NATIONAL AND KAPODISTRIAN UNIVERSITY OF ATHENS

SCHOOL OF SCIENCES
DEPARTMENT OF INFORMATICS AND TELECOMMUNICATIONS

PROGRAM OF POSTGRADUATE STUDIES
”INFORMATION AND DATA MANAGEMENT”

MASTER’S THESIS

Legal Text Classification based on Greek Legislation

Christos N. Papaloukas

SUPERVISORS: **Manolis Koubarakis**, Professor
Ilias Chalkidis, PhD Candidate

ATHENS

DECEMBER 2020

ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
”ΔΙΑΧΕΙΡΙΣΗ ΠΛΗΡΟΦΟΡΙΑΣ ΚΑΙ ΔΕΔΟΜΕΝΩΝ”

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Ταξινόμηση Νομικού Κειμένου βασισμένου στην Ελληνική Νομοθεσία

Χρίστος Ν. Παπαλουκάς

ΕΠΙΒΛΕΠΟΝΤΕΣ: **Μανόλης Κουμπαράκης**, Καθηγητής
Ηλίας Χαλκίδης, Υποψήφιος Διδάκτωρ

ΑΘΗΝΑ

ΔΕΚΕΜΒΡΙΟΣ 2020

MASTER'S THESIS

Legal Text Classification based on Greek Legislation

Christos N. Papaloukas
ID: M1602

SUPERVISORS: **Manolis Koubarakis**, Professor
Ilias Chalkidis, PhD Candidate

EXAMINING COMMITTEE: **Manolis Koubarakis**, Professor
Yannis Panagakis, Associate Professor

December 2020

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Ταξινόμηση Νομικού Κειμένου βασισμένου στην Ελληνική Νομοθεσία

Χρίστος Ν. Παπαλουκάς
Α.Μ.: M1602

ΕΠΙΒΛΕΠΟΝΤΕΣ: **Μανόλης Κουμπαράκης**, Καθηγητής
Ηλίας Χαλκίδης, Υποψήφιος Διδάκτωρ

ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ: **Μανόλης Κουμπαράκης**, Καθηγητής
Ιωάννης Παναγάκης, Αναπληρωτής Καθηγητής

Δεκέμβριος 2020

ABSTRACT

Legal text processing is an emerging and significant field in Natural Language Processing, with its applications being valuable not only to legal professionals but also to society. Towards this end, in this study, we engage in legal text classification based on Greek legislation. We introduce and make publicly available a novel dataset based on Greek legislation, consisting of more than 47k official, classified Greek legislation resources. We experiment with and evaluate a battery of advanced methods and classifiers, varying from traditional machine learning and recurrent models to state-of-the-art transfer learning models. Results evince the shortcoming of traditional machine learning classifiers against more sophisticated methods, despite setting adequate baselines for most of the considered tasks. Recurrent architectures with domain-specific word embeddings offer improved overall performance while being competitive even to Transformer-based models. Still, cutting-edge multilingual and monolingual Transformer-based models brawl on the top of the classifiers' ranking, inducing us to 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 project of Greek legal text classification is researched to such an extent. We anticipate this study to be a strong groundwork, contributing significantly to the future research of the Greek NLP community.

SUBJECT AREA: Natural Language Processing, Machine Learning, Artificial Intelligence

KEYWORDS: Text Classification, Legal Data, Neural Networks

ΠΕΡΙΛΗΨΗ

Η επεξεργασία νομικού κειμένου είναι ένας αναδυόμενος και σημαντικός τομέας στην Επεξεργασία Φυσικής Γλώσσας, με τις εφαρμογές του να είναι πολύτιμες όχι μόνο για τους νομικούς επαγγελματίες αλλά και για την κοινωνία. Για το σκοπό αυτό, στην παρούσα μελέτη, ασχολούμαστε με την ταξινόμηση νομικού κειμένου βασισμένου στην ελληνική νομοθεσία. Παρουσιάζουμε και διαθέτουμε στο κοινό ένα νέο σύνολο δεδομένων βασισμένο στην ελληνική νομοθεσία, που αποτελείται από περισσότερους από 47 χιλιάδες επίσημους, ταξινομημένους πόρους ελληνικής νομοθεσίας. Πραγματοποιούμε πειράματα και αξιολογούμε μια σειρά από υπερσύγχρονες μεθόδους και ταξινομητές, που κυμαίνονται από παραδοσιακή μηχανική μάθηση και επαναλαμβανόμενα μοντέλα έως υπερσύγχρονα μοντέλα μεταφοράς μάθησης. Τα αποτελέσματα αποδεικνύουν την αδυναμία των παραδοσιακών ταξινομητών μηχανικής μάθησης έναντι πιο εξελιγμένων μεθόδων, παρά τον καθορισμό επαρκών βάσεων για τις περισσότερες από τις ερευνηθείσες εργασίες. Οι επαναλαμβανόμενες αρχιτεκτονικές με εξειδικευμένες διανυσματικές παραστάσεις λέξεων προσφέρουν βελτιωμένη συνολική απόδοση, ενώ είναι ανταγωνιστικές ακόμη και σε μοντέλα που βασίζονται σε μετασχηματιστές. Ωστόσο, τα υπερσύγχρονα πολυγλωσσικά και μονογλωσσικά μοντέλα που βασίζονται σε μετασχηματιστές φιλονικούν στην κορυφή της κατάταξης των ταξινομητών, προκαλώντας μας να αμφισβητήσουμε την αναγκαιότητα εκπαίδευσης μονογλωσσικών μοντέλων μεταφοράς μάθησης κατά κανόνα. Από όσα γνωρίζουμε, αυτή είναι η πρώτη φορά που το έργο της ταξινόμησης ελληνικού νομικού κειμένου ερευνάται σε τέτοιο βαθμό. Αναμένουμε ότι αυτή η μελέτη θα αποτελέσει μια ισχυρή βάση, συμβάλλοντας σημαντικά στη μελλοντική έρευνα της ελληνικής κοινότητας επεξεργασίας φυσικής γλώσσας.

ΘΕΜΑΤΙΚΗ ΠΕΡΙΟΧΗ: Επεξεργασία Φυσικής Γλώσσας, Μηχανική Μάθηση, Τεχνητή Νοημοσύνη

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ: Ταξινόμηση Κειμένου, Νομικά Δεδομένα, Νευρωνικά Δίκτυα

”Words can be like X-rays if you use them properly – they’ll go through anything. You read and you’re pierced.” — Aldous Huxley, Brave New World

ACKNOWLEDGEMENTS

First of all, I would like to thank my supervisor Prof. Manolis Koubarakis for giving me the chance to work on such an interesting problem. His courses and guidance throughout all my academic years are invaluable and have “reinforced” my knowledge and interest in AI. Working as a member of the AI Research Group for almost two years was really inspiring and engaging. Secondly I would like to thank my supervisor and friend Ilias Chalkidis. His support, advice and patience throughout this thesis and all my postgraduate years worth a king’s ransom, and have undoubtedly contributed to my knowledge in NLP. I hope we join our forces again in the future as his research work is remarkable and really influential. Last but not least, I want to thank my partner, my family and especially my grandfather for their unconditional support all these years.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

PREFACE

The present thesis is part of the requirements for the acquisition of a Master's degree "Information and Data Management" in the Department of Informatics and Telecommunications of the National and Kapodistrian University of Athens.

1. INTRODUCTION

In recent years, there has been intensified activity in the adaptation of Artificial Intelligence technologies to the legal domain in which legal practitioners are required to analyze and review an overwhelming number 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 through data rationally. 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), which have recently received a significant upsurge of interest.

With legal text processing being a flourishing field in NLP, many relevant applications have been derived such as legal information extraction [1], entity recognition [2, 3, 4], court opinion generation and analysis [5, 6], legal judgement prediction [7, 8, 9] and many more. Along these lines, the current work focuses on the task of multi-class legal text classification. Text classification (also known as text categorization or text tagging) is the task of assigning a set of predefined categories to free-text, and text classifiers can be used to organize, structure and categorize pretty much anything.

Towards these efforts, we engage in a more specific and relatively unexplored problem, the task of multi-class legal text classification based on Greek legislation documents. To the best of our knowledge, this is the first time this task is undertaken to such an extent, with experimental methods varying from traditional machine learning to state-of-the-art transfer learning models. All the experiments are conducted on a novel dataset consisting of Greek legislation resources, which is publicly offered for further use. The main contributions of this thesis are listed below:

- We study the task of multi-class text classification in Greek Legislation by applying and evaluating a battery of advanced methods and classifiers, varying from traditional machine learning and recurrent models, to state-of-the-art transfer learning models. We discuss the results and lay the groundwork for future research and applications.
- We generate and publicly offer `RAPTARCHIS47k`[1], a novel dataset based on Greek legislation. Its origin is “Raptarchis - Permanent Greek Legislation Code”, a collection of Greek legislative documents classified into thematic categories. An explicit rule-based parser was developed to produce a JSON oriented dataset consisting of 47k legal resources, propitious for further research on NLP tasks over Greek legal text.

The rest of the thesis is structured as follows: in chapter 2, we provide background information about the origin of our dataset and discuss the approaches of related work. In chapter 3, we describe the whole process of producing our novel dataset, present its

[1] Available at: https://github.com/christospi/gltc-raptarchis47k.

quantitative analysis and give a more detailed insight on the undertaken classification task. In chapter 4, we specify the experimental methods and their setup and ultimately, we present and evaluate the obtained results. Finally, in chapter 5, we reiterate through the main contributions of this thesis and propose a couple of ideas for future work and applications.

2. BACKGROUND AND RELATED WORK

Initially, in this chapter, we introduce and describe “Raptarchis - Permanent Greek Legislation Code”, the source of our legal text dataset. Further on, we provide a preview on the background and applications of legal text classification. Finally, we cite and discuss related work in the area of legal text classification of NLP, with techniques based on Machine and Deep Learning.

2.1 Raptarchis - Permanent Greek Legislation Code

“Raptarchis - Permanent Greek Legislation Code” (hereafter GLC) contains the Greek Legislation since the constitution of the Greek State in 1834 approximately until 2015. It includes laws, decrees, regulations and decisions with their respective amendments such as replacements, modifications and deletions while its only source of information is the Official Government Gazette. Today, all the legislation, together with the relevant indexes, is contained in 111 numbered volumes (documents) and is divided into 47 main thematic categories.

The founder and creator of GLC was Pantelis Raptarchis who maintained GLC as a private enterprise until 1978. In 1978, by Law 805/1978, GLC was donated to the State. Thus GLC, from 01/01/1979, fell into the hands of the State and particularly, the Ministry of Presidency (today, Ministry of Interior and Administrative Reconstruction). At the beginning of its ”public life”, the GLC management service was an independent division. It was directly under the General Secretary of the Ministry of the Presidency, now part of the Finance Directorate of the Directorate General for Administrative Support of the General Secretariat for Public Administration and e-Government, of the Ministry of Public Administration and Decentralization.

Currently, GLC is publicly offered through e-Themis portal[1], the legal database and management service of GLC, under the administration of the Ministry of the Interior. E-Themis is primarily focused on providing the most recent legislation on a multitude of predefined thematic categories, as described in GLC. More specifically, legislation is listed in one of the 47 main topics (Volumes) with increasing numbering (e.g., Constitutional Law 1.1A, Public Administration 2, 2A, 2B etc.). Each topic is divided into chapters and subjects, in which the legislations are numbered in complete chronological order. The total number of chapters is 389 while the total number of subjects is 2285.

Customers of the service are individuals, legal professionals, legal entities or organizations whose professional activity requires access to all in-force laws at the time, as well as every citizen that wants to be informed about Greek legislation in a multitude of predefined subject areas. The main advantages of GLC and e-Themis accordingly are the chronological and thematic categorization of legislation, which operates as a guiding manner,

[1] See http://e-themis.gov.gr.

making it easier for the reader to find the requested article in a short time and maintaining the full validity of legislation. The inner structure of GLC legal-documents collection provides an index of legal resources based on the thematic section that each one is related to, offering their readers a swift and efficient method for searching the Greek legislative knowledge.

Based on these data, we created and publicly offer a new dataset named `RAPTARCHIS47k`, containing all the available legislation from GLC in JSON format. Each JSON document contains a legal resource along with its metadata like thematic topics, publication year, type etc. as they extracted from the original GLC documents. In section 3.1.2, we discuss in-depth the processing method we followed to generate `RAPTARCHIS47k` and describe the data structure.

2.2 Related Work on Legal Text Classification

In the legal domain, continually increasing and extensive document collections require advanced text processing methods to retrieve, organize and search their textual content. An essential operation towards these efforts is legal text classification, which has become an emerging application in the NLP field.

Text classification is widely used in the legal domain, where a plethora of legal documents must be reviewed and analyzed for each particular task. By applying classification techniques in these documents, legal experts are able to perform comparative studies more efficiently, while they also manage to focus on the appropriate examination field more wisely. Moreover, by having available the document's domain, organizing and archiving become uncomplicated as well. The outcome of these advantages is that legal professionals achieve to execute essential tasks effortlessly, while they have more time to spend on complicated and brain-challenging work. As it becomes easier for the reader to find the requested resource in a short time, the classification of legal documents also operates in a guiding manner. Human rights organizations, legal scholars or even ordinary citizens become capable of quickly exploring legislative acts, legal code, legal decisions or anything else related to the legal area that interests them.

In research, plenty of publications that focus on text classification exist for more than twenty years, indicating that it is a widely investigated area in the NLP field. From decision trees and logic-based rules in early years to statistical classifiers as the naive Bayes model and support vector machines later on and recently, to neural-network-based classifiers with a plethora of different models derived from Recurrent Neural Networks (RNN) to Transformers. Looking into legal text classification — which is the centre of our attention — we realize its application's impact through the uprise of relative scientific work and the growing interest of NLP researchers to continuously achieve superior results in it, even if it is binary, multi-class or multi-label classification.

In prior groundwork [10], Nallapati and Manning investigated the problem of binary text classification in the legal domain and presented domain-specific problems where machine

learning classifiers such as SVMs were insufficient. Expressly, they point out the importance of better feature selection in such specialized domains and expose the limitations of advanced classifiers to capture the intricacies of natural language, especially for non-traditional domains such as the legal.

Frequently, related work in the addressed task is built upon datasets of court decisions and legal cases. One of them that explores the use of text classification in the legal domain is [11], which aims to predict the ruling of the French Supreme Court and the law area to which a case belongs, by combining the output of multiple SVM classifiers. The reported results mention ~96% average F1 score for predicting the law area (out of 8 different classes) and the ruling of the case by utilizing a system based on SVM ensembles. Another similar work is that of [12], in which the authors propose an attention-based neural network method to jointly model the charge prediction task and the relevant article extraction task in a unified framework, evaluated on criminal cases in China. In fact, the task of determining the appropriate charges for a given case complies with the multi-label classification paradigm. Despite the class distribution imbalances, the results showed that SVM-based models perform adequately when fact descriptions are the only given input. However, they are outperformed by more sophisticated neural-network architectures when the input also includes the relevant extracted articles. Both these efforts, along with many related others in bibliography [8, 9, 13], suggest that SVM classifiers are a strong baseline for text classification tasks.

