Multi-granular Legal Topic Classification on Greek Legislation

Christos Papaloukas [†]   **Ilias Chalkidis**[‡]
Konstantinos Athinaios [†]   **Despina-Athanasia Pantazi** [†]   **Manolis Koubarakis** [†]
[†] National and Kapodistrian University of Athens, Greece
`[christospap, dpantazi, koubarak]@di.uoa.gr`
[‡] University of Copenhagen, Denmark
`ilias.chalkidis@di.ku.dk`

Abstract

In this work, we study the task of classifying legal texts written in the Greek language. We introduce and make publicly available a novel dataset based on Greek legislation, consisting of more than 47 thousand official, categorized Greek legislation resources. We experiment with this dataset and evaluate a battery of advanced methods and classifiers, ranging from traditional machine learning and RNN-based methods to state-of-the-art Transformer-based methods. We show that recurrent architectures with domain-specific word embeddings offer improved overall performance while being competitive even to transformer-based models. Finally, we show that cutting-edge multilingual and monolingual transformer-based models brawl on the top of the classifiers' ranking, making us question the necessity of training monolingual transfer learning models as a rule of thumb. To the best of our knowledge, this is the first time the task of Greek legal text classification is considered in an open research project, while also Greek is a language with very limited NLP resources in general.

1 Introduction

In recent years, there has been intensified activity in the adaptation of Artificial Intelligence technologies to the legal domain (Chalkidis and Kampas, 2018; Zhong et al., 2020; Chalkidis et al., 2021), in which legal practitioners are required to analyze and review an overwhelming amount of legal data, mostly being plain text documents. This process requires dedication and an extraordinary level of resources, both concerning human resources along with the use of automated techniques to sift rationally through data. However, more sophisticated automated techniques are able to assist legal experts in making obsolete many labour-intensive manual tasks. These techniques are mostly contained in the area of machine learning and natural language processing (NLP).

With legal text processing being an emerging subarea of NLP, many relevant applications have been derived such as legal topic classification (Nallapati and Manning, 2008; Chalkidis et al., 2020a) legal information extraction (O'Neill et al., 2017; Chalkidis et al., 2018), legal entity recognition (Chalkidis et al., 2017a; Angelidis et al., 2018; Leitner et al., 2019), court opinion generation and analysis (Wang et al., 2012; Ye et al., 2018), legal judgement prediction (Aletras et al., 2016; Xiao et al., 2018; Chalkidis et al., 2019a) and many more. However, current legal NLP studies are mainly focused in English and Chinese, with very limited resources being available in other languages.

Our work focuses on the task of *multi-class legal topic classification*, where the goal is to identify the relevant thematic topic that represents a document. In our case, thematic topics (categories) are available in a multi-level hierarchy from broader to more specialized ones. The main contributions of our work are listed below:

- We introduce Greek Legal Code (*GLC*), a dataset consisting of approx. 47k legal resources from Greek legislation. The origin of *GLC* is "Permanent Greek Legislation Code - Raptarchis", a collection of Greek legislative documents classified into multi-level (from broader to more specialized) categories.
- We study the task of multi-class legal topic classification for Greek legislation by examining a battery of advanced methods, ranging from traditional machine learning techniques and RNN-based methods, to state-of-the-art Transformer-based methods. We discuss the results and lay the groundwork for further research.

Considering that Greek is a language with few NLP resources, we anticipate that our study will be a significant contribution for the Greek NLP community. To enhance the available NLP resources and foster reproducible results, we make both our code and dataset publicly available.

2 Related Work

Mencia and Fürnkranzand (2007) introduced a legal topic classification task using a dataset obtained from the EUR-LEX[1] database, which includes EU laws that have been tagged with EUROVOC concepts. They used multiple binary Perceptrons, one for each label, and multi-label pairwise Perceptrons on top of Bag-of-Words (BoW) representations. While the methods seem primal and inefficient by today's standards, the EURLEX dataset is widely adopted as a notable benchmark in Large-scale Multi-label Text Classification (LMTC) literature.

Nallapati and Manning (2008) were also among the first who investigated the task of text classification in the legal domain where machine learning classifiers such as SVMs were insufficient. They experimented with a dataset of 5.5k US docket entries of court cases. The authors stress the importance of feature selection in such specialized domains and expose the limitations of classifiers relying on BoW featues to capture the intricacies of natural language, widespread in specialized domains such as the legal one.

Undavia et al. (2018) applied neural networks on legal document classification in a similar task, classification of legal court opinions. They used a dataset of 8k US Supreme Court (SCOTUS) opinions, where they targeted two sub-tasks, depending on the total output categories: 15 broad and 279 finer-grained categories. They experimented with shallow neural networks using different word embeddings, where their best model (word2vec + CNN) scored 72.4% accuracy in the 15-classes task and 31.9% accuracy in the 279-classes task. They concluded claiming that an RNN-based network together with domain-specific word embeddings could possibly tackle the task with higher accuracy.

Chalkidis et al. (2019c) experimented with several classifiers on a novel dataset of 57k legislative documents (EURLEX57k) from EUR-LEX in English. They demonstrated that that BiGRUs with self attention outperform CNN-based methods that employ the label-wise attention mechanism. Using domain-specific word embeddings and context-sensitive ELMO (Peters et al., 2018) embeddings improves the overall performance. Furthermore, the authors experimented with BERT (Devlin et al., 2019) obtaining the best results

In a more recent and extended version of this study, Chalkidis et al. (2020b) evaluated a battery of LMTC methods ranging from RNN-based Label-Wise Attention Networks (LWANs) to Probabilistic Label Trees (PLTs) (Prabhu et al., 2018; Khandagale et al., 2019; You et al., 2019) and Transformer-based models (BERT, ROBERTA) (Devlin et al., 2019; Liu et al., 2019) on three English datasets: EURLEX57k (Chalkidis et al., 2019c), MIMIC-III (Johnson et al., 2016) and AMAZON13k (Lewis et al., 2004). The experimental results show that PLT-based methods outperform LWANs, while Transformer-based approaches surpass state-of-the-art in two out of three datasets. Furthermore, a new state-of-the-art method is introduced which combines BERT and LWAN, giving the best results overall. Furthermore, the case of few and zero-shot learning is studied with new models that leverage the label hierarchy and yield better results.

