

Multi-Label Classification of Bills from the Italian Senate

Andrea De Angelis¹, Vincenzo di Cicco¹, Giovanni Lalle², Carlo Marchetti² and Paolo Merialdo¹

¹Roma Tre University, Italy

²Senato della Repubblica, Italy

Abstract

The classification of legal texts is usually carried out by domain experts in force at institutions. The classification process is very complex because the reference thesauri are very rich, both in terms of variety of concepts and in terms of numbers. In addition, they often contain very rarely used labels. In this paper we show how to implement a Machine Learning system that can support the domain experts of the Italian Senate, handling infrequently used labels (*Zero/Few-shot classification*) and making the output of the model *explainable* to humans.

Keywords

Legal texts classification, Zero/Few-shot classification, Explainability

1. Introduction

A relevant problem faced by legislative Institutions and by Parliaments is the organization of legal texts in ways fostering their accessibility and prompt consultation by MPs, domain experts, journalists, researchers, and citizens throughout the whole legislative and scrutiny activities. One of the key strategies adopted in this context is the classification of acts, and sometimes of their parts, according to some pre-defined Thesaurus. Classification, when available, enables users to access useful functionality such as topic search and topic-based browsing.

Classification is generally achieved involving legal domain experts reading each text and associating them Thesaurus labels; however, solutions have been developed that enable the task to be semi-automated using Machine Learning (ML) techniques.

Known state-of-the-art approaches are generally specific to one language [5] [10], although there is one study that can support several languages with the same architecture [1]. Also, not all nations or institutions necessarily use the same reference thesaurus. Despite these differences, the thesauri generally share some common features:


- they contain a large number of labels (generally thousands);
- labels represent a wide variety of topics (e.g., politics, transportation, medicine, etc.);
- most labels are used rarely, or even never.

AIxPA 2022: 1st Workshop on AI for Public Administration, December 2nd, 2022, Udine, IT

✉ and.deangelis@hotmail.com (A. De Angelis); vdiccco@os.uniroma3.it (V. di Cicco); giovanni.lalle@senato.it (G. Lalle); carlo.marchetti@senato.it (C. Marchetti); paolo.merialdo@uniroma3.it (P. Merialdo)



© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

This paper summarizes activities and results of a successful project aiming at the development of a ML-based system for supporting experts of the Italian Senate during the classification of legislative proposals (bills) using the labels contained in TESEO,¹ i.e., the official thesaurus designed and maintained by the Italian Senate since 1992. The system developed to support this time-consuming and skill-demanding processes has been designed to satisfy the following requirements and constraints:

- *Human-in-the-loop*. Experts must always be in full control of the labels selected for classifying a text, in order to avoid misclassifications. The system must limit to propose labels that must then be selected (or discarded) by users. The ultimate action of determining the set of labels to be associated to a given text must always be performed by humans.
- *Explainable*. The system must provide easy explanations of proposed labels to domain experts, in order to allow them to quickly assess and evaluate every suggestion.
- *Zero/Few Shot Classification*. TESEO is a huge thesaurus comprising thousands of labels that have an extremely varying rate of use: some of them have been associated to thousands of texts, while some have never been used, yet. As a consequence, the system must be able to handle labels for which few, and sometimes even zero, training examples are available.
- *Upgradable*. Bills and laws deals with a highly dynamic range of topics, continuously absorbing new concepts and themes (e.g., COVID-19 pandemic). The system must therefore embed methods to reliably update its knowledge-base from external resources, other than from its annotated data.

The contribution of our work can be summarized as follows:

- we report our experience in building a bill classification system designed to satisfies the above requirements;
- we publicly release a dataset of Italian bills across Italian Parliament terms XIII-XVIII spanning years 1999-2022, each one annotated with relevant labels coming from TESEO.

The remainder of this paper is organised as follows: Section 2 summarizes related work; Section 3 describes the dataset we release, and the TESEO thesaurus adopted for the classification; Section 4 deals with the overall architecture deployed; Section 5 presents experimental results, demonstrating the effectiveness of the proposed solution.

2. Related work

Applications in LegalAI have been studied intensively, with major contributions in recent years. *Zhong et al.* provide a survey, categorizing techniques and tasks [16]. They also report some of the main challenges in the implementation of a LegalAI system, among which they mention the importance of having an interpretable model to prevent fairness from being compromised.