A more recent study on classifying legal documents with neural networks is that of Undavia et al. [14], which deals with the problem of document classification of legal court opinions. The authors divide the task into two sub-tasks, depending on the number of total output classes: 15 broad and 279 finer-grained categories. Experimenting with different “vanilla” word embeddings (word2vec, fastText, GloVe and doc2vec) and shallow neural network models based on CNN, LSTM and GRU architectures, their best system (word2vec + CNN) achieves 72.4% accuracy in the 15-classes task and 31.9% accuracy in the 279-classes task. Once more, the baseline method of SVM with Bag-Of-Words features, reaches a noteworthy accuracy (64% and 30.5% respectively), proving to be a strong competitor to much more advanced models. The authors conclude by claiming that a fine-tuned GRU-based network together with domain-specific word embeddings could possibly complete the task with higher accuracy.

Another closely related work similar to [8] is [15], in which the authors tackle the problem of legal area classification, on a novel dataset of 6k judgements of the Singapore Supreme Court written in English, employing traditional statistical and state-of-the-art NLP models. To deal with data imbalance, they merged the 51 total different legal area labels into 31 final labels. Through a variety of benchmarked models and methods, including topic modelling, word embeddings and language modelling classifiers, the results showed that although data scarcity affects all the classifiers, a number of them and mostly pre-trained language models like BERT could perform pretty well even with a limited number of judgments. Wrapping up, the authors indicate the importance of having law-specific datasets and word embeddings as well as better methods to leverage transfer learning, in order to improve the adaptation of state-of-the-art NLP models for the legal domain.

A significant influence for the methods and models employed in our experiments comes from the research work of Ilias Chalkidis, Ion Androutsopoulos et al. on legal text classification based on European Court and European Legislation datasets. In [7], the authors evaluate a wide variety of neural models on a new dataset (consisting of European Court of Human Rights cases), establishing strong baselines that outperform prior feature-based models. The three main tasks of the evaluation are binary violation classification, multi-label classification and case importance prediction. Focusing on the multi-label violation classification task, the objective is to predict which (out of 66) specific human rights articles andor protocols have been violated. The results reveal that the developed Hierarchical-BERT model is close to the Hierarchical Attention Network (HAN) and surpasses BiGRU with self-attention and label-wise attention networks (LWANs; 60.0% vs 57.6% in micro-F1) which proved to be robust in multi-label classification [16]. Also, the authors point out the difficulty of few-shot learning (under-represented labels) in legal judgement prediction, through the poor performance of all the evaluated models.

In [17] I. Chalkidis, M. Fergadiotis et al. experiment with several neural classifiers on a novel dataset of 57k legislative documents (EUROLEX57k) from EUR-LEX[2]. They show that BiGRUs with self-attention outperform the current most advanced multi-label methods that employ label-wise attention. The use of domain-specific Word2Vec and context-sensitive ELMO embeddings improve the overall performance. Furthermore, the authors experiment with document zoning, considering only the title and recitals of each document for the classification task. Using a fine-tuned BERT-base model, they obtain the best results for all but zero-shot learning labels, establishing strong standards for LMTC on EURLEX57k. Still, few and zero-shot learning proves to be a challenging task that requires explicit handling.

A more recent and comprehensive version of this study comes in [18], in which the authors evaluate a battery of LMTC methods from pure LWANs to hierarchical classification and transfer learning approaches on three datasets from different domains. The experimental results show that hierarchical methods based on Probabilistic Label Trees (PLTs) 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. Finally, the case of few and zero-shot learning is tackled with the proposal of newer models that leverage the label hierarchy and yield better results.

As the approaches mentioned above prove to be quite efficient in text classification, specific models and methods applied in our experiments were built upon them. We investigate whether slightly modified versions of these methods can perform in multi-class classification just as well as they do in multi-label. In Chapter 4, we present a more technical view on the experimental setup and discuss the adjustments made on these models.

However, all the preceding efforts focus mostly on the English language. According to literature [19, 20, 21, 4, 22], NLP tasks focused on the Greek language are known to be more complicated and demanding due to the intricacies of the language itself. To the best

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

of our knowledge, this is the first time the task of Greek legal text classification is tackled to such an extent, as the experimental methods vary from traditional machine learning to state-of-the-art transfer learning models. We hope this study will offer a strong foundation and significant contribution to Greek NLP community for future work.

3. TASK DEFINITION

In this chapter, we describe the whole process of generating `RAPTARCHIS47k` dataset and demonstrate it through its quantitative analysis. Moreover, we give a more detailed insight into the undertaken task and discuss the emerging intricacies.

3.1 Generating a Novel Dataset on Greek Legislation

3.1.1 Original Documents Description

As already mentioned in section 2.1, “Raptarchis - Permanent Greek Legislation Code” contains classified Greek legislation from 1834 to 2015. Its primary goal was to create a serviceable and quick index of legal resources based on thematic sections rather than make a formal and strictly accurate representation of legislative knowledge. These data set up one of the main credible and publicly available sources of classified Greek legislation[1], suitable for this thesis.

As most of the input documents have a relatively standard structure, a rule-based parser is quite sufficient in order to generate a prosperous legal dataset. However, a different, more in-depth approach on parsing these legal documents and creating a more semantic dataset is presented in N. Mathioudakis thesis[2] but this is out-of-focus for our objective at the time being. So, the goal in this initial task is to separate each legal resource into a single document containing all its related metadata such as ID, publication year, title etc. along with its classification hierarchy (i.e. on which volume, chapter and subject it belongs to).

First, the structure of the legislative volumes (i.e. original documents) should be defined. Each legislative volume is related and structured according to a main thematic topic. Inside each volume, the main thematic topic is divided into thematic subcategories which are represented by chapters and subsequently, each chapter breaks down to subjects which contain the legal resources, creating an interlinked thematic chain. The performed classification experiments intend to predict the classes of this thematic chain.

In the text, volumes are being enumerated incrementally, followed by their title. Chapters are specified using capital letters of the Greek alphabet followed by their thematic title while subjects are specified with lower-case letters of the Greek alphabet also followed by their title. Finally, the legal resources contained in each subject are sorted chronologically in an incrementally enumerated list. In Figure 1, we can see the thematic hierarchy diagram of a legislative volume, while in the next figures, we can see the original format the documents follow.

[1] Another legitimate source is that of European Legislation written in modern Greek, see: https://eur-lex.europa.eu/browse/directories/legislation.html/.

[2] Available at: https://pergamos.lib.uoa.gr/uoa/dl/object/2899984/.

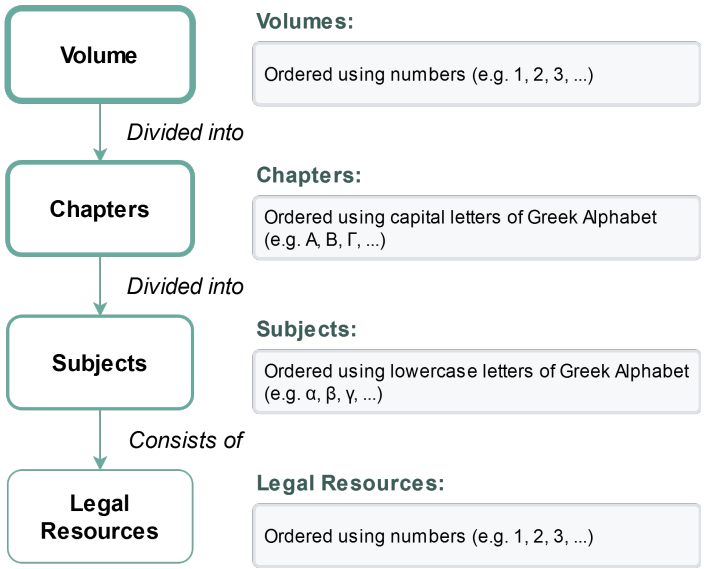
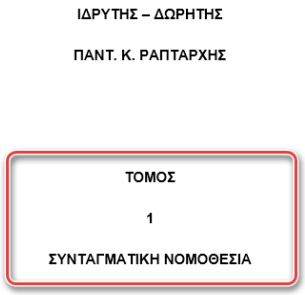


Figure 1: Original Raptarchis GLC thematic hierarchy


ΙΔΡΥΤΗΣ – ΔΩΡΗΤΗΣ

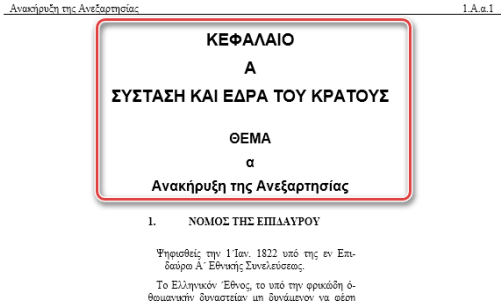
ΠΑΝΤ. Κ. ΡΑΠΤΑΡΧΗΣ

ΤΟΜΟΣ

1

ΣΥΝΤΑΓΜΑΤΙΚΗ ΝΟΜΟΘΕΣΙΑ

Figure 2: Example of Volume identification in original documents.


Ανακήρυξη της Ανεξαρτησίας 1.Α.α.1

ΚΕΦΑΛΑΙΟ

Α

ΣΥΣΤΑΣΗ ΚΑΙ ΕΔΡΑ ΤΟΥ ΚΡΑΤΟΥΣ

ΘΕΜΑ

α

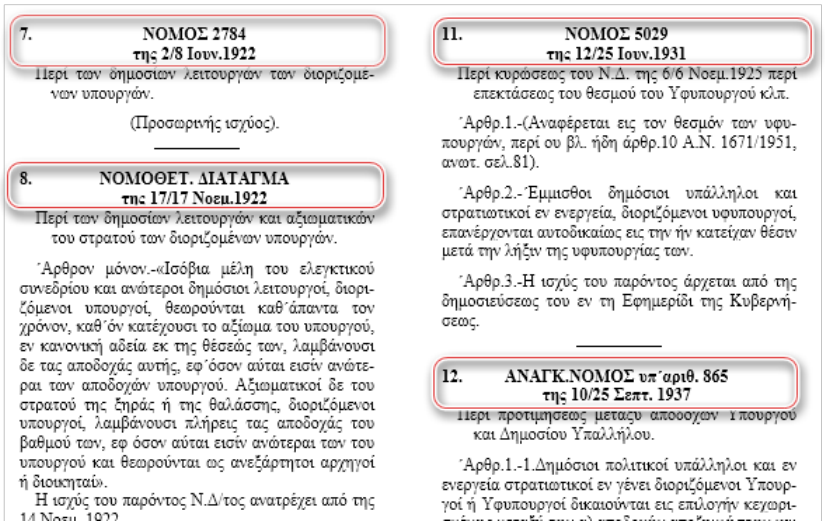
Ανακήρυξη της Ανεξαρτησίας

1. ΝΟΜΟΣ ΤΗΣ ΕΠΙΔΑΥΡΟΥ

Ψηφισθείς την 1 Ίαν. 1822 υπό της εν Επι-
δαύρω Α΄ Εθνικής Συνελεύσεως.

Το Ελληνικόν Έθνος, το υπό την φρικώδη ό-
θωμανικήν δυναστείαν μη δυνάμενον να φέση

Figure 3: Example of Chapter and Subject identification in original documents.


7. ΝΟΜΟΣ 2784
της 2/8 Ιουν.1922

Περί των δημοσίων λειτουργών των διοριζομέ-
νων υπουργών.

(Προσωρινής ισχύος).

8. ΝΟΜΟΘΕΤ. ΔΙΑΤΑΓΜΑ
της 17/17 Νοεμ.1922

Περί των δημοσίων λειτουργών και αξιωματικών του στρατού των διοριζομένων υπουργών.

΄Αρθρον μόνον.-«Ισόβια μέλη του ελεγκτικού συνεδρίου και ανώτεροι δημόσιοι λειτουργοί, διορι-
ζόμενοι υπουργοί, θεωρούνται καθ΄άπαντα τον χρόνον, καθ΄όν κατέχουσι το αξίωμα του υπουργού, εν κανονική αδεία εκ της θέσεώς των, λαμβάνουσι δε τας αποδοχάς αυτής, εφ΄όσον αύται εισίν ανώτε-
ραι των αποδοχών υπουργού. Αξιωματικοί δε του στρατού της ξηράς ή της θαλάσσης, διοριζόμενοι υπουργοί, λαμβάνουσι πλήρεις τας αποδοχάς του βαθμού των, εφ όσον αύται εισίν ανώτεραι των του υπουργού και θεωρούνται ως ανεξάρτητοι αρχηγοί ή διοικηταί».

Η ισχύς του παρόντος Ν.Δ/τος ανατρέχει από της

11. ΝΟΜΟΣ 5029
της 12/25 Ιουν.1931

Περί κυρώσεως του Ν.Δ. της 6/6 Νοεμ.1925 περί επεκτάσεως του θεσμού του Υφυπουργού κλπ.

΄Αρθρ.1.-(Αναφέρεται εις τον θεσμόν των υφυ-
πουργών, περί ου βλ. ήδη άρθρ.10 Α.Ν. 1671/1951, ανωτ. σελ.81).

΄Αρθρ.2.-΄Εμμισθοι δημόσιοι υπάλληλοι και στρατιωτικοί εν ενεργεία, διοριζόμενοι υφυπουργοί, επανέρχονται αυτοδικαίως εις την ήν κατείχαν θέσιν μετά την λήξιν της υφυπουργίας των.

΄Αρθρ.3.-Η ισχύς του παρόντος άρχεται από της δημοσιεύσεως του εν τη Εφημερίδι της Κυβερνή-
σεως.

12. ΑΝΑΓΚ.ΝΟΜΟΣ υπ΄αριθ. 865
της 10/25 Σεπτ. 1937

Περί προτιμήσεως μεταξύ αποσπασθέντων Υπουργού και Δημοσίου Υπαλλήλου.

΄Αρθρ.1.-1.Δημόσιοι πολιτικοί υπάλληλοι και εν ενεργεία στρατιωτικοί εν γένει διοριζόμενοι Υπουρ-
γοί ή Υφυπουργοί δικαιούνται εις επιλογήν κεχωρι-

Figure 4: Example of legal resources list in original documents.

3.1.2 Generating RAPTARCHIS47k - Parsing Procedure

As described, the original collection of GLC offered through e-Themis portal consists of 111 legislative volumes, each one encoded in .doc format (MS Word file 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[3], a python utility based on python-docx[4] that detects and extracts text from docx files.