Following literature, we examine traditional machine learning methods and the RNN-based methods used in Chalkidis et al. (2019c) and Chalkidis et al. (2020b) and investigate whether these methods can perform equally well in multi-class text classification as they do in the multi-label setting. Furthermore, we examine several BERT-based methods, including multi-lingual models (Devlin et al., 2019; Conneau et al., 2020) that have not been studied to date in the context of legal NLP.

Most of the preceding efforts focus on the English language and limited studies and resources exist considering languages other than English and Chinese. While recently there is a new wave of studies on NLP tasks focused on the Greek language (Athanasiou and Maragoudakis, 2017; Papantoniou and Tzitzikas, 2020; Pitenis et al., 2020; Koutsikakis et al., 2020), there is limited work for legal NLP in Greek (Angelidis et al., 2018). To the best of our knowledge, this is the first study on the task of Greek legal text classification, where experiments range from traditional machine learning to transfer learning models.

3 GLC Dataset

3.1 Original Data

The "Permanent Greek Legislation Code - Raptarchis[2]" is a thorough catalogue of Greek legislation since the creation of the Greek state in 1834 until 2015. It includes Laws, Royal and Presidential Decrees, Regulations and Decisions, retrieved from the Official Government Gazette,

---

[1]See https://eur-lex.europa.eu/.

[2]Mr. P. Raptarchis is the original curator of this collection.

where Greek legislation is published. This collection is one of the official, publicly available sources of classified Greek legislation suitable for our classification task.[3]

Currently, the original catalogue is publicly offered in MS Word (.doc) format through the portal e-Themis[4], the legal database and management service of it, under the administration of the Ministry of the Interior (Affairs)[5]. E-Themis is primarily focused on providing legislation on a multitude of predefined thematic categories, as described in the catalogue. The main goal is to help users find legislation of interest using the thematic index.

The original collection follows a bibliographic structure (Figure 1).[6] It consists of 47 legislative *volumes* and each volume corresponds to a main thematic topic. Inside each volume, the main thematic topic is divided into thematic subcategories which are called *chapters* and subsequently, each chapter breaks down to *subjects* which contain the legal resources. The total number of chapters is 389 while the total number of subjects is 2285, creating an interlinked thematic hierarchy. Thus the task is defined as a *multi-level text classification task*, where the goal is to predict the thematic category at each level (volume, chapter, subject).

An example of this thematic hierarchy is the volume of "Criminal Law", which is divided into 9 chapters. "International Criminal Law" and "Military Criminal Law" are two out of the nine chapters. Subsequently, "International Criminal Law" is subdivided into 8 subjects (e.g., "Genocide", "Counterfeiting" etc.) and "Military Criminal Law" is subdivided into 4 subjects (e.g., "Military Criminal Code", "Legal Remedies" etc.). Another example is the volume of "Labour Law", divided into 17 chapters. Two of those are "Collective Employment Contracts" and "Hygiene and Safety of Employees". "Collective Employment Contracts" is subdivided into 7 subjects (e.g., "Limits of Salaries", "Holiday Allowance" etc.) and "Hygiene and Safety of Employees" is subdivided into 3 subjects (e.g., "Hygiene And Safety Of Workplaces And Employees", "Work Health Books" etc.).
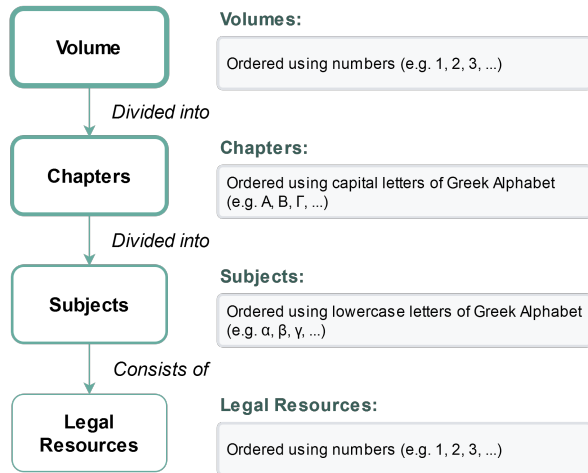
For the study presented in this paper, we curated

Figure 1: Original GLC thematic hierarchy

and publicly release a new dataset named Greek Legal Code (*GLC*), containing all the available legislation from the original catalogue in JSON format. Each JSON file contains a legal resource (law) along with its metadata, i.e., thematic topics, publication year, document type, etc., as they were extracted from the original documents.[7]

3.2 Dataset Statistics

In this section, we present a detailed quantitative analysis of the final dataset, consisting of 47,563 documents (i.e., categorized legal resources).

Data Split *GLC* is split into three subsets: training (60%), development (20%) and test (20%), as shown in Table 1. The documents are distributed uniformly for all levels of the class hierarchy in order to achieve the same level of partitioning from bottom to top (i.e., from each subject to the whole dataset).

Subset	Docs	Mean # of tokens / doc	(<100 tokens)
Train (60%)	28536	600	15412 (54%)
Dev. (20%)	9511	574	5175 (54.4%)
Test. (20%)	9516	595	5075 (53.3%)
Total:	47563	594	25662 (54%)

Table 1: Dataset split and statistics.

Document Size In Figure 2, we observe that more than half of the documents in our dataset have less than 100 tokens. We see that most of the documents are not that long, with their total number of tokens being below the mean number of tokens across all documents. Many records (documents) in GLC, especially the older ones have only

---

[3]Another official source is that of European Legislation written in modern Greek, available at: https://eur-lex.europa.eu/browse/directories/legislation.html.

[4]The portal is hosted at https://www.secdigital.gov.gr/e-themis/, where you can find the thematic index (hierarchy).

[5]https://www.ypes.gr

[6]In fact, it originated through a proper printed thesaurus.

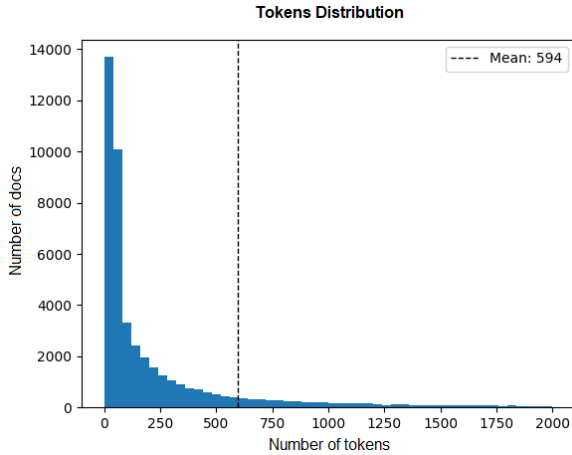[7]In Appendix A, we present in great detail the preprocessing procedure.