Several researches address the problem of classification of law texts. *Chalkidis et al.* study the Large-Scale Multi-Label Text Classification (LMTC) problem in the legal field [5], with

¹<https://www.senato.it/tesauro/teseo.html>

the release of a dataset, EURLEX57k, consisting of English legal documents from different European legislatures. They also study the problem of classifying such documents using labels from EUROVOC, the European Union’s multilingual thesaurus. The main problem they face concerns the size of the thesaurus, which contains thousands of labels and their distribution on legal documents: out of 7k labels, only about 4k have at least one annotated example available; of these, only 52% are used at least 10 times, making the problem challenging from a ML perspective. They test several Deep Learning solutions, including one that uses Zero/Few-shot classification techniques to attempt to address the label distribution problem. Later, the same authors delved into the problem of law text classification with a focus on rare label classification, proposing some ad-hoc models for this setup [4]. These approaches meet our requirements about Zero/Few-shot classification and explainable models, but unfortunately are designed specifically and only for the English language.

Avram et al. have proposed PyEurovoc, a tool capable of handling EUROVOC classification in 22 different languages, including Italian [1]. Their solution is based on the use of BERT [8]. This work shows that BERT achieves good performance in several classification task, but does not perform well Zero-shot classification. Also, BERT models are difficult to interpret and it is costly to update, making BERT-based approaches unsuitable for our requirements.

Papaloukas et al. address the problem of classifying legal texts written in Greek [10], a language for which there were few ready-to-use available resources. They release a dataset containing legislative documents annotated with thematic topics and experiment with different Deep Learning solutions for the task of multi-class legal topic classification.

Other authors study the classification of law texts, delving into a particular language. In 2005 *Bartolini et al.* presented SALEM, a tool using NLP techniques to assign a type to each law article and to tag parts of the article with entities in the legal world [2]. They use a very small thesaurus consisting of 8 classes representing the types of provisions in the legal text (e.g., whether it is an amendment to a previous text, whether it introduces an obligation that someone must comply with, etc.).

Finally, with a focus on the Italian language but on a different task from law text classification, in 2021 *Tagarelli & Simeri* presented LambERTa, a novel BERT-based language understanding framework for finding articles of interest out of a legal corpus (the Italian Civil Code) as a response to a query expressed in natural language [11].

3. Thesaurus and Bills Corpus

In this section, first we describe the characteristics of TESEO, the thesaurus of the Italian Senate. Then, we illustrate the bills corpus and how we have organized it in a suitable dataset.

3.1. TESEO: the Thesaurus of the Italian Senate

TESEO is a thesaurus created by the Italian Senate to classify Bills. The number of labels may vary over time as some label may be added, as well as some labels may be deprecated and therefore no longer applicable. At the time of writing, TESEO contains 3,398 labels, whose 100 have been deprecated over time.

Table 1
Dataset characteristics.

articles	avg tokens per article	labels	avg labels per article
28,616	232.38	2,556	2.73

Labels are ordered according to the logical structure of the Universal Decimal Classification² and are organized in a hierarchy, which aims at modeling a wide variety of concepts, such as sports, medicine, art, penal and civil law, and so on.

3.2. The Bills Corpus

Bills of the Italian Legislature are publicly available on the Web³. However, they are not released in a standard, structured format: depending on the legislature, they are either in HTML or XML, with a loose structure that makes it tricky to extract portions of interest such as the title, the articles and associated TESEO labels. For this reason, the first effort of our work was to organize the corpus of bills in a CSV dataset, which is publicly available online⁴.

The dataset contains the texts of the articles of the bills from legislatures XIII - XVIII and are obtained by extraction from XML files, if available, or by web scraping from HTML pages.