Ideally, the structure we expected these generated text files to have according to the description, was the following:

Table 1: Ideal structure of transformed text documents

VOLUME <VOLUME ID> <VOLUME TITLE> **CHAPTER** <CHAPTER ID> <CHAPTER TITLE> **SUBJECT** <SUBJECT ID> <SUBJECT TITLE> [Legal Resource 1] [Legal Resource 2] [Legal Resource 3] **SUBJECT** <SUBJECT ID> <SUBJECT TITLE> [Legal Resource 4] [Legal Resource 5] [Legal Resource 6] ...

However, starting working with these text files, we encountered problematic samples that

[3]Available at: https://github.com/ankushshah89/python-docx2txt/.
[4]Available at: https://github.com/python-openxml/python-docx/.

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’s title etc., mostly due to minor inaccuracies in text conversion. Furthermore, multiple white spaces, multiple new lines and special or corrupted characters occurred in the text. To overcome these complications, we performed an essential clean-up using heuristics and regular expressions to produce neat text files following the same normalized structural pattern. A text sample at this point seems like this:

Table 2: Sample of a Volume text document

Original (Greek)	*Translated (English)*
ΤΟΜΟΣ 1 ΣΥΝΤΑΓΜΑΤΙΚΗ ΝΟΜΟΘΕΣΙΑ ΚΕΦΑΛΑΙΟ Α ΣΥΣΤΑΣΗ ΚΑΙ ΕΔΡΑ ΤΟΥ ΚΡΑΤΟΥΣ ΘΕΜΑ α Ανακήρυξη της Ανεξαρτησίας 1. ΝΟΜΟΣ ΤΗΣ ΕΠΙΔΑΥΡΟΥ Ψηφισθείς την 1΄Ιαν. 1822 υπό της εν Επιδαύρω Α΄ Εθνικής Συνελεύσεως. Το Ελληνικόν ΄Εθνος, το υπό την φρικώδη οθωμανικήν δυναστείαν μη δυνάμενον να φέρη [...]	VOLUME 1 CONSTITUTIONAL LEGISLATION CHAPTER A STATE ESTABLISHMENT AND HQ SUBJECT a Declaration of Independence 1. PREFECTURE OF EPIDAURUS Voted on Jan 1 1822 under the National Assembly in Epidaurus. The Greek Nation, under the horrible Ottoman dynasty unable to bring [...]

Next, we built a parser module in Python, able to receive these text files as input and produce an object-oriented and robust dataset of JSON files, optimal for our classification experiments. Each final JSON file represents a unique legal resource, ready to be fed into the machine-learning models we built.

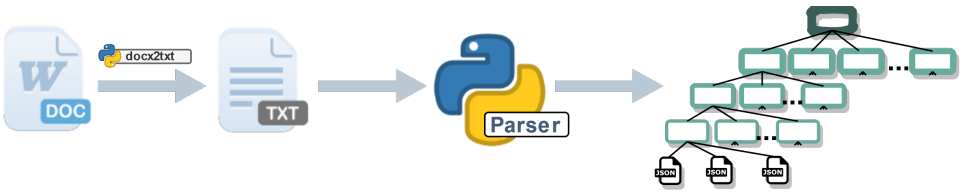


Figure 5: Parser module workflow

The parser module builds in memory a tree of depth four (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 nodes represent the legal resources. An overview of the tree can be found at Figure 6.

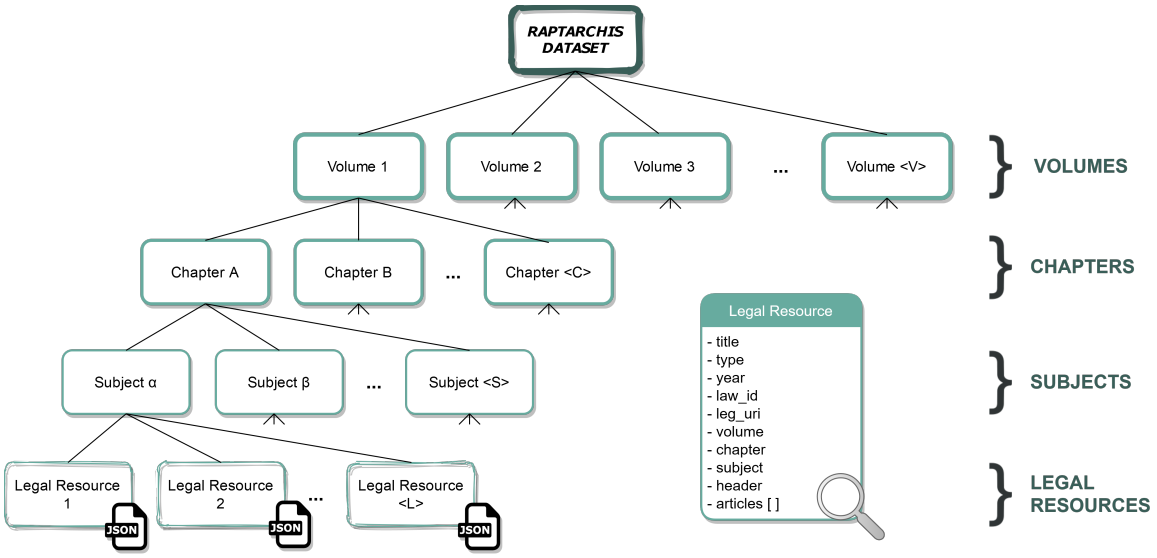


Figure 6: Tree representation of RAPTARCHIS47k thematic hierarchy

To achieve that, the parser iterates over the collection of text files and sequentially reads each one, building in that way the tree model. With the use of regular expressions, it divides the context into volumes, chapters and subsequently each chapter into thematic subjects. After gathering all the information about the thematic order up to this point, it proceeds on extracting each legal resource (under the specific thematic subdivision) by identifying the beginning and the end of it. The content of each legal resource is defined as the starting point of the numbered fragment that contains the resource's metadata and the ending point as the next sequential metadata section (as shown on Figure 4). Each legal resource may contain the whole original legislative document, parts of it (e.g. in most cases some of its articles) or even a short sentence of it (usually its original title or a short description). Hence, the parser attempts to identify it and separate its possibly existing articles. However, if this is not feasible, it just keeps its whole body as a text chunk. No deeper parsing is performed (i.e. in paragraphs, sentences) as this is out of scope.

After iterating over all the text files, the entire tree has been built. The main goal of identifying and extracting each legal resource and its metadata fragment, along with the whole thematic chain it belongs to, has been achieved. The final step is to populate the leaf nodes (i.e., the legal resources) with the appropriate metadata and enhance the available text samples. For this to be accomplished, the parser uses the metadata fragment of each legal resource to extract the necessary information. Specifically, the words of interest are shown in Figure 7, depicting an example of a metadata fragment.

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

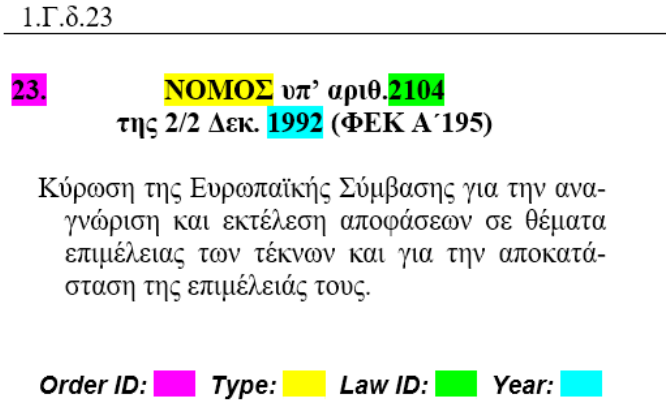


Figure 7: Legal resource's metadata identification

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[5]) by utilizing Nomothesia[6] [23], a platform built by our research group[7]. Nomothesia makes Greek legislation easily accessible to the public by offering government data 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.

Evaluating the final data, we noticed that many legal resources have limited tokens count (as shown in charts of the next section). 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. On the following figure, we can observe this claim:

[5]See the supported legislation types at: http://legislation.di.uoa.gr/search/.
[6]See: http://legislation.di.uoa.gr/.
[7]See: http://ai.di.uoa.gr/.

ΤΟΜΟΣ (4) ΑΣΤΥΝΟΜΙΚΗ ΝΟΜΟΘΕΣΙΑ	VOLUME (4) POLICE LEGISLATION
ΚΕΦΑΛΑΙΟ (ΑΑ) ΕΛΛΗΝΙΚΗ ΑΣΤΥΝΟΜΙΑ	CHAPTER (AA) HELLENIC POLICE
ΘΕΜΑ (Α) ΟΡΓΑΝΙΚΕΣ ΔΙΑΤΑΞΕΙΣ	SUBJECT (A) ORGANIC PROVISIONS
32. ΑΠΟΦΑΣΗ ΑΡΧΗΓΟΥ ΕΛΛΗΝΙΚΗΣ ΑΣΤΥΝΟΜΙΑΣ Αριθ. 7001/4/12-ρλστ΄ 23 Μαρτ. - 1 Απρ. 1998 (ΦΕΚ Β΄ 316) Καθορισμός τοπικής αρμοδιότητας των Ειδικών Αστυνομικών Υπηρεσιών Δίωξης Εγκληματικότητας Ηρακλείου, Ρεθύμνης και Χανίων.	**32. DECISION OF THE CHIEF OF GREEK POLICE No. 7001/4/12-ρλστ΄ March 23 - April 1, 1998 (Government Gazette B' 316)** Determination of local competence of the Special Police Services for the Prosecution of Crime in Heraklion, Rethymno and Chania.

ΤΟΜΟΣ (18) ΤΥΠΟΣ ΚΑΙ ΤΟΥΡΙΣΜΟΣ	VOLUME (18) PRESS AND TOURISM
ΚΕΦΑΛΑΙΟ (Β) ΤΟΥΡΙΣΜΟΣ	CHAPTER (B) TOURISM
ΘΕΜΑ (Α) ΣΧΟΛΗ ΤΟΥΡΙΣΤΙΚΩΝ ΕΠΑΓΓΕΛΜΑΤΩΝ	SUBJECT (A) SCHOOL OF TOURISM PROFESSIONS
14. ΒΑΣΙΛΙΚΟΝ ΔΙΑΤΑΓΜΑ της 21 Μαΐου/26 Ιουν. 1959 Περί επιβολής ποινών και χαρακτηρισμού διαγωγής των σπουδαστών της Ανωτέρας Εκπαιδεύσεως της Σχολής Τουριστικών Επαγγελμάτων και διαβαθμίσεως κλπ.	**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 8: Small-sized samples of RAPTARCHIS47k (original:left and translated:right) indicating highly representative words

Conclusively, the complete dataset, consisting of JSON files following the format of Figure 10, is stored in a hierarchical directory structure as earlier described in Figure 6. On the final step, the parser iterates and distributes the JSON files to split the dataset into train, development and test subsets.

17. ΠΡΟΕΔΡΙΚΟ ΔΙΑΤΑΓΜΑ υπ΄αριθ. 234 της 26/31 Ιουλ. 1996 (ΦΕΚ Α΄177) Αύξηση κατά δεκαπέντε δισεκατομμύρια (15.000.000.000) δραχμές του μετοχικού κεφαλαίου της «ΔΗΜΟΣΙΑΣ ΕΠΙΧΕΙΡΗΣΕΩΣ ΠΕΤΡΕΛΑΙΟΥ Α.Ε.».	17. PRESIDENTIAL DECREE no. 234 of 26/31 July 1996 (Government Gazette A' 177) Fifteen billion increase (15,000,000,000) drachmas of the share capital of "PUBLIC PETROLEUM ENTERPRISE S.A.".

Figure 9: Original legal resource sample (left) and its translation to English (right)

```
{
  "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 δισεκατομμύρια [...] Α.Ε.».",
  "articles": [
    "\nΕγκρίνεται η από 22 Απριλίου 1996 απόφαση [...] δραχμών η κάθε μία.",
    "\nΤο Διάταγμα αυτό ισχύει από την δημοσίευσή [...] διατάγματος."
  ]
}
```


Figure 10: Transformation of the previous sample to JSON format. The legal resource has been parsed and enhanced with two articles as fetched from Nomothesia web-platform

3.1.3 RAPTARCHIS47k in numbers

In this section, we present a more detailed and quantitative analysis of the dataset. To begin with, the final dataset consists of *47563* JSON files (i.e., classified legal resources). Initially, the total number of legal resources the parser managed to identify was *53052*. Out of the total, *1873* were interlinked with existing legal resources from Nomothesia, and *857* of them were successfully enhanced with text fragments from the web-platform. Eventually, *5489* were totally removed mostly due to multiple appearances as previously described in section 3.1.2 or due to their incomplete or almost empty text fragment.

`RAPTARCHIS47k` is split into three distributed subsets in the following chunks: train (60%), development (20%) and test (20%) as shown in Table 3. Distribution was performed consistently to all the levels of class hierarchy in order to achieve the same level of partitioning from bottom to top (i.e., from each subject to the whole dataset).

Table 3: Total number of documents and tokens(words) per document data

Subset	Documents (D)	Mean of Words/D	Low-in-words D (<100 words)
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%)

For each document, the text content consists of the header along with the body, either the body consists of multiple articles or just a single text passage. Evaluating the dataset, more than half of it consists of documents with no more than 100 words. However, the mean of words per document is almost 600, indicating that a few extensive documents are also included. This matter becomes evident in the following chart:

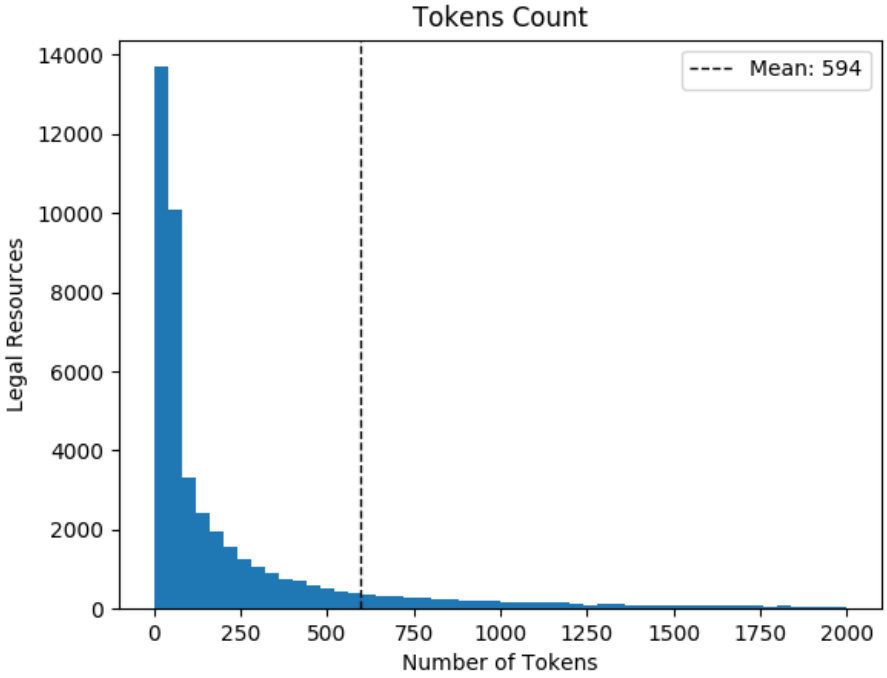


Figure 11: RAPTARCHIS47k tokens chart. The vertical line shows the mean token-size (594 tokens) of documents

Table 4: Documents size per dataset's percentage. Maximum and mean size is reported

DATASET (10%): DOCS: 4756 LENGTH: MAX: 26 MEAN: 21.40
DATASET (20%): DOCS: 9512 LENGTH: MAX: 33 MEAN: 25.30
DATASET (30%): DOCS: 14268 LENGTH: MAX: 41 MEAN: 29.00
DATASET (40%): DOCS: 19025 LENGTH: MAX: 53 MEAN: 33.27
DATASET (50%): DOCS: 23781 LENGTH: MAX: 79 MEAN: 39.41
DATASET (60%): DOCS: 28537 LENGTH: MAX: 142 MEAN: 50.76
DATASET (70%): DOCS: 33294 LENGTH: MAX: 246 MEAN: 70.52
DATASET (80%): DOCS: 38050 LENGTH: MAX: 464 MEAN: 104.24
DATASET (85%): DOCS: 40428 LENGTH: MAX: 690 MEAN: 131.39
DATASET (90%): DOCS: 42806 LENGTH: MAX: 1132 MEAN: 173.26
DATASET (95%): DOCS: 45184 LENGTH: MAX: 2363 MEAN: 249.75
DATASET (100%): DOCS: 47563 LENGTH: MAX: 146309 MEAN: 594.23

Around 80% of the total documents have a mean of 100 tokens/words, proving that only

the 5-10% of documents consist of lengthy text segments.