Figure 2: Tokens' distribution over docs in GLC

the descriptive title of the legal resource or a small part of it and not the full text.

	Total	Frequent	Few-shot (<10)	Zero-shot
Volume	47	47 (100%)	0	0
Chapter	389	333 (85.6%)	53 (13.6%)	3 (00.7%)
Subject	2285	712 (31.2 %)	1431 (62.6%)	142 (06.2%)

Table 2: Number of classes per thematic level and their distribution to frequent-few-zero categories

Label Frequency *GLC* classes are divided into three categories for each thematic level: *frequent classes*, which occur in more than 10 training documents and can be found in all three subsets (training, development and test); *few-shot classes*[8] which appear in 1 to 10 training documents and also appear in the documents of the development and test sets, and *zero-shot classes* which appear in the development and/or test, but not in the training documents. As demonstrated in Table 2, many classes are under-represented, especially in the thematic level of subjects, causing the appearance of few- and zero-shot categories. The appearance of under-represented classes increases as we move towards more specific thematic levels. Table 3 shows the total number of documents per category and thematic level.

	Total	Frequent	Few-shot (<10)	Zero-shot
Volume	47563	47563 (100%)	0	0
Chapter	47563	47108 (99.0%)	445 (00.9%)	10 (<00.1%)
Subject	47563	38475 (80.9%)	8870 (18.6%)	218 (00.5%)

Table 3: Number of documents as labeled per thematic level and their distribution to frequent-few-zero categories

---

[8]The terms few-shot and zero-shot are used to express the underrepresentation of classes in *GLC* rather than transfer learning approaches as in other deep learning literature.

In the volume level, all the classes belong to the frequent category and are sufficiently represented, as more than 10 documents per class exist in the training data. In the chapter level, few-shot classes appear and are rather underrepresented as most documents are classified among frequent classes, leaving less than 1% of the total documents to be associated with ~14% of the total classes. In the subject level, data are even more unequally distributed over classes. The majority of documents are classified into frequent classes, leaving more than half of the total classes (~63%) to be associated with less than 20% of the total documents, along with 142 classes having zero representation at the training subset.

4 Task and Methods

4.1 Multi-class Legal Topic Classification for Greek Legislation

We study multi-class legal topic classification for Greek legislation using the *GLC* dataset[9]. *GLC* offers 3 hierarchical levels of thematic categorization, forming a tree structure that follows the original GLC organization as explained in Section 3.1. We divide the classification task for *GLC* into three separate classification tasks. Each task deals with a different level of the thematic hierarchy (i.e., volume, chapter or subject) and the classifier's objective is to predict the correct class out of all the classes in this specific thematic level.

4.2 Methods

In our study, we consider an arsenal of nine methods ranging from traditional machine learning techniques (2) and RNN-based methods (3), to state-of-the-art Transformer-based methods (4).

SVM-BOW: Support Vector Machines (SVMs) is a strong baseline for the multi-class classification problem as it is one of the highest performing traditional ML methods. We represent the legal documents using Bag-of-Words (BoW) features, getting the most frequent n-grams across all training data, weighted by TF-IDF.

XGBOOST-BOW: XGBoost (Chen and Guestrin, 2016) is a scalable, fast and robust open-source implementation[10] of the Gradient Boosting decision tree algorithm. In XGBoost, the model is fitted on the gradient of loss generated from the previous

---

[9]Available at: https://huggingface.co/datasets/greek_legal_code
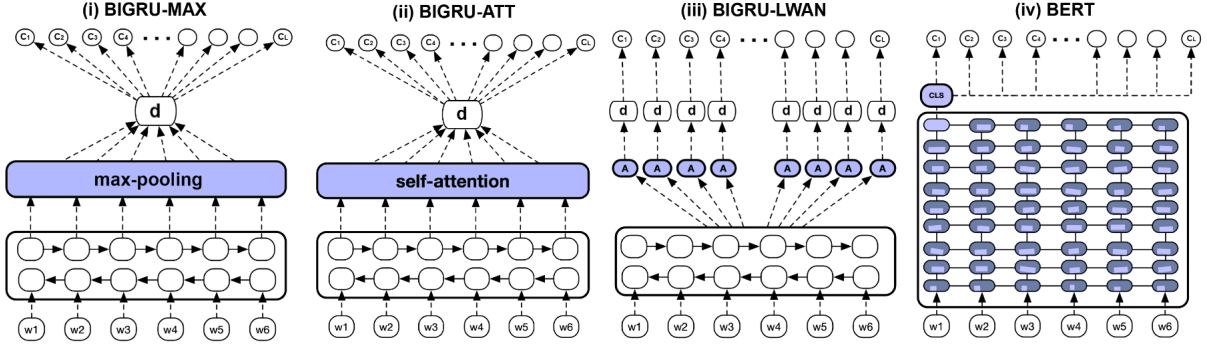[10]Available at: https://github.com/dmlc/xgboost/

Figure 3: Illustration of (i) BIGRU-MAX, (ii) BIGRU-ATT, (iii) BIGRU-LWAN and (iv) BERT

step and the gradient boosting algorithm is modified so that it works with any differentiable loss function. In our case, the documents are represented using BoW features weighted by TF-IDF (similarly to SVM-BOW).

BIGRU-MAX: The first RNN-based method that we examine is BiGRU with max-pooling, employing pre-trained, domain-specific word embeddings (Angelidis et al., 2018).[11] The stacked BiGRU encoder converts the pre-trained word embeddings $\mathbf{w}_i$ into context-aware embeddings $\mathbf{h}_i$. Context-aware token embeddings pass through a max-pooling layer to produce the final document representation $\mathbf{d}$ by reducing the initial matrix (token-wise vectors) into a single vector but hopefully, keep the most salient information. Finally, a dense layer with L output units and the *softmax()* activation is deployed, to transform the document representation $\mathbf{d}$ into a probability distribution over L classes. L is defined according to the examined task, i.e., $L=47$ for the volume-level task, $L=389$ for the chapter-level task and $L=2285$ for the subject-level task.