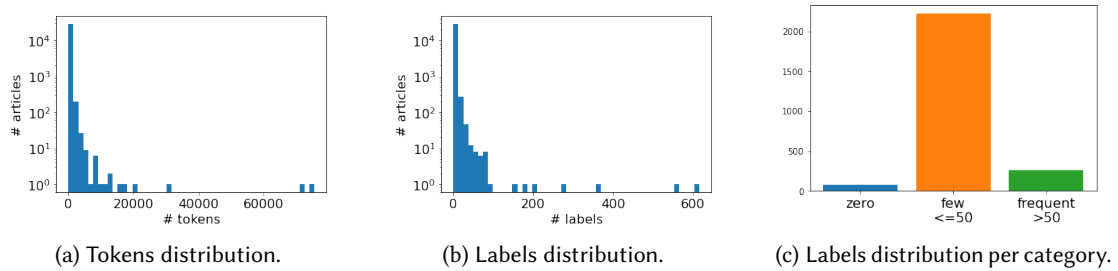


Figure 1: Characteristics of the Corpus of Bills.

Table 1 reports the main features of the dataset: number of the available articles, average number of tokens per article, total number of unique labels associated with the article, and average number of labels per article. Figure 1a shows the distribution of tokens; it is worth observing that the range is very wide, varying between a minimum of 2 tokens and a maximum of $\sim 75k$; the latter are outliers that were excluded from the experimental phase. Figure 1b shows the distribution of labels on articles. Observe that most of the labels are used a few times: there are more than 600 labels that were used only once, while less than 10 labels were used

²<https://udcc.org/index.php/site/page?view=about>

³For example, from the websites of the Italian Senate (<https://www.senato.it/ric/sddl/nuovaricerca.do?params.legislatura=18>, of the Chamber of Deputies <https://www.camera.it/leg18/76>, of the Official Gazette <https://www.gazzettaufficiale.it/>

⁴<https://github.com/SenatoDellaRepubblica/MultiLabelBillClassification>

more than 1000 times. In addition, there are about 800 labels from TESEO (25% of the total) that have never been used to classify an article. According to [5], we used 3 categories to define labels: (i) *frequent*, if they occur more than 50 times in the training set; (ii) *few*, if they occur minimum 1 and at most 50 times in the training set; (iii) *zero*, if they never occur in the training set, but occur at least once in the validation or test set. Figure 1c shows the distribution of labels over categories.

4. System Architecture

The goal of the system is to help domain experts in the classification of bills texts. Given the size of TESEO and the challenges of manually navigating its hierarchy, the proposed solution aims to suggest a shortlist of the most relevant labels for a selected article of a bill. Domain experts can interact with the shortlist in a *human-in-the-loop* fashion, through three different actions: (i) *confirming a label*, that is, moving a suggested label from the shortlist to the final set of labels; (ii) *requiring an explanation*, that is, asking for evidence in the text for the suggested label; (iii) *adding a label*, that is, manually integrating missing labels searching through TESEO. A screenshot of the overall system can be seen in Figure 2.

We tackle the problem of building the shortlist in two steps: first, the system estimates the relevance of each TESEO label through a trained multi-label classifier; then the labels are ranked and selected creating the final shortlist.

This section describes the components of the system. Section 4.1 describes the model used, explaining why it is suitable for Zero/Few-shot classification, and how it enables *explainable* predictions. Section 4.2 describes two different word embedding systems with which we experimented, and how they can enable regular *updates* of the system. Finally, Section 4.3 describes how the labels are selected and ranked to create the final shortlist shown to the domain expert.

4.1. Zero/Few-Shot Multi-Label Classifier

We frame the task of estimating the relevance of each TESEO label to a given article, as a multi-label classification problem.

The main challenges in building such a component lie in handling the large number of TESEO labels despite having few, or even zero, training samples for the majority of them, as described in Section 3.

We build our work on top of ZERO-BIGRU-LWAN, a neural classifier proposed by Chalkidis *et al.* to handle similar challenges [5]. The main feature of the model is that it learns to classify a main text (e.g., an article text) against the tokens of a short text (e.g., a label descriptor), thus exploiting the semantic meaning of each label. In addition, since a label is a generic short text, a trained model can be used on new, unseen, labels allowing to handle additions to TESEO.