RAPTARCHIS47k 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, test); few-shot classes which appear in 1 to 10 training documents and zero-shot classes which appear in the development and/or test, but not in the training documents. As shown on Table 5, many classes are under-represented, especially in the thematic level of subjects, causing the appearance of few and zero-shot samples.

Table 5: Total number of Classes per thematic level and their distribution to frequent-few-zero categories

	Total classes	Frequent	Few-shot (<10 occ.)	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%)

The appearance of underrepresented classes — especially in Subject level — makes the dataset also appropriate for few- and zero-shot learning experiments. Focusing more on frequent, few- and zero-shot samples, the following table shows the total number of documents per category:

Table 6: Total number of Documents per thematic level and their distribution to frequent-few-zero categories

	Total Docs.	Frequent	Few-shot (<10 occ.)	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%)

In volume-level, all the classes belong to the frequent category, as more than 10 documents per class exist in the training data. In chapter-level, things are getting more interesting, as 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. Subsequently, in subject-level, data are even more unequally distributed over classes as 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.

Focusing more on the legal resources' distribution over classes in all the thematic levels, the following charts illustrate the number of documents per cluster of classes in Volume, Chapter and Subject levels:

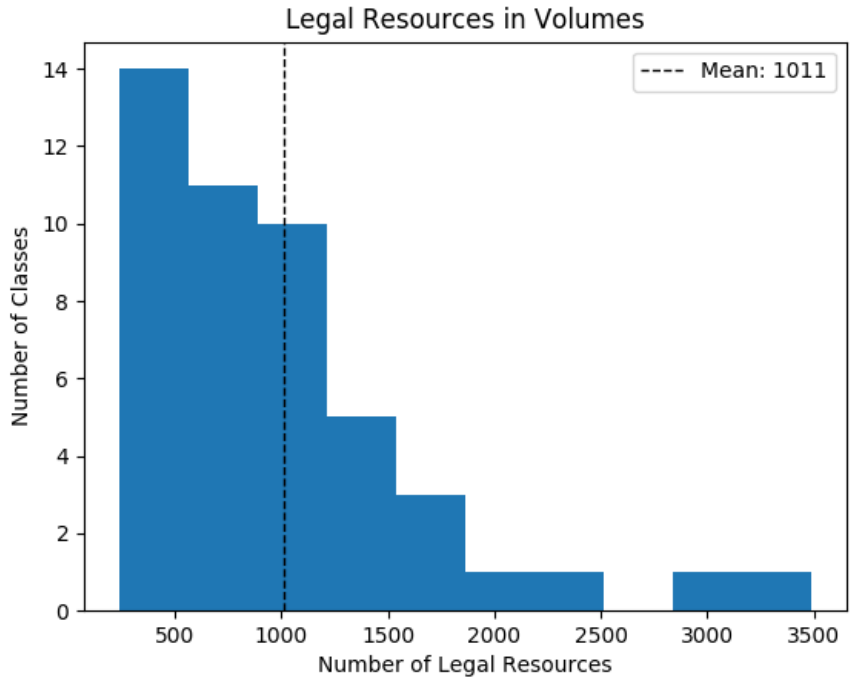


Figure 12: Number of legal resources per cluster of classes in Volume level. The mean of documents per class is also reported

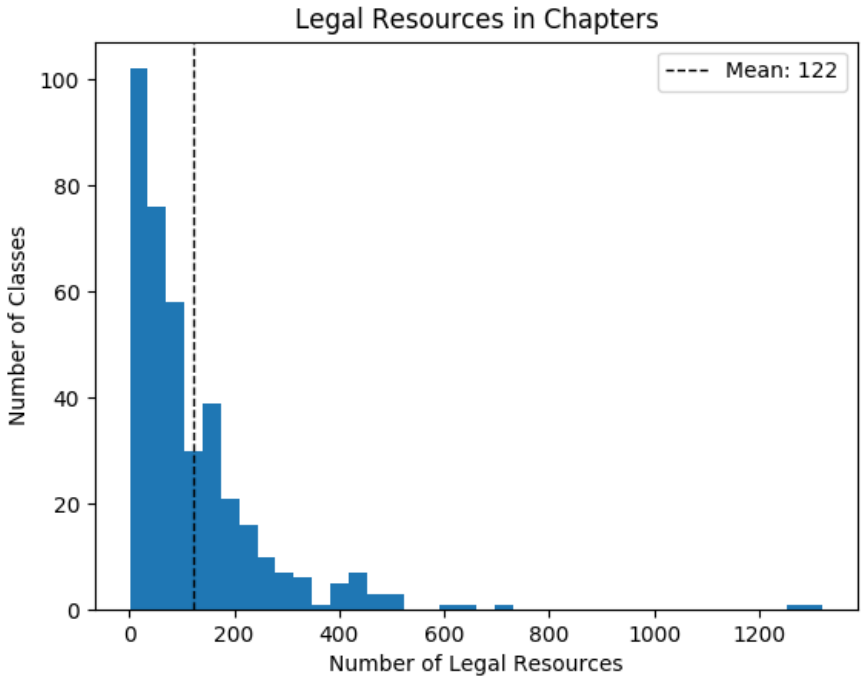


Figure 13: Number of legal resources per cluster of classes in Chapter level. The mean of documents per class is also reported

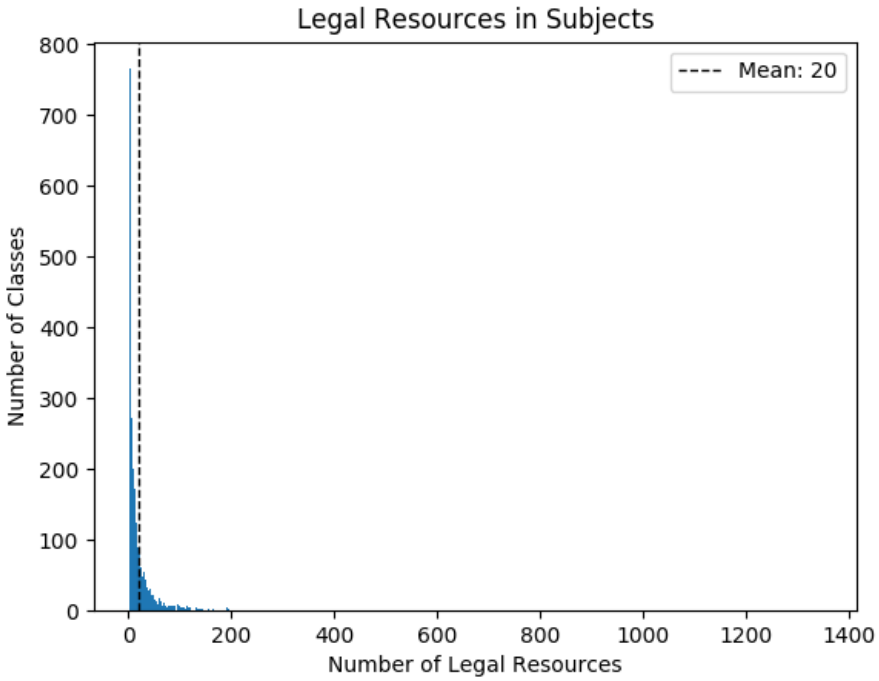


Figure 14: Number of legal resources per cluster of classes in Subject level. The mean of documents per class is also reported

Moving from the broader thematic level of volumes to the most specific level of subjects, it is becoming noticeable that the data follow an aggressive Zipfian distribution. This has been noted again in the legal domain [24] where Extreme Multi-Label Text Classification has been applied to label legal documents with concepts from Eurovoc vocabulary, as well as in other domains, like medical examinations [25] where LMTC has been applied to index documents with concepts from medical thesauri.

Finally, another interesting fact about `RAPTARCHIS47k` is that although the constituent legal resources have a nominal year range from 1834 to 2015, the most of them are published between 1960-2000 (Figure 15). The early years of this range disclose the political changeover that happened in Greece, with the disgraceful rise of the far-right military junta that ruled Greece from 1967 to 1974. During and after this period, legislation in Greece changed radically, indicating the transformation of the State from dictatorial to democratic. As for the later years, the decrease in legal resources implies the end of the actual support of Raptarchis project.

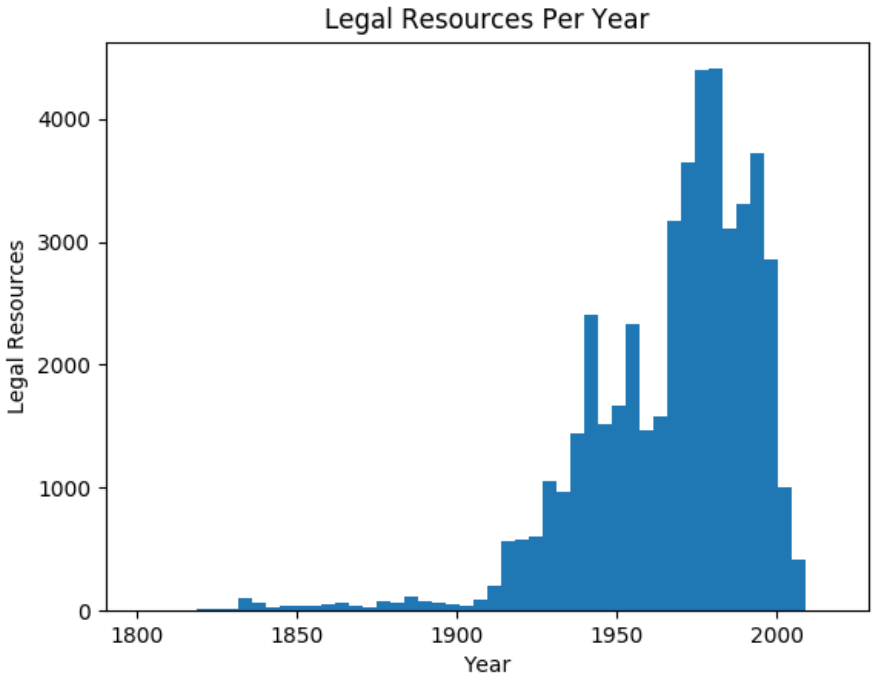


Figure 15: Number of legal resources per year in RAPTARCHIS47k

3.2 Multi-class Legal Text Classification based on Greek Legislation

Multi-class text classification is the task of classifying documents into one out of three or more classes while assuming that each sample is assigned to one and only one label; e.g., a document may talk about philosophy or mathematics but not both at the same time (unless Plato[8] authored the document). Still, multi-class classification should not be confused with multi-label classification, where multiple labels are to be predicted for each instance. In this work, we study the application of multi-class classification on Greek legislation.

As previously described, `RAPTARCHIS47k` offers 3 hierarchical levels of thematic categorization, which, for the purpose of the experiments, are labelled as:

- **Volume:** it is the first and broader level of the thematic categorization. It consists of 47 different classes and is divided into Chapters.
- **Chapter:** it is the second level of the thematic categorization. It consists of 389 different classes and is divided into Subjects.
- **Subject:** it is the third, final and more specialized level of the thematic categorization. It consists of 2285 different classes.

[8] https://www.britannica.com/biography/Plato.

Henceforth and to make it clear, when we refer to any of these labels, we do so to indicate the thematic level of categorization rather than the structural form of the dataset (as described earlier in subsection 3.1.1). On the following table we present an example of a volume-level class along with its subdivisions:

Table 7: ”Criminal Legislation” Volume along with Chapter and Subject subdivisions

Volume	Chapters	Subjects
Criminal Legislation	Criminal Law	Penal Code
	Special Criminal Laws	Betrayal, Animal Theft And Animal Killing, Robbery, Drugs, Fleeing From Justice, False Certificate Before The Authority, Various Offenses
	International Criminal Law	Counterfeiting, Women’s Trafficking, Bondage, Genocide, Protection Of Internationally Protected Persons, Terrorism, Organized Crime and Drug Trafficking, Corruption Of Foreign Public Officers, Offenses Committed In The Member States Of The European Union
	Extraordinary Criminal Laws	Crimes Against Public Security
	Military Criminal Legislation	Military Criminal Code, Criminal Proceedings, Legal Remedies, Execution Of The Death Penalty
	Special Military Offenses	Desertion
	Military Court	Permanent Military Court, Extraordinary Military Court, Military Courts Regulation, Judgment Of The Armed Forces, Secretaries Of Military Justice, Military Prisons
	Marine Criminal Law	Naval Criminal Legislation, Judicial Consultants Of The Navy, Admiralty Court Rules
	Aviation Criminal Law	Air Court

For each legal resource, the classifier constructs a representative text chunk on which it

is trained and seeks to predict the proper class. In case of successful parsing, this text chunk includes the header and all the articles concatenated. Otherwise, only the header is considered (which also contains the body text passage).

Despite the hierarchical formation of the thematic categorization of our dataset, we chose to address the classification problem in flat logic. That is to say, for each classification prediction, the classifier does not take into account the thematic hierarchy. Towards this approach, the classification task is subdivided into three separate tasks. Each task engages in a different level of the thematic hierarchy (i.e. in volumes, chapters, or subjects) where the classifier attempts to predict the correct class out of all the classes in this specific thematic level.