BIGRU-ATT: The second RNN-based method is a BiGRU network with self-attention (Xu et al., 2015; Chalkidis et al., 2019c). This methods uses the very same word embedding and BiGRU encoder, similarly to the previously described method (BIGRU-MAX). Instead of max-pooling, it uses the self-attention mechanism to produce a final document representation $\mathbf{d}$. This representation is computed as the weighted sum of the BiGRU context-aware embeddings $\mathbf{h}_i$, weighted by the self-attention scores a_i produced as:

$$a_i = \frac{\exp\left(\mathbf{h}_i^\top \mathbf{u}\right)}{\sum_j \exp\left(\mathbf{h}_j^\top \mathbf{u}\right)} \qquad (1)$$

$$\mathbf{d} = \frac{1}{T} \sum_{i=1}^{T} a_i \mathbf{h}_i \qquad (2)$$

T represents the document's length in words while $\mathbf{u}$ is a trainable vector used to compute the attention scores a_i over $\mathbf{h}_i$. Similarly to BIGRU-MAX, a final dense layer with L output units and *softmax()* activation is deployed to predict the correct output class using a probability distribution over all the classes.

BIGRU-LWAN: The third RNN-based method replaces the self-attention mechanism of BIGRU-ATT with the label-wise attention mechanism. The original Label-Wise Attention Network (LWAN) was introduced in Mullenbach et al. (2018) using a CNN-based encoder. Later on, (Chalkidis et al., 2019c,b) replaced it with a BiGRU encoder. In contrast with BIGRU-ATT, this label-wise attention technique uses L independent attention heads, one per class, generating L label-wise document representations $\mathbf{d}_l$ from the sequence of $\mathbf{h}_i$ vectors produced by the BiGRU encoder. The intuition is that each label-wise document embedding is dedicated in predicting the corresponding class, focusing on possibly different aspects of each representation $\mathbf{h}_i$. In effect, different parts of the representation may be more relevant for different classes.

$$a_{li} = \frac{\exp\left(\mathbf{h}_i^\top \mathbf{u}_l\right)}{\sum_{i'} \exp\left(\mathbf{h}_{i'}^\top \mathbf{u}_l\right)} \qquad (3)$$

$$\mathbf{d}_l = \frac{1}{T} \sum_{i=1}^{T} a_{li} \mathbf{h}_i \qquad (4)$$

Again, T represents the document's length in words while $\mathbf{u}_l$ ($l = 1, 2, ..., L$) is a trainable vector

---

[11] Angelidis et al. (2018) released a Word2Vec model trained on Greek legal corpora, including Greek national and EU legislation.

used to compute the attention scores a_{li} of the l-th attention head. Then, each attention head $\mathbf{d}_l$ goes through an independent dense layer for each class and similarly to the previous methods, *softmax()* activation is deployed to transform the document representation $\mathbf{d}$ into a probability distribution over L classes.

M-BERT: BERT (Devlin et al., 2019) is a Transformer-based (Vaswani et al., 2017) language model initially developed by Google. In BERT, deep bidirectional representations are pre-trained from unlabeled text by jointly conditioning on both left and right context. As a result, for any new task, the pre-trained BERT model can be fine-tuned with just one additional task-specific output layer trained with task-specific data. We employ the multilingual version of the 12-layer BERT-BASE-CASED model, which supports modern Greek, alongside 99 languages, out-of-the-box. We add linear layer on top of M-BERT encoder followed by a *softmax()* activation. This extra dense layer is fed with the so-called "classification token" ([cls]) of the BERT encoder as described in (Devlin et al., 2019), serving as the final document representation.

XLM-ROBERTA: The next BERT-based model we examine is XLM-RoBERTa (Conneau et al., 2020), a multilingual adaptation of RoBERTa (Liu et al., 2019). RoBERTa is built following BERT's architecture, while removed the next-sentence pre-training objective and trained with much larger batches and higher learning rate. Based on this study, Conneau et al. (2020) proposed the XLM-RoBERTa model, which supports 100 different languages (Greek included) and is trained on 2.5TB of filtered Common Crawl data.

GREEK-BERT: We also experiment with GREEK-BERT (Koutsikakis et al., 2020), a native monolingual version of BERT, trained solely on modern Greek, achieving state-of-the-art results in several NLP tasks. GREEK-BERT was pre-trained on 29GB of text from a corpus consisting of the Greek part of Wikipedia, the Greek part of the European Parliament Proceedings Parallel Corpus (Europarl) (Koehn, 2005) and OSCAR (Ortiz Suárez et al., 2019), a clean version of Common Crawl. Even though multilingual models like the previous two offer exceptional performance, monolingual models usually surpass them in most downstream tasks. Again, GREEK-BERT follows the previous BERT models configuration.

GREEK-LEGAL-BERT: Finally, we experiment with GREEK-LEGAL-BERT (Athinaios, 2020), another BERT-flavored model for modern Greek. Its pre-training corpus is based entirely on Greek legal documents and consists of around 5GB of Greek and EU legislation documents retrieved by Nomothesia[12] (Chalkidis et al., 2017b), a Greek legislative Knowledge Base. It also follows the previously described configuration.

Method	Hyperparameters
SVM-BOW	Kernel: [linear, rbf] C: [0.25, 0.50, 0.75, 1] N-Grams: [(1,3), (1,5)] Max-Features: [200k, 400k]
XGBOOST-BOW	Booster: gbtree with softmax N-Estimators: 800 Max-Depth: [4, 5, 7, 10] Min-Child-Weight: [2, 5, 10]
BIGRU-*	HYPEROPT[13]on: Layers: [1, 2] Units: [200, 300, 400] Dropout: [0.1, 0.2, 0.3, 0.4] Word Dropout: [0, 0.01, 0.02] Batch Size: [8, 16]
BERT	Dropout: 0.1 Learning Rate: [1e-5, 2e-5, 3e-5, 5e-5] Batch Size: 8

Table 4: Hyper parameters search space for all models

5 Experiments

5.1 Experimental SetUp

We tune all methods by grid-searching over the core hyper-parameters (per method) presented in Table 4, except the RNN-based methods where we used HyperOpt due to the large search space. We select the models with the best performance on the development set. We use early stopping on the development loss using the Adam (Kingma and Ba, 2015) optimizer. All BERT models follow the base configuration, i.e., 12 layers with 786 units and 12 attention heads each.[14] We report micro-averaged Precision (P), Recall (R), and F1-score (F1) on the test set. For each method, we perform 5 runs with different seeds and report the average. We report results for the frequent and few-shot classes (when available), omitting results for zero-shot classes since our methods are incapable of zero-shot learning. We release our code for reproducibility:[15]

---

[12]http://legislation.di.uoa.gr.
[13]See: https://hyperopt.github.io/hyperopt/
[14]All models are available at https://huggingface.co/models.
[15]See: https://github.com/christospi/glc-nllp-21.