First, the model represents each label by encoding their descriptors (i.e., the short textual content of each label), computing the centroid of the Word Embeddings of its tokens:

$$u_l = \frac{1}{E} \sum_{e=1}^E w_{le} \quad (1)$$

Classificazione Teseo

S. 2495 Intero DDL Articolato DDL

Art. 4. (Istituzione del Registro per la trasparenza dell'attività di rappresentanza di interessi, cause di esclusione e di incompatibilità)

1. È istituito presso l'Autorità garante della concorrenza e del mercato il Registro pubblico per la trasparenza dell'attività di rappresentanza di interessi, di seguito denominato « Registro ». Il Registro è tenuto in forma **digitale** ed è articolato distintamente in una parte ad accesso riservato ai soggetti iscritti e alle amministrazioni pubbliche e in una parte ad accesso pubblico, consultabile per via **telematica**. I dati inseriti nel Registro sono di tipo aperto ai sensi dell'articolo 1, comma 1, lettera l-ter) del **codice** dell'amministrazione **digitale**, di cui al decreto legislativo 7 marzo 2005, n. 82. Tutti possono consultare la parte del Registro ad accesso pubblico mediante i sistemi di identificazione informatica previsti all'articolo 64, commi 2-*quater* 2-*nonies*, del **codice** dell'amministrazione **digitale**, di cui al decreto legislativo 7 marzo 2005, n. 82.

2. Il Registro sostituisce ogni altro registro per l'iscrizione di rappresentanti di interessi già istituito alla data di entrata in vigore della presente legge.

Descrittori selezionati Teseo

ALBI ELENCHI E REGISTRI - PUBBLICITA' DI ATTI E DOCUMENTI -

Descrittori suggeriti Teseo Nomi propri Nomi geopolitici

INCARICHI + GIORNALISTI + REGOLAMENTI + PENE DETENTIVE +

DIRIGENTI E PRIMI DIRIGENTI + MINORI +

AUTORITA' INDIPENDENTI DI CONTROLLO E GARANZIA + TELEMATICA + CONCORRENZA +

REATI + VIGILANZA + CONDANNE PENALI + CONTROLLO DELLE NASCITE +

Annulla Copia negli appunti

Figure 2: Screenshot of the system processing the Senate bill n.2495 (<https://www.senato.it/leg/18/BGT/Schede/Ddliter/54699.htm>). In this example, the two suggested labels ALBI ELENCHI E REGISTRI and PUBBLICITA' DI ATTI E DOCUMENTI are confirmed, while the remaining shortlist is shown on the bottom. The domain expert can manually add new labels using the TESEO button, or require an explanation of any label by clicking on it. The highlighted keywords show the explanation for the label TELEMATICA.

where w_{le} is the Word Embedding associated with the e -th token of the l -th descriptor, and E is the total number of tokens in the descriptor. Then, it processes the text by applying a Bidirectional (Bi-GRU) [13, 12] to the text Word Embeddings, producing context-aware representations h_t for each token. Finally, it compares the text and the descriptor content through an attention layer:

$$v_t = \tanh(W h_t + b) \quad (2)$$

$$a_{lt} = \frac{\exp(v_t^\top u_l)}{\sum_{t'} \exp(v_{t'}^\top u_l)} \quad (3)$$

$$d_l = \sum_{t=1}^T a_{lt} v_t \quad (4)$$

where v_t represents the context-aware embeddings processed by a feed-forward layer, a_{lt} is the attention score of the l -th descriptor and the t -th text token, T is the text length, d_l is the final representation of the text for the l -th descriptor.

The final descriptor probability, used as relevance score, is computed as:

$$p_l = \text{sigmoid}(u_l^\top d_l) \quad (5)$$

Training ZERO-BIGRU-LWAN requires that Word Embeddings are kept frozen, leading to a similar representation for u_l for seen and unseen (zero-shot) labels descriptor, giving the model

a mechanism to handle *Zero/Few-shot classification* exploiting labels textual content. Moreover, by using only a single attention layer, it enables a visualization of the most important keywords in the text linked to label, providing an *explanation* to assist Domain Experts. Additionally, Word Embeddings, provides a simple and cheap mechanism to regularly *upgrade* the system by adding new prior knowledge from unlabeled data sources, without expensive pre-training tasks such as the one used by BERT.