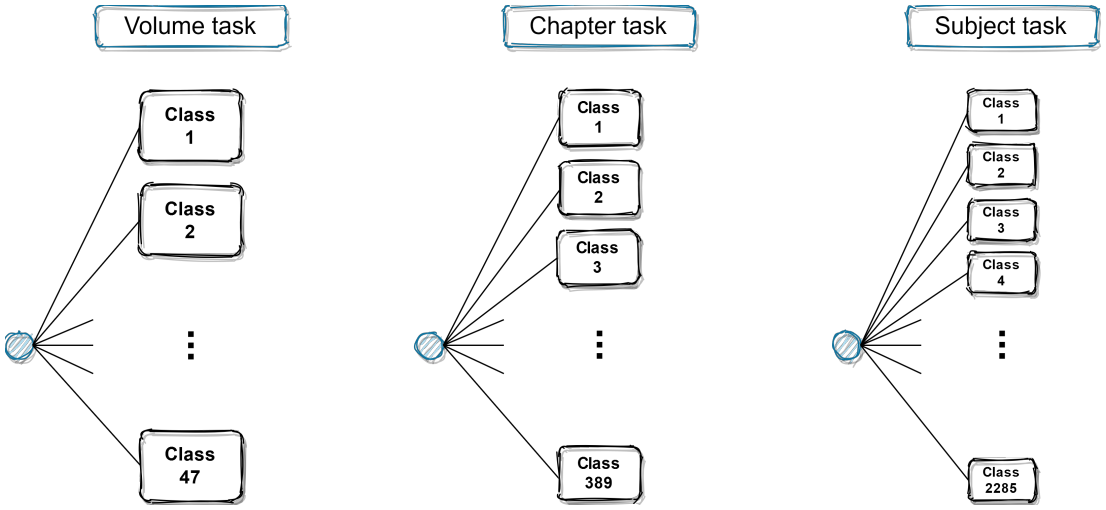


Figure 16: The main classification task is divided into (3) different sub-tasks, one per thematic level

Examining the quantitative data about the dataset in section 3.1.3, we notice that each thematic level has different samples distribution over its classes. To recap, in volume-level, all the classes belong to the frequent category, meaning that each class is sufficiently represented. In chapter-level we see that some of its classes are underrepresented (13.6%), while in subject-level that percentage reaches an excessive rate of 62.6%, along with 142 classes having zero representation at the training process.

At this point, we thought of shaping the dataset differently and more conveniently. As [15] suggests, we could have chosen to merge few- and zero-shot classes into a single, more generic “other” class. Another option would be to limit our experiments only in frequent classes or entirely remove zero-shot classes and samples. However, we chose not to perform any of these alternatives. Intending to offer a baseline in Greek legal text classification — and especially in this particular dataset — we prefer the experiments to be performed on the “vanilla” version of `RAPTARCHIS47k`, without any transformation. Therefore, except for the different number of classes, the partitioning of the main task also offers an interesting look at how the classifiers perform in setups with diverse and intricate data distribution. The evaluation and review of the applied methods will offer a strong basis for future study and more sophisticated experiments, as they proposed in chapter 5.

3.3 Few-shot and Zero-shot Learning Approaches

As earlier described, the appearance of underrepresented classes makes RAPTARCHIS47k appropriate for few- and zero-shot learning experiments. Usually, the methods employed in these experiments are fine-grained to deal with this data complication.

In the relevant task of multi-label text classification, a significant breakthrough has been made towards this effort. In [25], the authors discuss the challenge of few-shot and zero-shot learning over the MIMIC datasets. They achieve promising results investigating the effect of encoding the hierarchy in these settings. Despite their proposed sophisticated attention-based architectures, other more refined factors like alternative hierarchy encodings and deeper neural networks were not examined. In [26] Zikun Hu et al. focus on the task of charge prediction in few-shot and confounding charges, introducing discriminative legal attributes into consideration. To that end, they propose a novel attribute-based multitask learning model to predict those charges.

Along the same line, in [17] Chalkidis et al. also tackle the problem of few and zero-shot learning in large scale multi-label text classification but without taking into consideration the label hierarchy. Instead, in [18], they present a more refined and extended work. By utilizing multiple datasets, the authors methodically compare flat, PLT-based and hierarchy-aware LMTC methods in few and zero-shot learning multi-label classification, as they also explore the effect of transfer learning in that task for the first time.

The approaches mentioned above are sound and well-reported, motivating us to apply them in our task. Indeed, for this to happen, a few adjustments should be made to adapt these methods into carrying out the task of multi-class classification. Nevertheless, at the time being, none of these methods was considered in our study for two reasons. Primarily, since most of the conducted experiments have adequate sample support and also, because our main objective was to examine how the most prevailing methods perform in this particular task. Thus, we only report results in frequent-shot and few-shot categories. Future support of these approaches is undoubtedly considered as a subsequent task.

4. EXPERIMENTS

To begin with, in this chapter, we describe the methods employed in our experiments, ranging from traditional machine learning to state-of-the-art transfer-learning models. Later on, we analyze the experimental setup and finally, we present the experimental evaluation through the demonstration and interpretation of the obtained results.

4.1 Methods

4.1.1 Support Vector Machines with Bag-of-Words Features

The lead-off method that puts forward a strong baseline for the multi-class classification problem as it is one of the highest performing traditional ML methods is Support Vector Machines (SVM). SVM is able to learn non-linear separations between classes as its kernel function takes a low-dimensional input space and transforms it into a higher-dimensional space, converting non-separable problems to separable. Its objective is to search for an optimal hyperplane in a higher-dimensional space, which is capable of separating the classes in the data by a maximum possible margin.

For this implementation, 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. We apply an SVM classifier using Scikit-learn's SVM package and experiment with most of its hyperparameters.

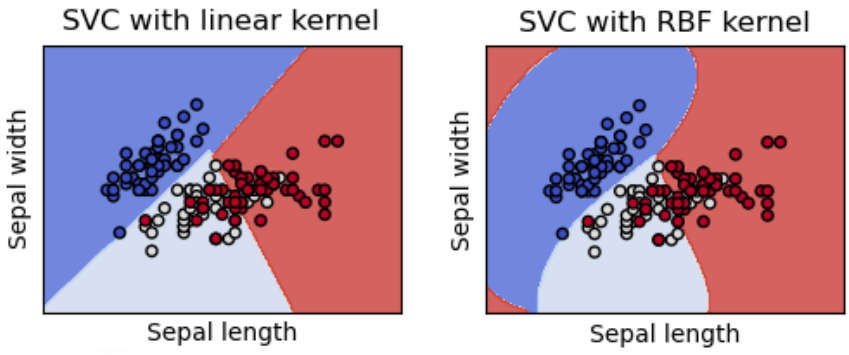


Figure 17: 2D projection of toy dataset (about sepals), showcasing Linear and RBF kernels. We get an intuitive understanding of their respective expressive power, with linear vs flexible non-linear decision boundaries

4.1.2 XGBoost with Bag-of-Words Features

The next model utilized for our experiments is XGBoost [27] (eXtreme Gradient Boosting), a scalable, fast and robust open-source implementation[1] of the Gradient Boosting decision tree algorithm. Since its introduction, this algorithm has been credited with winning numerous Kaggle competitions while it also composes the backbone of several pioneering applications.

First, let us explain some of its basic concepts. Boosting is an ensemble method, seeking to create a robust classifier based on weak classifiers. In this context, robust and weak are references to a measure of how correlated are the learners to the actual target variable. By adding models on top of each other iteratively, errors of the previous model are corrected by the next predictor, until the model accurately predicts the training data. Subsequently, Gradient Boosting also comprises an ensemble method that sequentially adds predictors and corrects previous models. However, instead of assigning different weights to the classifiers after every iteration, this method fits the new model to new residuals of the previous prediction. It then minimizes the loss when adding the latest prediction. Eventually, the model is being updated using gradient descent and hence the name, gradient boosting. This method supports both regression and classification problems.

XGBoost implements this algorithm explicitly for decision tree boosting with an additional custom regularization term in the objective function. So, we can say that in XGBoost, the model is fitted on the gradient of loss generated from the previous step and the gradient boosting algorithm is modified so that it works with any differentiable loss function. As for our implementation, the documents are represented using BoW features (similarly to SVM approach), and we explore the fine-tuning of its more radical hyperparameters.

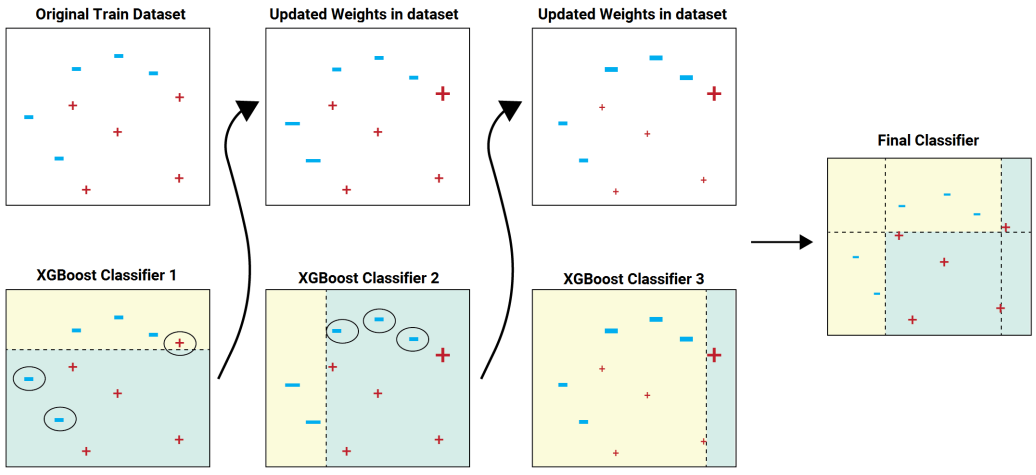


Figure 18: XGBoost classifier training process. Each sub-classifier learns from previous errors and re-adjusts the weights. This process continues until we have a combined final classifier that predicts all the data points correctly

[1]Available at: https://github.com/dmlc/xgboost/

4.1.3 BiGRUs - MaxPooling with Word2Vec Embeddings

The first neural, recurrent method is a BiGRU with Max-Pooling, employing pre-trained word embeddings. These word embeddings originate from a domain-specific Word2Vec [28, 29] model, as described in [4]. This specific legal Word2Vec model was trained on a corpus of over 200k Greek legislative and legal documents, producing 100-dimensional word embeddings for a vocabulary of 428,963 words (types), based on 615 millions of tokens (words). Both methods described in subsections 4.1.4 and 4.1.5 also utilize these pre-trained word embeddings.

Following, the word encoder that converts the pre-trained word embeddings (w_i) into context-aware embeddings (h_i) is a stack of BiGRUs. BiGRU is a sequence processing model that consists of two GRUs, one taking the input in a forward direction and the other in a backwards direction. In fact, it is a bidirectional recurrent neural network with only the update and reset gates. Directly after, each context-aware token embedding is passed on a max-pooling layer. Max-pooling is deployed over the complete output of BiGRUs layer to produce the final representation (d) of the document, providing a fixed size output layer. The aim is to reduce the output dimensionality but hopefully, keep the most salient information.

Finally, a dense layer with (L) output units and $softmax()$ activations is deployed, to transform the document representation (d) into a probability distribution over (L) classes and predict the proper one. The (L) parameter is defined according to the attempted task, as described in section 3.2: L=47 for the Volume task, L=389 for the Chapter task and L=2285 for the Subject task.

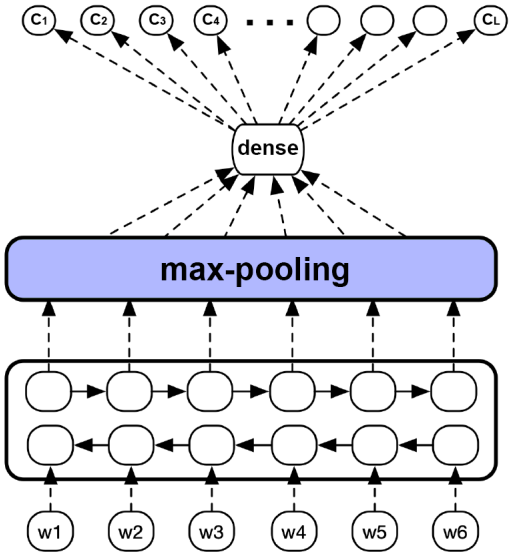


Figure 19: Architecture of BiGRUs - MaxPooling with Word2Vec Embeddings

4.1.4 BiGRUs - Self-Attention with Word2Vec Embeddings

The second recurrent method is a BiGRU with self-attention [17, 30], using the aforementioned Greek legal word embeddings through the Word2Vec model. Initially, each document is represented as the sequence of its word embeddings (w_i), which subsequently go through a stack of BiGRUs that converts them into context-aware ones (h_i).

Next, a self-attention mechanism is employed by the document encoder to produce a final, solid representation (d) of the document. This representation is computed as the sum of the resulting context-aware embeddings (h_i), weighted by the self-attention scores (a_i):

$$a_i = \frac{\exp\left(h_i^\top u\right)}{\sum_j \exp\left(h_j^\top u\right)} \tag{4..1}$$

$$d = \frac{1}{T}\sum_{i=1}^{T} a_i h_i \tag{4..2}$$

In the above formulas, (T) represents the document's length in words while (u) is a trainable vector used to compute the attention scores (a_i) over (h_i). Intuitively, this method gives even more contextual information to the embeddings as it enables the model to learn the correlation between the words in each document. This feature empowers a word embedding to receive enough attention that will make an observable change in its embedding and consequently, to the whole document's representation.

Eventually and similarly to 4.1.3, a dense layer with (L) output units and $softmax()$ activations is deployed to predict the appropriate output class using a probability distribution over all the classes.

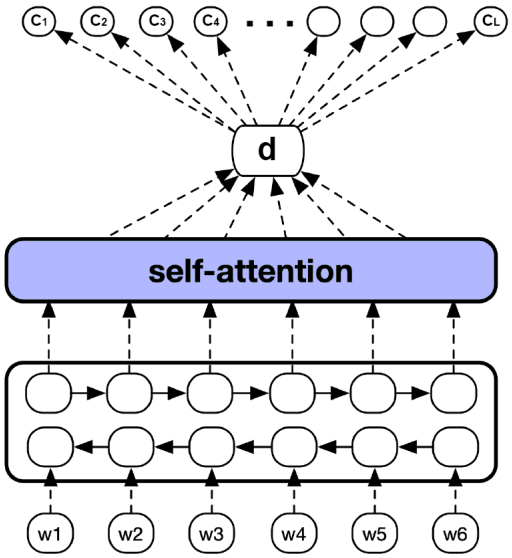


Figure 20: Architecture of BiGRUs - Self-Attention with Word2Vec Embeddings

4.1.5 BiGRUs - Label-Wise Attention Network with Word2Vec Embeddings

The third and final method based on BiGRUs and Word2Vec embeddings also employs a self-attention mechanism, but this time with a different, label-wise (or class-wise) approach. The original Label-Wise attention network was introduced by Mullenbach et al. [16], and according to it, the word embeddings of each document are initially converted to a sequence of vectors (h_i) by a CNN encoder. Guided by this effort, Chalkidis et al. [17, 24] replaced the CNN encoder with a BiGRU to produce context-sensitive embeddings (h_i), similarly to the method described in subsection 4.1.4.