MODEL	Volume (47 classes)						Chapter (389 classes)								
	ALL LABELS			FREQUENT			ALL LABELS			FREQUENT			FEW-SHOT		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
SVM-BOW	85.3	85.3	85.3	85.3	85.3	85.3	77.9	77.9	77.9	77.9	78.6	78.2	**90.0**	09.3	16.8
XGBOOST-BOW	77.2	77.2	77.2	77.2	77.2	77.2	67.5	67.5	67.5	67.8	68.1	67.9	19.2	10.3	13.4
BIGRU-MAX	84.3	84.3	84.3	84.3	84.3	84.3	77.5	77.5	77.5	77.9	77.9	77.9	44.9	45.4	45.1
BIGRU-ATT	86.4	86.4	86.4	86.4	86.4	86.4	81.1	81.1	81.1	81.1	81.6	81.3	86.7	40.2	54.9
BIGRU-LWAN	84.1	84.1	84.1	84.1	84.1	84.1	76.8	76.8	76.8	76.9	77.3	77.1	63.8	30.9	41.7
M-BERT	85.8	85.8	85.8	85.8	85.8	85.8	80.2	80.2	80.2	80.1	80.6	80.4	81.0	43.9	56.6
XLM-ROBERTA	85.0	85.0	85.0	85.0	85.0	85.0	80.6	80.6	80.6	80.6	81.0	80.8	80.3	37.9	51.3
GREEK-BERT	87.5	87.5	87.5	87.5	87.5	87.5	82.4	82.4	82.4	82.4	82.8	82.6	81.0	45.2	**57.8**
GREEK-LEGAL-BERT	**89.4**	**89.4**	**89.4**	**89.4**	**89.4**	**89.4**	**84.3**	**84.3**	**84.3**	**84.4**	**84.8**	**84.6**	79.0	**45.8**	57.8

Table 5: Experimental results for *Volume* and *Chapter* levels.

5.2 Volume-level Classification Evaluation

In volume-level classification, all the 47 possible classes are sufficiently represented and belong to the frequent category, while the mean number of documents per class is approximately 1k. However, we acknowledge class imbalance (label skewness) as a challenge in our dataset, though not so evident here. Table 5 (first zone) demonstrates the results for this task. SVM-BOW proves to be a strong competitor in the first classification task (85.3 in F1). Interestingly, it outperforms two of our neural methods with domain-specific word embeddings, namely BIGRU-MAX and BIGRU-LWAN with 84.3 and 84.1 F1 scores, respectively. As for XG-Boost, although it seemed quite promising and very fast at training, its inadequate F1 score of 77.2 places it at the bottom of the list for this task.

Among the RNN-based neural methods, BIGRU-ATT achieves an F1 score of 86.4 outperforming the rest of the RNN-based methods. Its results indicate the significance of two of its fundamental features: (i) the domain-specific word2vec embeddings and (ii) the cumulative self-attention head that provides an advantageous final document representation. Compared to BIGRU-MAX, we believe that its max-pooling layer hinders some of the document's particularities and thus, it yields a lower score. Likewise, the BIGRU-LWAN method with L different attention heads seems to be more tailor-made for multi-label classification tasks, as it does not offer any performance improvement compared to BIGRU-ATT.

GREEK-LEGAL-BERT proves to be the best method we experimented with, achieving a score of 89.4 in F1, followed by the generic GREEK-BERT with a score of 87.5 in F1. The two multilingual models (M-BERT and XLM-ROBERTA) also achieve satisfying F1 scores (85.8 and 85.0), confirming their claim to offer top-notch results in most downstream NLP tasks. The results demonstrate that monolingual models are able to surpass other advanced multilingual transformer-based models. We hypothesize that the fact that GREEK-LEGAL-BERT is pre-trained entirely on Greek legal corpora is the main reason of its superiority, while the specialization in the Greek language, alongside GREEK-BERT, seems even more critical.

5.3 Chapter-level Classification Evaluation

In Table 5 (second zone) we observe again that XGBOOST-BOW has the lowest performance with 67.5 in overall F1 score with SVM-BOW being the second worst overall (77.9) alongside BIGRU-MAX. The ranking of BERT-based models is similar to the previous task (Volume-level classification), i.e., monolingual models outperform their multilingual counterparts, while GREEK-LEGAL-BERT is better by 2% compared to the generic GREEK-BERT model. These results further supporting our intuition on the importance of native language support and domain knowledge.

Inspecting the few-shot label results, we observe that BIGRU-ATT and BERT models outperform BIGRU-MAX by 10%, while the traditional machine learning approaches have terrible performance (approx. 15 F1). These results highlight the importance of the attention mechanism and its capability to focus in specific parts of the text that are more prominent in relation to the downstream task and the specific labels.

5.4 Subject-level Classification Evaluation

In the last sub-task, where there are many more classes (2285 in total), we observe a similar ranking in the methods (Table 6). However, the absolute difference in few-shot labels' scores is higher this time, when comparing either the attention-based method with BIGRU-MAX (approx. +20% F1) or monolingual with multilingual BERT models (approx. +9% F1).

MODEL	Subject (2285 classes)								
	ALL LABELS			FREQUENT			FEW-SHOT		
	P	R	F1	P	R	F1	P	R	F1
SVM-BOW	37.9	37.9	37.9	37.9	47.8	42.3	00.0	00.0	00.0
XGBOOST-BOW	55.3	55.3	55.3	56.1	64.8	60.1	46.9	19.1	27.2
BIGRU-MAX	62.9	62.9	62.9	66.0	70.5	68.1	47.1	37.8	42.0
BIGRU-ATT	74.8	74.8	74.8	75.3	79.6	77.4	72.6	61.1	66.3
BIGRU-LWAN	65.2	65.2	65.2	68.1	72.8	70.4	50.7	40.4	45.0
M-BERT	76.8	76.8	76.8	79.8	82.5	81.1	64.4	59.7	62.0
XLM-ROBERTA	78.0	78.0	78.0	80.3	83.4	81.8	68.2	62.1	65.0
GREEK-BERT	79.4	79.4	79.4	80.5	83.9	82.2	**74.6**	67.0	70.6
GREEK-LEGAL-BERT	**81.2**	**81.2**	**81.2**	**83.0**	**85.5**	**84.2**	73.6	**69.4**	**71.4**

Table 6: Experimental results for *Subject* level.

