philippe Cudré-Mauroux (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 722–735.

[3] Ilias Chalkidis, Emmanouil Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2019. 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, 78–87. https://doi.org/10.18653/v1/W19-2209

[4] Ilias Chalkidis, Emmanouil Fergadiotis, Prodromos Malakasiotis, and Ion Androutsopoulos. 2019. 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, 6314–6322. https://doi.org/10.18653/v1/P19-1636

[5] Ilias Chalkidis, Manos Fergadiotis, and Ion Androutsopoulos. 2021. MultiEURLEX - A multi-lingual and multi-label legal document classification dataset for zero-shot cross-lingual transfer. *CoRR* abs/2109.00904 (2021). arXiv:2109.00904 https://arxiv.org/abs/2109.00904

[6] Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The Muppets straight out of Law School. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, Online, 2898–2904. https://doi.org/10.18653/v1/2020.findings-emnlp.261

[7] Ilias Chalkidis and Anders Søgaard. 2022. Improved Multi-label Classification under Temporal Concept Drift: Rethinking Group-Robust Algorithms in a Label-Wise Setting. https://doi.org/10.48550/ARXIV.2203.07856

[8] Wei-Cheng Chang, Hsiang-Fu Yu, Kai Zhong, Yiming Yang, and Inderjit S. Dhillon. 2019. A Modular Deep Learning Approach for Extreme Multi-label Text Classification. *CoRR* abs/1905.02331 (2019). arXiv:1905.02331 http://arxiv.org/abs/1905.02331

[9] Kunal Dahiya, Deepak Saini, Anshul Mittal, Ankush Shaw, Kushal Dave, Akshay Soni, Himanshu Jain, Sumeet Agarwal, and Manik Varma. 2021. DeepXML: A Deep Extreme Multi-Label Learning Framework Applied to Short Text Documents. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*. ACM. https://doi.org/10.1145/3437963.3441810

[10] Giovanni Comandè Daniele Licari. 2022. ITALIAN-LEGAL-BERT: A Pre-trained Transformer Language Model for Italian Law. In *The Knowledge Management for Law Workshop (KM4LAW), CEUR Workshop Proceedings*. https://ceur-ws.org/Vol-3256/km4law3.pdf

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *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)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, 4171–4186. https://doi.org/10.18653/v1/n19-1423

[12] Sarah Friedrich and Tim Friede. 2022. On the role of benchmarking data sets and simulations in method comparison studies. https://doi.org/10.48550/ARXIV.2208.01457

[13] David Grangier and Dan Iter. 2022. The Trade-offs of Domain Adaptation for Neural Language Models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Dublin, Ireland, 3802–3813. https://doi.org/10.18653/v1/2022.acl-long.264

[14] Dan Hendrycks, Collin Burns, Anya Chen, and Spencer Ball. 2021. CUAD: An Expert-Annotated NLP Dataset for Legal Contract Review. *CoRR* abs/2103.06268 (2021). arXiv:2103.06268 https://arxiv.org/abs/2103.06268

[15] Eduard Hovy, Laurie Gerber, Ulf Hermjakob, Chin-Yew Lin, and Deepak Ravichandran. 2001. Toward Semantics-Based Answer Pinpointing. In *Proceedings of the First International Conference on Human Language Technology Research*. https://www.aclweb.org/anthology/H01-1069

[16] Xin Huang, Boli Chen, Lin Xiao, and Liping Jing. 2019. Label-aware Document Representation via Hybrid Attention for Extreme Multi-Label Text Classification. *CoRR* abs/1905.10070 (2019). arXiv:1905.10070 http://arxiv.org/abs/1905.10070

[17] Deepali Jain, Malaya Dutta Borah, and Anupam Biswas. 2021. Summarization of legal documents: Where are we now and the way forward. *Computer Science Review* 40 (2021), 100388. https://doi.org/10.1016/j.cosrev.2021.100388

[18] Taehee Jung, Joo-kyung Kim, Sungjin Lee, and Dongyeop Kang. 2023. Cluster-Guided Label Generation in Extreme Multi-Label Classification. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, Dubrovnik, Croatia, 1670–1685. https://aclanthology.org/2023.eacl-main.122

[19] Sujay Khandagale, Han Xiao, and Rohit Babbar. 2019. Bonsai - Diverse and Shallow Trees for Extreme Multi-label Classification. *CoRR* abs/1904.08249 (2019). arXiv:1904.08249 http://arxiv.org/abs/1904.08249

[20] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. *CoRR* abs/1909.11942 (2019). arXiv:1909.11942 http://arxiv.org/abs/1909.11942

[21] Xin Li and Dan Roth. 2002. Learning Question Classifiers. In *COLING 2002: The 19th International Conference on Computational Linguistics*. https://www.aclweb.org/anthology/C02-1150

[22] Ilya Loshchilov and Frank Hutter. 2017. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.