In contrast with BiGRU with a self-attention network, this label-wise attention technique uses (L) independent attention heads, one per class, generating (L) document representations (d_l) from the sequence of (h_i) vectors produced by the BiGRU encoder. The intuition is that each final document embedding (i.e. each attention-head d_l) is dedicated to predicting the corresponding class, focusing on possibly different aspects of each representation (h_i). In effect, different parts of the base representation may be more relevant for different classes.

$$a_{li} = \frac{\exp\left(h_i^\top u_l\right)}{\sum_{i'} \exp\left(h_{i'}^\top u_l\right)} \tag{4..3}$$

$$d_l = \frac{1}{T} \sum_{i=1}^{T} a_{li} h_i \tag{4..4}$$

Then, each attention head (d_l) goes through an independent dense layer with $sigmoid()$ activation (L total dense layers) to produce a probability for the corresponding class. Eventually, to infer the concluding class, $argmax()$ is applied over all the generated probabilities to obtain the most possible one.

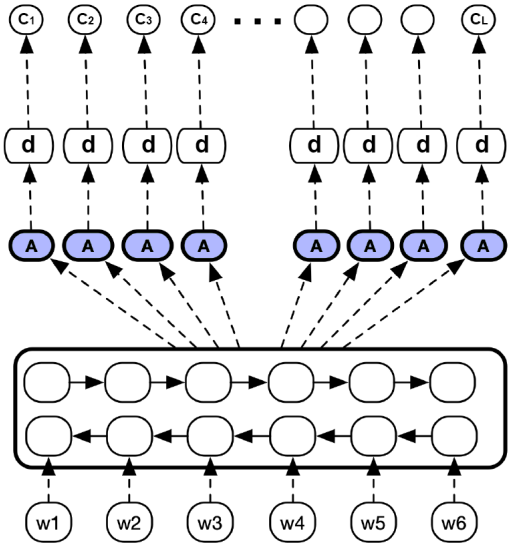


Figure 21: Architecture of BiGRUs - Label-Wise Attention Network with Word2Vec Embeddings

4.1.6 BERT-Base-Multilingual

Recent studies have shown that language representation models pre-trained on vast unlabelled corpora and fine-tuned to undertake specific tasks, have superior performance compared to models trained only on task-specific data. This approach is identified as transfer-learning and is particularly useful in domains like legal, where labelled data are limited. Following this paradigm, we also experiment with advanced transfer-learning models.

BERT [31] is a Transformer-based [32] language model developed by Google. In the inner workings, 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 layer jointly trained with task-specific data. The original BERT model comes with two pre-trained general types, both of them trained on large corpora:

- The BERT-Base model, a 12-layer, 768-hidden, 12-heads, 110M parameter neural network architecture
- The BERT-Large model, a 24-layer, 1024-hidden, 16-heads, 340M parameter neural network architecture.

On account of limited resources, in our implementation, we employ the multilingual version of the BERT-Base-uncased model, which supports modern Greek out-of-the-box. For the attempted multi-class task, we fine-tune the training process with `RAPTARCHIS47k` train-data and add a linear layer on top of BERT-Base-ML with $softmax()$ activation function, for probability distribution to be applied over the classes. This extra dense layer is fed with the so-called “classification token” of the BERT-Base model as described in [31], serving as the final document representation.

Although BERT (or BERT-based flavours) seems to be the king of NLP tasks — reporting state-of-the-art results — it has a size limitation in terms of tokens at 512 wordpieces. This aspect highlights an important disadvantage in processing long documents, a relatively common attribute in domains like legal text processing. However, we consider this not to be the case in our study, as most documents are below or near this limit (see subsection 3.1.3). For the rest of them that exceed this cap, the typical solution is to truncate their length to meet BERT’s maximum.

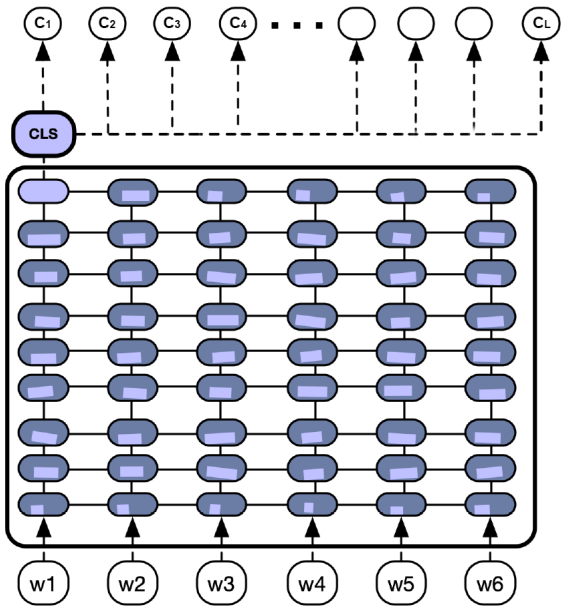


Figure 22: Architecture of BERT-Base-Multilingual

4.1.7 XLM-RoBERTa

The next BERT-based model we employed is XLM-RoBERTa [33], a multilingual adaptation of RoBERTa [34]. RoBERTa is built pretty much following BERT's architecture while showing that hyperparameter choices have a significant impact on the final results. It modifies key hyperparameters, removing the next-sentence pre-training objective and training with much larger mini-batches and learning rates. Based on this study, Conneau et al. proposed the XLM-RoBERTa model, which supports 100 different languages and is trained on 2.5TB of filtered CommonCrawl data.

The required fine-tuning for this model to meet the specifications of our task was the one previously described in the BERT-Base-Multilingual approach; enhancement of the training process with our task-specific data and application of a final dense layer with $softmax()$ activations fed with the "classification token", able to generate the requested probability distribution.

4.1.8 GREEK-BERT

To avoid all being Greek to us, the final model we experimented with is GREEK-BERT [35], a monolingual version of BERT-base for modern Greek. Although multilingual models like the previous two offer exceptional performance even in zero-shot configurations (e.g., fine-tune a pre-trained model in one language for a particular task and use it in another language for the same task without further training), monolingual models usually surpass them in most downstream tasks.

Towards that end, Koutsikakis et al. introduced and made publicly available GREEK-BERT, a monolingual pre-trained Transformer-based language model for modern Greek based on the architecture of BERT-BASE-UNCASED, achieving state-of-the-art results. Their model 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) and a clean version of Common Crawl in the Greek language. Again, the final model configuration we used for our task follows the paradigm of 4.1.6.

4.2 Experimental Setup

To begin with, nearly all the experiments were deployed on one machine with a single NVIDIA GeForce RTX 2080ti GPU, i7-9700K CPU and 32GB of RAM. A minor part of the experiments was deployed on a secondary machine with a single NVIDIA GeForce GTX 1080ti GPU, i5-7600 CPU and 32GB of RAM. The experimental setup per method follows in detail.

Support Vector Machines with Bag-of-Words Features:
For this method, we used Scikit-learn's SVM package[2] (SVC implementation) with Bag-of-Words (BoW) features weighted by TF-IDF, adopting the "one-versus-one" approach for multi-class classification. The optimal hyper-parameters were detected by performing grid-search over a search space of the most indicative of them. For the *kernel* we experimented with *{linear, rbf}*, while the misclassification cost *C* was set in the range *{0.25, 0.50, 0.75, 1}*. Next, we tuned the *n-gram* order in the range *{1, 2, 3, 4, 5}* (i.e. (1, 3), (1, 5)) and the number of *n-gram features* in the range *{200k, 400k}*. When n>1 we use n-grams up to order of n, e.g. for n=3 we use 1-grams, 2-grams and 3-grams.

XGBoost with Bag-of-Words Features:
We used the Python version of the XGBClassifier from the official library of XGBoost[3], utilizing Bag-of-Words (BoW) features weighted by TF-IDF. All the experiments were performed using the default booster *gbtree*, with the *objective* being *multi-class with softmax* function and the *number of estimators* set to *800*. A grid-search was performed over the following search space: the *maximum depth of a tree* was set in the range *{4, 5, 7, 10}* and the *minimum child weight* was tuned in the range *{2, 5, 10}*. Lastly, an *early stopping mechanism* per 10 rounds was employed, using the *mlogloss evaluation metric*.

BiGRUs - MaxPooling, Self-Attention, Label-Wise Attention with Word2Vec:
All the BiGRUs-based neural methods were implemented over Keras API[4] with tensorflow-gpu[5] as its backend, utilizing *100-D Word2Vec* embeddings pre-trained on Greek legal data [4]. We used *Glorot initialization* [36], *categorical cross-entropy loss* and the *Adam optimizer* [37] with default *learning rate 1e-3* to train the classifiers with an *early stopping mechanism* by examining the development loss (max. epochs were set to 50). HY-

[2]See: https://scikit-learn.org/stable/modules/svm.html
[3]See: https://xgboost.readthedocs.io/en/latest/python/index.html
[4]See: https://keras.io/
[5]See: https://www.tensorflow.org/install/gpu/

PEROPT[6] library was used to tune hyper-parameters by sampling 50 combinations per method/task (out of 144) and selecting the values with the best development loss in each task. The following sets of hyper-parameters were examined: *number of stacked BiGRU layers {1, 2}, GRU hidden units {200, 300 400}, batch size {8, 16}, dropout rate {0.1, 0.2, 0.3, 0.4}* and *word dropout rate {0, 0.01, 0.02}*.

BERT-Base-Multilingual, XLM-RoBERTa, GREEK-BERT:
BERT-based neural methods were developed in Tensorflow 2.0[7], also relying on the HuggingFace[8] Transformers library for BERT-based models. All three experimental models were built upon the *BERT-base-uncased* version (12-layers, 768-hidden, 12-heads, 110M parameters) and the *batch size* was set to *8* due to hardware resources limitations. Initially, the *dropout rate* and the *learning rate* were set to *0.1* and *5e-5* respectively, as suggested by Devlin et al. [31]. However, to enhance the range of experimental setups and yield improved results, we performed grid search over the *learning rate* hyper-parameter: *{1e-5, 2e-5, 3e-5, 5e-5}* for BERT-Base-Multilingual and Greek-BERT, *{1e-5, 2e-5}* for XLM-RoBERTa. All these values were tuned according to the best loss on the development data, employing *Adam optimizer*. Eventually, we noticed that the models did not necessarily converge in the fourth epoch, as suggested by Devlin et al. Therefore, we used an *early-stopping mechanism* and trained the models for twelve to fifteen epochs on average (while rarely, some configurations needed more than 35 epochs to converge).

4.3 Evaluation Measures

Apart from the methodology followed and before the demonstration of the results, it is essential to properly showcase the employed measures we used to evaluate our models' performance.

The first metric is Precision, also known as *Positive Predictive Value (PPV)*. It is the fraction of successfully retrieved instances (positive) among the retrieved instances (both positive and negative). Intuitively, precision is the answer to the question: *"What proportion of predicted positives is truly positive?"*.

$$Precision = TruePositives/(TruePositives + FalsePositives)$$

The next complementary metric is *Recall*, also known as *True Positive Rate (TPR)* or *Sensitivity*. It is the fraction of the successfully retrieved instances over the total number of relevant instances (retrieved or not). Recall offers an answer to the question: *"What proportion of actual positives is correctly classified?"*.

$$Recall = TruePositives/(TruePositives + FalseNegatives)$$

[6]Available at: https://github.com/hyperopt/hyperopt/
[7]See: https://www.tensorflow.org//
[8]See: https://huggingface.co//

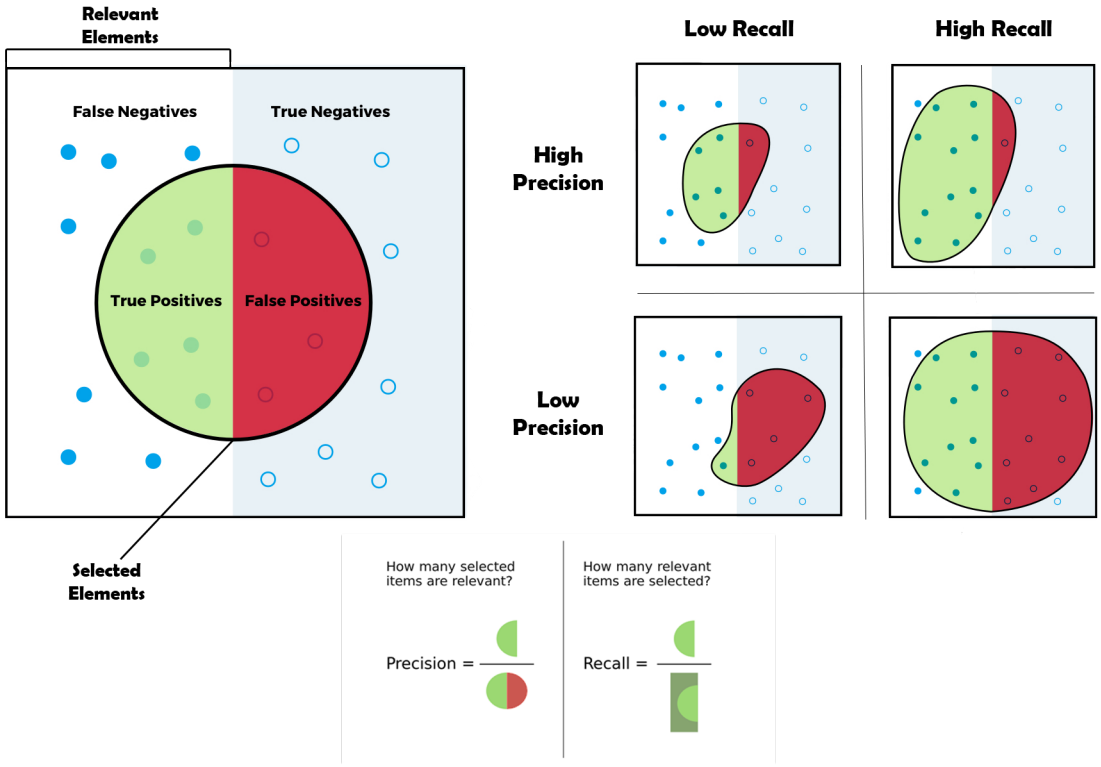


Figure 23: Precision vs Recall example

Generally, classifiers with higher precision and recall scores are preferable. However, there is a trade-off between precision and recall, as when tuning a classifier, the improvement of the recall score often results in the decrease of the precision score and vice versa — there is no such thing as a free lunch. So, to get an overall performance, it is necessary to combine both these metrics into a single one. Hence, a third metric is used named *F1-Score*. F1-Score provides a way to combine precision and recall into a single number. It is computed using a mean ("average"), but not the usual *arithmetic mean* but the *harmonic mean*, which is given by this formula:

$$F1 - score = 2{\times}(Precision{\times}Recall)/(Precision + Recall)$$

Finally, as we are dealing with a multi-class problem, we ought to combine the per-class F1-scores into a single number to get the classifier's overall F1-score. There are a few ways of doing that such as *macro-averaged, weighted-average* or *micro-averaged F1-Score*. The first one is computed as the arithmetic mean of the per-class F1-scores, the second one as the arithmetic mean of the weighted per-classes F1-scores (usually by the number of samples in that class) and the third one aggregates the contributions of all classes to compute the average F1-Score.