5.5 General Observations

There are two main observations:

- Pre-trained Transformer-based models perform exceptionally and the performance gains compared to traditional machine learning methods and RNN-based methods are increasing in relation to the number of labels. We can only speculate the following: (a) sub-word units used by Transformer-based models potentially play an important role in morphological rich languages like Greek, compared to full words processed by the rest of the examined models; and (b) the multi-head attention mechanism can better distill important detailed (specialized) information to distinguish categories (labels).

- The Label-wise Attention Network (LWAN) under-performs in the newly introduced multi-class text classification task. Again, we can only speculate that the produced label-wise document representations that improve results over the standard universal attention mechanism in the multi-label setting (Chalkidis et al., 2020a) lead to a "greedy" over-scoring (produced logits) followed by an aggressive label competition (softmax over logits) in the examined multi-task setting leading to negative (poor) results.

We aim to further review and validate these observations, while also study the reasons behind those thoroughly in the future, possibly using more datasets and ablation studies (e.g., train RNN-based methods with pretrained sub-word embeddings).

6 Conclusions and Future Work

We introduced Greek Legal Code (*GLC*), a new publicly available dataset consisting of 47k Greek legislation resources. Relying on this dataset, we experimented with several classifiers, ranging from traditional machine learning and recurrent models to state-of-the-art transfer learning models. Through their performance evaluation, we realized that although traditional machine learning classifiers (e.g., SVM-BOW) set strong baselines for

some of the considered tasks, they fall short against more sophisticated methods. In contrast, RNN-based methods relying on BiGRUs provide improved overall performance and were competitive to multilingual transformer-based architectures (M-BERT, XLM-ROBERTA). Beyond doubt, monolingual transformer-based models (GREEK-BERT and GREEK-LEGAL-BERT) lead to state-of-the-art results, especially when they are pre-trained on in-domain corpora.

Nonetheless, more emphasis should be given to the qualitative and quantitative characteristics of the examined datasets. Intricacies like class imbalance, data scarcity and diversity apparently need special handling. Regarding our study, we noticed that few-shot and especially zero-shot classes need to be properly handled with appropriate methods, as standard classifiers are insufficient. As for the recent trend to develop novel monolingual BERT-based models, results show that already established multilingual models are incredibly powerful even in monolingual tasks. While research is on-going and these models are continuously being improved, also taking into consideration the computational costs, it is quite challenging to motivate researchers into making an effort to train monolingual models for medium or small-sized languages; especially when multilingual models can perform equally well or occasionally, even better.

In future work, we plan to investigate specialized methods with improved few-shot and zero-shot performance (Hu et al., 2018; Rios and Kavuluru, 2018; Chalkidis et al., 2019c) that leverage various data properties (e.g., label descriptors and label hierarchy). Also, we intend to apply deep learning techniques that take into account the hierarchy of classes in datasets like *GLC*, where there is an underlying taxonomy (Kowsari et al., 2017; Chalkidis et al., 2020b; Manginas et al., 2020). Furthermore, experimenting with similar datasets like that of EU Legislation written in Greek will allow us to confirm our current conclusions. For example in the future, we could consider Cypriot legislation[16] to evaluate the out-of-domain generalization of models trained on Greek legislation. Finally, our long term goal is to support and encourage further research in NLP for the Greek language by publishing novel datasets, introducing and experimenting with state-of-the-art methods.

---

[16]http://www.cylaw.org/nomoi/

Acknowledgements

References

N. Aletras, D. Tsarapatsanis, D. Preotiuc-Pietro, and V. Lampos. 2016. Predicting judicial decisions of the european court of human rights: a natural language processing perspective. *PeerJ Comput. Sci.*, 2.

Iosif Angelidis, Ilias Chalkidis, and Manolis Koubarakis. 2018. Named entity recognition, linking and generation for greek legislation. In *JURIX*.

V. Athanasiou and M. Maragoudakis. 2017. A novel, gradient boosting framework for sentiment analysis in languages where nlp resources are not plentiful: A case study for modern greek. *Algorithms*, 10:34.

Konstantinos Athinaios. 2020. Named Entity Recognition using a Novel Linguistic Model for Greek Legal Corpora based on BERT model. *BS Thesis, School of Science, Department of Informatics and Telecommunications*.

Ilias Chalkidis, Ion Androutsopoulos, and Nikolaos Aletras. 2019a. Neural legal judgment prediction in English. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4317–4323, Florence, Italy. Association for Computational Linguistics.

Ilias Chalkidis, Ion Androutsopoulos, and Achilleas Michos. 2017a. Extracting contract elements. In *Proceedings of the 16th Edition of the International Conference on Articial Intelligence and Law*, ICAIL '17, page 19–28, New York, NY, USA. Association for Computing Machinery.

Ilias Chalkidis, Ion Androutsopoulos, and Achilleas Michos. 2018. Obligation and prohibition extraction using hierarchical RNNs. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 254–259, Melbourne, Australia. Association for Computational Linguistics.

Ilias Chalkidis, Emmanouil Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2019b. Extreme multi-label legal text classification: A case study in EU legislation. In *Proceedings of the Natural Legal Language Processing Workshop 2019*, pages 78–87, Minneapolis, Minnesota. Association for Computational Linguistics.

Ilias Chalkidis, Emmanouil Fergadiotis, Prodromos Malakasiotis, and Ion Androutsopoulos. 2019c. Large-scale multi-label text classification on EU legislation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6314–6322, Florence, Italy. Association for Computational Linguistics.

Ilias Chalkidis, Manos Fergadiotis, Sotiris Kotitsas, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020a. An empirical study on large-scale multi-label text classification including few and zero-shot labels. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7503–7515, Online. Association for Computational Linguistics.