[23] Eneldo Loza Mencía and Johannes Fürnkranz. 2010. *Efficient Multilabel Classification Algorithms for Large-Scale Problems in the Legal Domain*. Springer-Verlag, Berlin, Heidelberg, 192–215.

[24] Jinghui Lu, Maeve Henchion, Ivan Bacher, and Brian Mac Namee. 2021. A Sentence-Level Hierarchical BERT Model for Document Classification with Limited Labelled Data. In *Discovery Science*, Carlos Soares and Luis Torgo (Eds.). Springer International Publishing, Cham, 231–241.

[25] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning Word Vectors for Sentiment Analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Portland, Oregon, USA, 142–150. https://aclanthology.org/P11-1015

[26] Ugo Mattei. 1997. Three Patterns of Law: Taxonomy and Change in the World's Legal Systems. *The American Journal of Comparative Law* 45, 1 (01 1997), 5–44. https://doi.org/10.2307/840958 arXiv:https://academic.oup.com/ajcl/article-pdf/45/1/5/10485274/ajcl0005.pdf

[27] Christos Papaloukas, Ilias Chalkidis, Konstantinos Athinaios, Despina-Athanasia Pantazi, and Manolis Koubarakis. 2021. Multi-granular Legal Topic Classification on Greek Legislation. *CoRR* abs/2109.15298 (2021). arXiv:2109.15298 https://arxiv.org/abs/2109.15298

[28] Mobashir Sadat and Cornelia Caragea. 2022. Hierarchical Multi-Label Classification of Scientific Documents. arXiv:2211.02810 [cs.CL]

[29] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *ArXiv* abs/1910.01108 (2019).

[30] Yanyao Shen, Hsiang-Fu Yu, Sujay Sanghavi, and Inderjit S. Dhillon. 2020. Extreme multi-label classification from aggregated labels. In *ICML 2020*. https://www.amazon.science/publications/extreme-multi-label-classification-from-aggregated-labels

[31] Pang-Ning Tan, Michael S. Steinbach, and Vipin Kumar. 2005. *Introduction to Data Mining*. Addison-Wesley. http://www-users.cs.umn.edu/%7Ekumar/dmbook/

[32] Francesco Tarasconi, Milad Botros, Matteo Caserio, Gianpiero Sportelli, Giuseppe Giacalone, Carlotta Uttini, Luca Vignati, and Fabrizio Zanetta. 2020. Natural Language Processing Applications in Case-Law Text Publishing. In *International Conference on Legal Knowledge and Information Systems*.

[33] Marc Van Opijnen and Cristiana Santos. 2017. On the Concept of Relevance in Legal Information Retrieval. *Artif. Intell. Law* 25, 1 (mar 2017), 65–87. https://doi.org/10.1007/s10506-017-9195-8

[34] Ronghui You, Suyang Dai, Zihan Zhang, Hiroshi Mamitsuka, and Shanfeng Zhu. 2018. AttentionXML: Extreme Multi-Label Text Classification with Multi-Label Attention Based Recurrent Neural Networks. *CoRR* abs/1811.01727 (2018). arXiv:1811.01727 http://arxiv.org/abs/1811.01727

[35] Hsiang-Fu Yu, Kai Zhong, Inderjit S. Dhillon, Wei-Cheng Wang, and Yiming Yang. 2019. X-BERT: eXtreme multi-label text classification using bidirectional encoder representations from transformers. In *NeurIPS 2019 Workshop on Science Meets Engineering of Deep Learning*. https://www.amazon.science/publications/x-bert-extreme-multi-label-text-classification-using-bidirectional-encoder-representations-from-transformers

[36] Wei Zhao, Haiyun Peng, Steffen Eger, Erik Cambria, and Min Yang. 2019. 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, 1549–1559. https://doi.org/10.18653/v1/P19-1150

[37] Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020. How Does NLP Benefit Legal System: A Summary of Legal Artificial Intelligence. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 5218–5230. https://doi.org/10.18653/v1/2020.acl-main.466