According to relevant academic work and since our multi-class problem deals with class imbalances, we chose to report the *micro-averaged* versions of the measures and mainly,

the *micro-averaged F1-Score* as the classifier's overall F1-score. Specifically, for a fair-minded evaluation, we compare the models' performance on development data, and for the best one, we report the average performance on test data after one rerun. Micro-F1 intuitively makes sure that most cases, regardless of the class, are assigned to the correct class. Of course, that means that the more representative the class is, the more influence it has in the metric while poor results in low-representative classes will not have a significant impact on the score. However, we consider this not to be a botheration as we mostly care for the overall data performance and not preferably to any class.

Furthermore, we enhance the overall performance report with two additional ranking metrics — introduced in Information Retrieval [38] — that have also been used in LMTC tasks [18, 17]: *Recall at top-K predictions (R@K)* and *Normalized Discounted Cumulative Gain at top-K predictions (nDCG@K)*. According to literature, these metrics are most appropriate in tasks mentioned earlier or when dealing with ranked results. Nevertheless, we consider them being particularly interesting also for our task. The macro-averaged versions of *R@K* and *nDCG@K* are defined as follows:

$$R@K = \frac{1}{T}\sum_{t=1}^{T}\frac{S_t(K)}{R_t} \tag{4..5}$$

$$nDCG@K = \frac{1}{T}\sum_{t=1}^{T}\sum_{k=1}^{K}\frac{2^{S_t(k)}-1}{\log(1+k)} \tag{4..6}$$

Where (T) is the total number of test documents, (K) is the number of labels to be selected per document, $S_t(K)$ is the number of correct labels among those ranked as top-K for the *t-th* document, and (R_t) is the number of gold labels for each document. Specifically for our task, (R_t) always equals 1 as there is only one correct class for each document and $S_t(K)$ equals either 1 or 0 (i.e. the correct class exists or not among those ranked as top-K).

A more intuitive description for both these metrics is that *R@K* denotes the proportion of relevant items found in the top-K predictions while *nDCG@K* is a measure of ranking quality, yielding a better score when the gold label(s)/class(es) are ranked highly.

4.4 Experimental Results

For the evaluation of the optimal setup per method/task, micro-F1 on the development data was audited. For the final and presented experimental results, we report micro-averaged precision (P), recall (R) and F1, calculating the average performance on the test data after a rerun of the optimal configuration. We only report frequent and few-shot categories, omitting zero-shot (for reasons discussed in section 3.3), as none of our methods is capable of undertaking the classification task on zero-shot learning setup. A more thorough interpretation of the following results is expected through the correlation of the dataset's quantitative analysis in subsection 3.1.3. Finally yet importantly, we believe that each task (i.e. per classification level) should be evaluated independently, as a different classification

problem with no correlation between the others; mostly because we do not adopt a hierarchical classification approach and because data are diversely distributed over classes in each task. Besides, there is not a single classifier that consistently surpasses the others; it mostly depends on the problem and the data.

4.4.1 Volume-level Classification Evaluation

In volume-level classification, all the 47 possible classes are sufficiently represented and belong to the frequent-classes category, while the mean of documents per class is approximately 1k. However, we acknowledge that class imbalance is an existing complication in our dataset, though not so evident here. Thus, this first task provides a convenient classification setup, with a typical number of different classes and relatively straightforward results. On Table 8, we report micro P, R, F1 scores for the volume-level task while on Table 9, we report the corresponding R@{3,5} and nDCG@{3,5} scores.

Table 8: Volume-level classification experiments: P-R-F1 scores

	VOLUME								
	ALL LABELS			FREQUENT			FEW-SHOT		
	P	R	F1	P	R	F1	P	R	F1
SVM-BOW	85.3	85.3	85.3	85.3	85.3	85.3	-	-	-
XGBOOST-BOW	77.2	77.2	77.2	77.2	77.2	77.2	-	-	-
BIGRU-MAX	84.3	84.3	84.3	84.3	84.3	84.3	-	-	-
BIGRU-ATT	86.4	86.4	86.4	86.4	86.4	86.4	-	-	-
BIGRU-LWAN	84.1	84.1	84.1	84.1	84.1	84.1	-	-	-
BERT-BASE-ML	87.7	87.7	87.7	87.7	87.7	87.7	-	-	-
XLM-ROBERTA	87.7	87.7	87.7	87.7	87.7	87.7	-	-	-
GREEK-BERT	88.2	88.2	**88.2**	88.2	88.2	**88.2**	-	-	-

Table 9: Volume-level classification experiments: R@K and nDCG@K scores

	VOLUME											
	ALL LABELS				FREQUENT				FEW-SHOT			
	R@3	R@5	nDCG@3	nDCG@5	R@3	R@5	nDCG@3	nDCG@5	R@3	R@5	nDCG@3	nDCG@5
SVM-BOW	95.2	97.0	91.2	92.0	95.2	97.0	91.2	92.0	-	-	-	-
XGBOOST-BOW	90.6	93.4	85.2	86.3	90.6	93.4	85.2	86.3	-	-	-	-
BIGRU-MAX	94.0	96.3	90.1	91.0	94.0	96.3	90.1	91.0	-	-	-	-
BIGRU-ATT	94.7	96.5	91.3	92.1	94.7	96.5	91.3	92.1	-	-	-	-
BIGRU-LWAN	94.2	96.4	90.1	91.0	94.2	96.4	90.1	91.0	-	-	-	-
BERT-BASE-ML	95.2	96.8	92.2	92.9	95.2	96.8	92.2	92.9	-	-	-	-
XLM-ROBERTA	95.7	97.1	92.5	93.1	95.7	97.1	92.5	93.1	-	-	-	-
GREEK-BERT	**95.8**	**97.2**	**92.8**	**93.3**	**95.8**	**97.2**	**92.8**	**93.3**	-	-	-	-

At first, we notice that P, R and micro-F1 scores per method are equal; this observation holds for every multi-class problem when considering all the data and compute the micro-

averaged versions of these metrics. Also, inspecting the ranked metrics table, we see that classifiers' scores follow (almost) the same ranking as that in the P-R-F1 table, with rates getting better when K increases.

To the point, SVM-BOW proves to be a strong competitor in our classification task (85.3 in F1), even when utilizing simplified word representations as BoW. Interestingly, it outperforms even two of our neural methods with domain-specific word embeddings; BIGRU-MAX and BIGRU-LWAN with 84.3 and 84.1 F1 scores, respectively. As for XGBoost, although it seemed quite promising and very fast at training, its inadequate F1 score of 77.2 places it to the bottom of our classifiers list for this task. Perhaps the skewed class distribution affects its overall performance, and so, the training algorithm does not weigh the misclassification of minority classes appropriately.

Among the neural methods based on BiGRU encoders, BiGRU with self-attention (BIGRU-ATT) achieves a remarkable F1 score of 86.4. 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 assume that its max-pooling layer diminishes 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 its architecture does not offer any performance improvement compared to BIGRU-ATT.

Beyond doubt, transformer-based methods go up to the volume-task classification podium, with GREEK-BERT being the best method we experimented with, achieving a score of 88.2 in F1. The other two multilingual models (BERT-Base-Multilingual and XLM-RoBERTa) also achieve an exceptional score of 87.7 both, confirming their claim to offer top-notch results in most downstream NLP tasks. In like manner, GREEK-BERT verifies its claim too, as proves that monolingual models are able to surpass other advanced multilingual transformer-based models. The fact that it is pre-trained entirely on Greek corpora grants it an advantage over the other methods, especially in the current task where the samples per class are adequate.

4.4.2 Chapter-level Classification Evaluation

In chapter-level classification, the number of the total classes is 389; 333 are frequent classes, and 53 are few-shot classes (with the rest classes being zero-shot). With class imbalance popping up in this task, we expect the results to be reasonably intriguing. On Table 10 and Table 11, we present the scores of this level's experiments.

Table 10: Chapter-level classification experiments: P-R-F1 scores

	CHAPTER								
	ALL LABELS			FREQUENT			FEW-SHOT		
	P	R	F1	P	R	F1	P	R	F1
SVM-BOW	77.9	77.9	77.9	77.9	78.6	78.2	90.0	09.3	16.8
XGBOOST-BOW	67.5	67.5	67.5	67.8	68.1	67.9	19.2	10.3	13.4
BIGRU-MAX	77.5	77.5	77.5	77.9	77.9	77.9	44.9	45.4	45.1
BIGRU-ATT	81.1	81.1	81.1	81.1	81.6	81.3	86.7	40.2	54.9
BIGRU-LWAN	76.8	76.8	76.8	76.9	77.3	77.1	63.8	30.9	41.7
BERT-BASE-ML	82.4	82.4	**82.4**	82.4	82.7	**82.6**	84.1	54.6	**66.3**
XLM-ROBERTA	81.0	81.0	81.0	81.0	81.4	81.2	78.7	38.1	51.4
GREEK-BERT	81.4	81.4	81.4	81.4	81.8	81.6	81.3	40.2	53.8

Table 11: Chapter-level classification experiments: R@K and nDCG@K scores

	CHAPTER											
	ALL LABELS				FREQUENT				FEW-SHOT			
	R@3	R@5	nDCG@3	nDCG@5	R@3	R@5	nDCG@3	nDCG@5	R@3	R@5	nDCG@3	nDCG@5
SVM-BOW	89.8	92.5	85.0	86.1	90.6	93.2	85.8	86.8	81.4	83.5	76.0	76.8
XGBOOST-BOW	81.4	85.0	75.8	77.2	82.2	85.8	75.5	78.0	59.8	67.0	53.0	56.0
BIGRU-MAX	88.9	92.1	84.2	85.6	89.4	92.5	84.8	86.1	89.7	92.8	85.0	86.3
BIGRU-ATT	90.8	93.1	86.9	87.8	91.2	93.5	87.3	88.2	**92.8**	**93.8**	**88.7**	**89.1**
BIGRU-LWAN	88.1	91.2	83.6	84.8	88.6	91.7	84.1	85.4	75.3	81.4	68.4	70.9
BERT-BASE-ML	**91.9**	**93.8**	**88.0**	**88.9**	**92.2**	**94.1**	**88.4**	**89.2**	85.6	87.6	82.9	83.7
XLM-ROBERTA	91.1	93.7	87.0	88.1	91.5	**94.1**	87.5	88.6	91.8	92.8	87.9	88.4
GREEK-BERT	91.4	**93.8**	87.4	88.4	91.7	**94.1**	87.8	88.8	**92.8**	**93.8**	88.6	89.0

Initially, we see that XGBOOST-BOW has the lowest performance with 67.5 in overall F1 score. Looking more closely at few-shot evaluation, we notice that its performance (13.4) is far below the other neural methods, being only comparable with that of SVM (16.8). We strongly believe that two different aspects mainly induce its incompetent scores: i) the data scarcity, causing the classifier not to be adequately trained in predicting few-shot classes and ii) the BOW features that fail to provide context-aware and meaningful word representations. Hence, XGBOOST-BOW seems not to be the recommended method to handle small or imbalanced datasets — at least without thorough fine-tuning — as it proves to be incapable of understanding the semantics of a few samples efficiently.

The same conclusion also applies to SVM-BOW for the few-shot category. Despite its surprisingly high few-shot precision score of 90.0, its low recall score of 09.3 indicates its inability to classify most of the few-shot samples correctly. These rates could also be interpreted by taking into consideration the highly imbalanced dataset, something that is also discussed in related work [13, 39]. Probably, a more all-encompassing words' representation would improve their overall performance; an observation that becomes apparent in the following neural methods that utilize domain-specific word embeddings. Nevertheless, its great overall F1 score of 77.9 could mainly be explained by the classifier's ability to highly correlate certain words (even uncommon ones) with specific classes.

At this point, it is also essential to justify the high overall scores of these methods regardless of their pretty low scores in the few-shot category. As most of the documents in the chapter-level task belong to frequent classes (~99% of the total documents as shown in Table 6), the few-shot category slightly affects the rates of the overall micro-averaged versions of the selected evaluation measures. The fact that the less-representative classes marginally influence micro-averaged metrics is an aspect that does not concern us that much, as we mostly care for the overall data performance and not preferably to any class. What is more, due to the small number of documents belonging to the few-shot category, we assess these results as not that revealing.

Moving on to recurrent architectures, although two of them (BIGRU-MAX and BIGRU-LWAN with 77.5 and 76.8 F1 scores respectively) are outperformed by SVM-BOW, the impact of pre-trained word embeddings becomes evident in the few-shot category where the recurrent architectures surpass by far SVM-BOW and XGBOOST-BOW methods. The characteristic of the BiGRU component to operate on a single sequence of concatenated and context-aware facts also is of major importance. Furthermore, BIGRU-ATT method succeeds in outperforming even XLM-RoBERTa, achieving a slightly higher F1 score (81.1 against 81.0) along with a wider margin in the few-shot's F1 score (54.9 against 51.4). Hence, adjusted RNN methods should still be considered as strong competitors even to most recent transfer-learning models.

Once more, the two best methods in this task are transformer-based. BERT-BASE-ML outmatches its counterpart monolingual implementation GREEK-BERT with an F1 score of 82.4 against 81.4 while in the few-shot category, the F1 score margin in favour of BERT-BASE-ML becomes even wider (66.3 against 53.8). This comes as no surprise, as recent studies have shown that multilingual models are quite competitive and its performance is remarkably close to that of monolingual models on monolingual benchmarks, despite handling 100 languages or more. In some cases like ours, they even manage to surpass them. We assume that in setups with less data, BERT-BASE-ML manages to offer a more robust sense of language context due to its pre-training data diversity.