Ilias Chalkidis, Manos Fergadiotis, Sotiris Kotitsas, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020b. An empirical study on large-scale multi-label text classification including few and zero-shot labels. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7503–7515, Online. Association for Computational Linguistics.

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Martin Katz, and Nikolaos Aletras. 2021. LexGLUE: A Benchmark Dataset for Legal Language Understanding in English.

Ilias Chalkidis and Dimitrios Kampas. 2018. Deep learning in law: early adaptation and legal word embeddings trained on large corpora. *Artificial Intelligence and Law*, 27:171–198.

Ilias Chalkidis, Charalampos Nikolaou, Panagiotis Soursos, and Manolis Koubarakis. 2017b. Modeling and querying greek legislation using semantic web technologies. In *The Semantic Web*, pages 591–606, Cham. Springer International Publishing.

T. Chen and C. Guestrin. 2016. Xgboost: A scalable tree boosting system. *ACM-SIGKDD*.

A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. *ArXiv*, abs/1911.02116.

J. Devlin, M. Chang, K. Lee, and K. Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*.

Z. Hu, X. Li, C. Tu, Z. Liu, and M. Sun. 2018. Few-shot charge prediction with discriminative legal attributes. In *COLING*.

A. Johnson, T. J. Pollard, L. Shen, L. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Celi, and R. G. Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Sujay Khandagale, Han Xiao, and Rohit Babbar. 2019. Bonsai - Diverse and Shallow Trees for Extreme Multi-label Classification. *CoRR*, abs/1904.08249.

D. P. Kingma and J. Ba. 2015. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MTSUMMIT*.

John Koutsikakis, Ilias Chalkidis, Prodromos Malakasiotis, and Ion Androutsopoulos. 2020. Greek-bert: The greeks visiting sesame street. In *11th Hellenic Conference on Artificial Intelligence*, SETN 2020, page 110–117, New York, NY, USA. Association for Computing Machinery.

K. Kowsari, D. Brown, M. Heidarysafa, K. Meimandi, M. Gerber, and L. Barnes. 2017. Hdltex: Hierarchical deep learning for text classification. *ICMLA*.

Elena Leitner, Georg Rehm, and Julian Moreno-Schneider. 2019. Fine-grained Named Entity Recognition in Legal Documents. In *Semantic Systems. The Power of AI and Knowledge Graphs. Proceedings of the 15th International Conference (SEMANTiCS 2019)*, number 11702 in Lecture Notes in Computer Science, pages 272–287, Karlsruhe, Germany. Springer. 10/11 September 2019.

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

Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Nikolaos Manginas, Ilias Chalkidis, and Prodromos Malakasiotis. 2020. Layer-wise guided training for BERT: Learning incrementally refined document representations. In *Proceedings of the Fourth Workshop on Structured Prediction for NLP*, pages 53–61, Online. Association for Computational Linguistics.

Eneldo Loza Mencia and Johannes Fürnkranzand. 2007. An Evaluation of Efficient Multilabel Classification Algorithms for Large-Scale Problems in the Legal Domain. In *Proceedings of the 1st Linguistic Annotation Workshop*, pages 126–132, Halle, Germany.

J. Mullenbach, S. Wiegreffe, J. Duke, J. Sun, and J. Eisenstein. 2018. Explainable prediction of medical codes from clinical text. *ArXiv*, abs/1802.05695.

R. Nallapati and C. D. Manning. 2008. Legal docket classification: Where machine learning stumbles. *EMNLP*.

J. O'Neill, P. Buitelaar, C. Robin, and L. O' Brien. 2017. Classifying Sentential Modality in Legal Language: A Use Case in Financial Regulations, Acts and Directives. In *Proceedings of the 16th International Conference on Artificial Intelligence and Law (ICAIL)*, pages 159–168, London, UK.

Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019, pages 9 – 16, Mannheim. Leibniz-Institut für Deutsche Sprache.

K. Papantoniou and Y. Tzitzikas. 2020. Nlp for the greek language: A brief survey. *11th Hellenic Conference on Artificial Intelligence*.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Z. Pitenis, M. Zampieri, and T. Ranasinghe. 2020. Offensive language identification in greek. In *LREC*.

Yashoteja Prabhu, Anil Kag, Shrutendra Harsola, Rahul Agrawal, and Manik Varma. 2018. Parabel: Partitioned Label Trees for Extreme Classification with Application to Dynamic Search Advertising. In *Proceedings of the 2018 World Wide Web Conference*, WWW '18, pages 993–1002, Republic and Canton of Geneva, Switzerland.

A. Rios and R. Kavuluru. 2018. Few-shot and zero-shot multi-label learning for structured label spaces. *EMNLP*.

S. Undavia, A. Meyers, and J. Ortega. 2018. A comparative study of classifying legal documents with neural networks. *FedCSIS*.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. 2017. Attention is all you need. In *NIPS*.

W. Yang Wang, E. Mayfield, S. Naidu, and J. Dittmar. 2012. Historical analysis of legal opinions with a sparse mixed-effects latent variable model. In *ACL*.

Chaojun Xiao, Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Zhiyuan Liu, Maosong Sun, Yansong Feng, Xianpei Han, Zhen Hu, Heng Wang, and Jianfeng Xu. 2018. CAIL2018: A Large-Scale Legal Dataset for Judgment Prediction. *arXiv:1807.02478 [cs]*. ArXiv: 1807.02478.

K. Xu, J. Ba, R. Kiros, K. Cho, A. C. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*.

H. Ye, X. Jiang, Z. Luo, and W. Chao. 2018. Interpretable charge predictions for criminal cases: Learning to generate court views from fact descriptions. *ArXiv*, abs/1802.08504.

Ronghui You, Zihan Zhang, Ziye Wang, Suyang Dai, Hiroshi Mamitsuka, and Shanfeng Zhu. 2019. AttentionXML: Label Tree-based Attention-Aware Deep Model for High-Performance Extreme Multi-Label Text Classification. In *Advances in Neural Information Processing Systems 32*, pages 5812–5822. Curran Associates, Inc.

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*, pages 5218–5230, Online. Association for Computational Linguistics.

A Dataset Curation

We describe in-detail the processing method we followed to generate *GLC* and describe the data structure of the final JSON documents.