4.4.3 Subject-level Classification Evaluation

The final task is that of subject-level, with data being even more unequally distributed over classes. To recap, the number of the total classes is 2285; 712 of them are frequent, and 1431 are few-shot (the rest 142 belong to the zero-shot category). The interesting thing here is that the majority of documents (~81%) are classified among the frequent classes (~one-third of total classes), leaving the rest two-thirds of total classes to be associated with no more than 20% of the total documents (see Table 5 and Table 6 in 3.1.3). The tables that follow show the obtained results of this task.

Table 12: Subject-level classification experiments: P-R-F1 scores

	SUBJECT								
	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
BERT-BASE-ML	79.5	79.5	**79.5**	81.6	84.2	**82.9**	70.9	66.5	68.6
XLM-ROBERTA	63.5	63.5	63.5	69.3	70.8	70.1	40.1	39.1	39.6
GREEK-BERT	79.3	79.3	79.3	80.8	83.4	82.1	73.3	68.7	**70.9**

Table 13: Subject-level classification experiments: R@K and nDCG@K scores

	SUBJECT											
	ALL LABELS				FREQUENT				FEW-SHOT			
	R@3	R@5	nDCG@3	nDCG@5	R@3	R@5	nDCG@3	nDCG@5	R@3	R@5	nDCG@3	nDCG@5
SVM-BOW	54.6	61.3	47.6	50.4	68.9	77.2	60.1	63.5	33.8	38.2	29.6	31.4
XGBOOST-BOW	67.7	71.8	62.6	64.3	78.4	82.4	73.1	74.8	43.3	47.6	39.2	41.0
BIGRU-MAX	76.5	80.8	70.9	72.7	85.2	88.9	79.7	81.2	64.3	69.5	58.8	60.9
BIGRU-ATT	83.7	86.5	80.1	81.3	88.9	91.5	85.5	86.6	78.0	81.6	74.6	76.1
BIGRU-LWAN	76.1	79.5	71.7	73.1	84.6	87.7	80.2	81.5	60.3	64.8	56.1	58.0
BERT-BASE-ML	87.9	90.2	84.6	85.5	**92.4**	**94.2**	**89.4**	**90.1**	82.8	85.6	79.4	80.5
XLM-ROBERTA	75.3	79.4	70.4	72.1	83.6	87.0	78.9	80.3	59.3	65.9	53.6	56.3
GREEK-BERT	**88.6**	**90.7**	**84.9**	**85.7**	92.1	93.9	89.0	89.7	**87.0**	**89.4**	**82.8**	**83.8**

To begin with, SVM-BOW achieves the lowest performance with 37.9 in overall F1 score along with rock-bottom, zero few-shot scores. Its failure to classify few-shot classes is also evident in Table 13, where its ranked metrics show its inefficiency to predict the correct class even among top-K predictions. Possibly, its aspect of associating certain words with uncommon classes did not come quite effective in this task. Following, XGBOOST-BOW gets a greater F1 score of 55.3, alongside a rather low few-shot F1 score of 27.2. Regardless of being by far superior to SVM-BOW, its performance again falls short of the other, more sophisticated classifiers.

Directly after, all the recurrent methods offer competent F1 scores, with BIGRU-ATT being outstanding with 74.8 in overall F1 score. It seems that the self-attention head offers a remarkably beneficial document representation, combining information from multiple facts. Surprisingly, BIGRU-LWAN with multiple attention heads manages to beat even XLM-ROBERTA with 65.2 against 63.5 in overall F1 score, signifying the impact pre-trained and context-aware word embeddings can have in such tasks. Perhaps, the different pre-training approach XLM-ROBERTA follows in contrast with BERT (as described in 4.1.7) has a significant impact when dealing with modern Greek. As for BIGRU-MAX, its performance is once more lightly behind that of BIGRU-LWAN with 62.9 in F1, revealing that its max-pooling-based architecture is not able to exceed the performance of akin self-

attention-based methods.

Finally, on the highest rank of this task's evaluation is BERT-BASE-ML, with GREEK-BERT following with a marginally lower F1 score (79.5 vs 79.3). This lineup is reversed in the few-shot category, where GREEK-BERT oversteps BERT-BASE-ML with 70.9 against 68.6 in F1 score. However, due to BERT-BASE-ML's advantage in the frequent category, the overall score remains in his favour. The fact that these transformer-based methods achieve once more the two best overall F1 scores confirm the claim that transfer-learning is the optimal approach to follow in NLP tasks. On top of that, the results highlight the capability of multilingual transformer-based models to compete similar monolingual models even in monolingual tasks, making us question if the training of monolingual models as a rule of thumb is essential.

5. CONCLUSION AND FUTURE WORK

To recapitulate, we introduce RAPTARCHIS47k, a new publicly available dataset consisting of 47k Greek legislation resources, based on an original collection of categorized Greek legislation. Relying on this dataset, we experiment with and evaluate different advanced methods and classifiers, ranging from traditional machine learning and recurrent models to state-of-the-art transfer learning models. Through their performance evaluation, we realize that although traditional machine learning classifiers like SVM set adequate baselines for most of the considered (and ”uncomplicated”) tasks, they fall short against more sophisticated methods. In contrast, fine-tuned recurrent architectures based on bidirectional GRUs provide improved overall performance and even manage to go up against multilingual transformer-based architectures like XLM-RoBERTa. Needless to say, a critical factor of BiGRUs enhanced performance is the pre-trained, domain-specific and context-aware word embeddings that exceptionally capture the semantic similarity of words. Beyond doubt, state-of-the-art multilingual and monolingual Transformer-based models offer top-notch results providing a prosperous sense of language context, with BERT-BASE-ML prevailing in the chapter- and subject-level tasks. At the same time, GREEK-BERT dominates all the tested classifiers in the straightforward volume-level classification task.

Furthermore, we think that emphasis should always be given to the qualitative and quantitative characteristics of the examined datasets before even being utilized in NLP experiments. Intricacies like class imbalance, data scarcity and diversity apparently need special handling. Regarding our study, we noticed that few-shot and zero-shot learning needs to be properly handled with appropriate methods, as standard classifiers are insufficient. As for the urge 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. Perhaps, it would be more fruitful to study the feasibility of extracting smaller and more robust monolingual models from multilingual ones.

Coming up, a practical application of our best model — on which we are currently working — is its deployment in Nomothesia web platform, offering classification predictions for its existing legal resources. Our future work includes experimenting with different approaches, utilizing RAPTARCHIS47k in a better-shaped and more efficient way as described in section 3.2. We plan to investigate other promising methods like dilated CNNs [40], recurrent models with improved few- and zero-shot support [18, 26, 25], different multilingual and monolingual Transformer-based models and mostly, neural methods with a hierarchical-classification approach [41, 42]. Also, experimenting with similar datasets like that of EU Legislation written in Greek will allow us to confirm our current conclusions. Finally, our utter goal in the long run through this study is to support and encourage further research in NLP on Greek. By publishing novel datasets, introducing and experimenting with state-of-the-art methods and supporting reproducibility, research comes into bloom.

ABBREVIATIONS - ACRONYMS

AI	Artificial Intelligence
API	Application Programming Interface
ATT	Attention
BERT	Bidirectional Encoder Representations from Transformers
BIGRU	Bidirectional Gated Recurrent Unit
BOW	Bag Of Words
GLC	Greek Legislation Code
GRU	Gated Recurrent Unit
HAN	Hierarchical Attention Network
JSON	JavaScript Object Notation
LMTC	Large-Scale Multi-Label Text Classification
LSTM	Long Short Term Memory
LWAN	Label-Wise Attention Network
ML	Machine Learning
NDCG	Normalized Discounted Cumulative Gain
NLP	Natural Language Processing
PPV	Positive Predictive Value
RNN	Recurrent Neural Network
SVM	Support-Vector Machines
TF-IDF	Term Frequency-Inverse Document Frequency
TPR	True Positive Rate
URI	Uniform Resource Identifier
W2V	Word2Vec

BIBLIOGRAPHY

[1] I. Chalkidis, I. Androutsopoulos, and A. Michos, "Obligation and prohibition extraction using hierarchical rnns," ArXiv, vol. abs/1805.03871, 2018. 14

[2] I. Chalkidis, I. Androutsopoulos, and A. Michos, "Extracting contract elements," in *ICAIL '17*, 2017. 14

[3] C. Cardellino, M. Teruel, L. Alemany, and S. Villata, "Legal nerc with ontologies, wikipedia and curriculum learning," in *EACL*, 2017. 14

[4] I. Angelidis, I. Chalkidis, and M. Koubarakis, "Named entity recognition, linking and generation for greek legislation," in *JURIX*, 2018. 14, 19, 39, 44

[5] H. Ye, X. Jiang, Z. Luo, and W. Chao, "Interpretable charge predictions for criminal cases: Learning to generate court views from fact descriptions," *ArXiv*, vol. abs/1802.08504, 2018. 14

[6] W. Y. Wang, E. Mayfield, S. Naidu, and J. Dittmar, "Historical analysis of legal opinions with a sparse mixed-effects latent variable model," in *ACL*, 2012. 14

[7] I. Chalkidis, I. Androutsopoulos, and N. Aletras, "Neural legal judgment prediction in english," *ArXiv*, vol. abs/1906.02059, 2019. 14, 19

[8] O.-M. Sulea, M. Zampieri, M. Vela, and J. Genabith, "Predicting the law area and decisions of french supreme court cases," in *RANLP*, 2017. 14, 18

[9] N. Aletras, D. Tsarapatsanis, D. Preotiuc-Pietro, and V. Lampos, "Predicting judicial decisions of the european court of human rights: a natural language processing perspective," *PeerJ Com- put. Sci.*, vol. 2, p. e93, 2016. 14, 18

[10] R. Nallapati and C. D. Manning, "Legal docket classification: Where machine learning stum- bles," in *EMNLP*, 2008. 17

[11] O.-M. Sulea, M. Zampieri, S. Malmasi, M. Vela, L. P. Dinu, and J. Genabith, "Exploring the use of text classification in the legal domain," *ArXiv*, vol. abs/1710.09306, 2017. 18

[12] B. Luo, Y. Feng, J. Xu, X. Zhang, and D. Zhao, "Learning to predict charges for criminal cases with legal basis," *ArXiv*, vol. abs/1707.09168, 2017. 18

[13] T. Gonçalves and P. Quaresma, "Is linguistic information relevant for the classification of legal texts?," in *ICAIL '05*, 2005. 18, 50

[14] S. Undavia, A. Meyers, and J. Ortega, "A comparative study of classifying legal documents with neural networks," *2018 Federated Conference on Computer Science and Information Sys- tems (FedCSIS)*, pp. 515–522, 2018. 18

[15] J. S. T. Howe, L. H. Khang, and I. E. Chai, "Legal area classification: A comparative study of text classifiers on singapore supreme court judgments," *ArXiv*, vol. abs/1904.06470, 2019. 18, 35

[16] J. Mullenbach, S. Wiegreffe, J. Duke, J. Sun, and J. Eisenstein, "Explainable prediction of medical codes from clinical text," *ArXiv*, vol. abs/1802.05695, 2018. 19, 41

[17] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, and I. Androutsopoulos, "Large-scale multi-label text classification on eu legislation," *ArXiv*, vol. abs/1906.02192, 2019. 19, 36, 40, 41, 47

[18] I. Chalkidis, M. Fergadiotis, S. Kotitsas, P. Malakasiotis, N. Aletras, and I. Androutsopoulos, "An empirical study on large-scale multi-label text classification including few and zero-shot labels," in *EMNLP*, 2020. 19, 36, 47, 54

[19] K. Papantoniou and Y. Tzitzikas, "Nlp for the greek language: A brief survey," *11th Hellenic Conference on Artificial Intelligence*, 2020. 19

[20] Z. Pitenis, M. Zampieri, and T. Ranasinghe, "Offensive language identification in greek," in *LREC*, 2020. 19

[21] D. Kouremenos, K. Ntalianis, G. Siolas, and A. Stafylopatis, "Statistical machine translation for greek to greek sign language using parallel corpora produced via rule-based machine trans- lation," in *CIMA@ICTAI*, 2018. 19

[22] V. Athanasiou and M. Maragoudakis, "A novel, gradient boosting framework for sentiment analysis in languages where nlp resources are not plentiful: A case study for modern greek," *Algorithms*, vol. 10, p. 34, 2017. 19

[23] I. Chalkidis, C. Nikolaou, P. Soursos, and M. Koubarakis, "Modeling and querying greek leg- islation using semantic web technologies," in *ESWC*, 2017. 26

[24] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, and I. Androutsopoulos, "Extreme multi-label legal text classification: A case study in eu legislation," *ArXiv*, vol. abs/1905.10892, 2019. 32, 41

[25] A. Rios and R. Kavuluru, "Few-shot and zero-shot multi-label learning for structured label spaces," *Proceedings of the Conference on Empirical Methods in Natural Language Process- ing. Conference on Empirical Methods in Natural Language Processing*, vol. 2018, pp. 3132– 3142, 2018. 32, 36, 54

[26] Z. Hu, X. Li, C. Tu, Z. Liu, and M. Sun, "Few-shot charge prediction with discriminative legal attributes," in *COLING*, 2018. 36, 54

[27] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016. 38

[28] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *NIPS*, 2013. 39

[29] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, 2017. 39

[30] K. Xu, J. Ba, R. Kiros, K. Cho, A. C. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *ICML*, 2015. 40

[31] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” in *NAACL-HLT*, 2019. 42, 45

[32] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” in *NIPS*, 2017. 42

[33] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, “Unsupervised cross-lingual representation learning at scale,” *ArXiv*, vol. abs/1911.02116, 2020. 43

[34] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, “Roberta: A robustly optimized bert pretraining approach,” *ArXiv*, vol. abs/1907.11692, 2019. 43

[35] J. Koutsikakis, I. Chalkidis, P. Malakasiotis, and I. Androutsopoulos, “Greek-bert: The greeks visiting sesame street,” *11th Hellenic Conference on Artificial Intelligence*, 2020. 43

[36] X. Glorot and Y. Bengio, “Understanding the difficulty of training deep feedforward neural networks,” in *AISTATS*, 2010. 44

[37] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *CoRR*, vol. abs/1412.6980, 2015. 44

[38] C. D. Manning, P. Raghavan, and H. Schütze, “Introduction to information retrieval,” 2005. 47

[39] N. Japkowicz, “The class imbalance problem: Significance and strategies,” 2000. 50

[40] N. Kalchbrenner, L. Espeholt, K. Simonyan, A. Oord, A. Graves, and K. Kavukcuoglu, “Neural machine translation in linear time,” *ArXiv*, vol. abs/1610.10099, 2016. 54

[41] N. Manginas, I. Chalkidis, and P. Malakasiotis, “Layer-wise guided training for bert: Learning incrementally refined document representations,” *ArXiv*, vol. abs/2010.05763, 2020. 54

[42] K. Kowsari, D. Brown, M. Heidarysafa, K. Meimandi, M. Gerber, and L. Barnes, “Hdltex: Hierarchical deep learning for text classification,” *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pp. 364–371, 2017. 54