A.1 Parsing the Original Documents

The original legislative volumes are encoded in MS Word (.doc) format. While most of them follow the double-column format, there are cases where they also include text in single-column format or even include scanned documents or images as legal resources, making the initial data quite noisy. Considering our objective, these abnormalities should be revised, and all the additional metadata contained in the .doc file (e.g., font style, size, page margins) should be removed. Thus, converting these documents into plain text files was of significant importance. To achieve that, we used docx2txt[18], a python utility that detects and extracts text from .doc files.

Examining the output text files, we encountered problematic samples that needed special handling. For example, we found significant keywords (e.g., "ΘΕΜΑ" which means subject) missing from text or even having typos, subject IDs found inline with subject titles etc., mostly due to minor inaccuracies in the text conversion. Furthermore, multiple white spaces, multiple new lines and special or corrupted characters occurred in the text. To overcome these complications, we performed data cleaning using heuristics and regular expressions to produce neat text files that follow the same normalized structural pattern.

Next, we built a rule-based parser in Python which receives these text files as input and produces JSON files that will be the inputs to our classifiers. The goal here is to separate legal resources into single documents, along with all their related metadata such as ID, publication year, title etc. and their classification hierarchy (i.e., in which volume, chapter and subject they belong to). Each final JSON file represents a unique legal resource, ready to be fed into the machine-learning models we built. The parser builds in memory a tree of depth 4 that represents the whole GLC hierarchy. The first level consists of the thematic volumes while the second level contains all the thematic chapters for each volume. The third level includes the thematic subjects of the individual chapters, and finally, the leaf

---

[18]Available at: https://github.com/ankushshah89/python-docx2txt/.

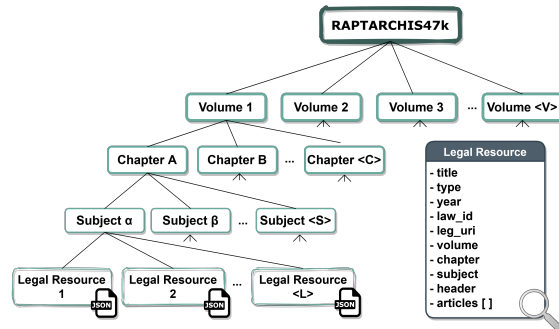nodes represent the legal resources. An overview of the tree can be found in Figure 4.

Figure 4: Tree representation of GLC thematic hierarchy

With the use of regular expressions, the parser organizes the content to its thematic hierarchy. Each legal resource may contain the whole original legislative document, some articles of it or even a short sentence (usually its original title or a short description). Hence, the parser attempts to identify and separate any existing articles. However, if this is not feasible, it just keeps the whole body as a text chunk. No deeper parsing is performed (i.e., in paragraphs, sentences) as this is out of scope. The final step is to populate the leaf nodes (i.e., the legal resources) with the appropriate metadata and enhance the available text samples. To accomplish this, the parser uses the metadata fragment of each legal resource to extract the necessary information. Specifically, the words of interest are shown in Figure 5, depicting an example of a metadata fragment.

Figure 5: Legal resource's metadata (translated)

Again, with proper regular expressions, the parser manages to retrieve the requested information. Also, having available the type, the year of publication and the ID of each legal resource, the parser is able to uniquely identify each one of them by using the following pattern: *{type}/{year}/{id}*. Exploiting that, it searches for duplicate legal resources that may exist in the dataset. For example,

one law may be present in more than one subject due to the thematic variety of its articles. To avoid any complexities and because our task is multi-class and not multi-label classification, the parser removes these resources entirely from the dataset.

Moreover, the parser manages to enhance the content of some legal resources (depending on their type[19]) by utilizing the platform Nomothesia[20] (Chalkidis et al., 2017b). Nomothesia makes Greek legislation available as open linked data using semantic web technologies. Through its RESTful API and by adopting the following URI template: *http://www.legislation.di.uoa.gr/eli/ {type}/{year}/{id}/data/json* the parser manages to retrieve the text of any legal resource in JSON format, as offered through Nomothesia. Then, it compares the number of tokens of the original and the fetched text fragments and eventually keeps the more extensive. In that way, the parser succeeds in enhancing the size and quality of the dataset.

For each final document, the text content consists of the header along with the body. In case of successful parsing the body consists of multiple articles. Otherwise, it only contains a single text passage. Evaluating the final data, we noticed that many legal resources have limited tokens count (as shown in charts of Section 3.2). However, we consider this not to be a crucial problem since meaningful information (e.g., highly representative words) is quite dense in most of these samples as shown in Figure 6.

```
{
  "title": "17. ΠΡΟΕΔΡΙΚΟ ΔΙΑΤΑΓΜΑ υπ´αριθ. 234",
  "type": "ΠΡΟΕΔΡΙΚΟ ΔΙΑΤΑΓΜΑ",
  "year": "1996",
  "law_id": "234",
  "leg_uri": "http://legislation.di.uoa.gr/eli/pd/1996/234",
  "volume": "ΒΙΟΜΗΧΑΝΙΚΗ ΝΟΜΟΘΕΣΙΑ",
  "chapter": "ΔΙΑΦΟΡΕΣ ΒΙΟΜΗΧΑΝΙΕΣ",
  "subject": "ΔΗΜΟΣΙΑ ΕΠΙΧΕΙΡΗΣΗ ΠΕΤΡΕΛΑΙΟΥ",
  "header": "Αύξηση κατά δεκαπέντε [...] Α.Ε.».",
  "articles": [
    "\nΕγκρίνεται η από 22 [...] δραχμών η κάθε μία.",
    "\nΤο Διάταγμα αυτό ισχύει από την [...] διατάγματος."
  ]
}
```

Figure 7: Final legal resource as JSON. The legal resource has been parsed and enhanced with two articles, as fetched from Nomothesia web platform

VOLUME (18) PRESS AND TOURISM
CHAPTER (B) TOURISM
SUBJECT (A) SCHOOL OF TOURISM PROFESSIONS
14. ROYAL DECREE **of 21 May / 26 June 1959**
On the imposition of penalties and the characterization of conduct of the students of the Higher Education of the School of Tourism Professions and graduation, etc.

Figure 6: Small-sized sample of GLC indicating highly representative words (translated)

Finally, the complete dataset, consisting of JSON files following the format of Figure 7 is distributed to train, development and test subsets as described in Section 3.2.

---

[19]See the supported legislation types at: http://legislation.di.uoa.gr/search/.

[20]See: http://legislation.di.uoa.gr/.