4.2. Word Embeddings

ZERO-BIGRU-LWAN requires pre-trained Word Embeddings to provide semantic representations for each token of the article and for the ones inside the label descriptor. In addition, given the wide range of concepts in TESEO and the continuous evolution of legislation texts (for example, consider the word “COVID-19”, which appeared from February 2020 onward), choosing which Word Embeddings to use is an opportunity to add useful prior knowledge to the system. Specialized Word Embeddings for the legal domain have been proposed (e.g., Law2Vec[7]), but they require extensive data collection from many legal sources of the same language; also, specialized knowledge does not perform better than general prior knowledge from sources such as Wikipedia [1]. For this reason, our system uses word vectors learned from plain Wikipedia, taking advantage of its extensive repository of knowledge, its availability in many languages, and its monthly dump release that enables recurring updates.

We experimented with two Word Embeddings models that operate at different levels of word representation: FASTTEXT [3] and WIKIPEDIA2VEC [15]. The former represents words by learning sub-word embeddings (i.e., characters n-grams) through an extension of the skip-gram model [9], thus it is able to model morphological information. The latter, on the other hand, emphasizes semantic information by learning both words and entity embeddings, using a loss function with three components: (1) a *word-based skip-gram model*, (2) a *knowledge base graph model* that learns entity vectors from Wikipedia’s hyperlink graph, and (3) an *anchor context model* that learns to predict words related to a given entity using anchors and their words, encouraging both types of embeddings to lie in the same d -dimensional vector space.

Our final system uses fastText word vectors, learned by means of the efficient open-source library.⁵

4.3. Shortlist Creation

Each label, with its estimated probability, is further processed by a selection component that outputs the final shortlist shown to the domain expert.

We experimented with two selection strategies. First, with a simple top- k , which selects the best k labels according to the model, observing low performance for the group of zero labels. Then, with a custom strategy, named *Ratio Threshold Strategy* (RTS), which attempts to replace unlikely frequent/few labels with promising zero ones, obtaining more balanced results, as we report in Section 5.1.

With RTS we first select the top- k frequent/few labels creating an initial shortlist S . Then, we create a second shortlist Z with the top- m zero labels. Finally, we attempt to replace the

⁵<https://fasttext.cc/>

tail of S (e.g: S_k) with the head of Z (e.g: Z_0). We accept a replacement if $\frac{p_{S_k}}{p_{Z_0}} \leq t$, effectively boosting likely zero labels over unlikely few/frequent ones. The replacement continues as long as the condition holds (up to S_{k-m} and Z_m). The resulting shortlist S , containing k labels of which up to m zero, is then displayed to the domain expert. It is worth observing that we hide label probabilities to eliminate potential biases in experts decision.

By a preliminary user study, we found that $k = 15$, $m = 3$ and $t = 40$ provide a good balance between the effort of the users and the overall performance.

5. Experiments

This section reports the main results of the experimental activity that we have conducted to evaluate the system.

Training and evaluation setup To train and test ZERO-BIGRU-LWAN, we partition both datasets into three training-validation-test sets following an 80%-10%-10% partitioning. We consider all the labels without training examples as zero labels, following the same approach as [5], obtaining the zero/few/frequent distribution shown in Figure 1c. We process both text and label descriptors, making them lowercase while keeping only alphanumeric characters. Additionally, to avoid affecting the centroid representation of each label with non-qualifying words, we remove the most common Italian stopwords from label descriptors. Also, both Wikipedia2Vec and FastText have been trained on a recent dump of the Italian Wikipedia downloaded⁶ in July 2022. For training we use word embeddings with 300 dimensions, the Adam [18] optimizer with a learning rate of 0.001 and batch size of 16, and employ early stopping on the validation loss to reduce overfitting.

We evaluate the system based on the number of true labels retrieved for a given text, regardless of their rank. Following the same argument of [5], we avoided using Precision@K and Recall@K, which can under- or over-estimate performance when K differs from the actual number of true labels (~ 3 in our case, as shown in Table 1). For this reason, we report the R-Precision@K (RP@K) of the model that achieved the best loss on the validation set:

$$\text{RP@K} = \frac{1}{N} \sum_{n=1}^N \frac{S_n(K)}{\min(K, R_n)}$$

Here, N is the total number of articles in the test set, R_n is the number of true labels for the n -th article, and $S_n(K)$ is the number of true labels retrieved in the shortlist S of size K for the current article.

It is worth observing that previous works reports the RP@K of each label group, evaluated against the top- k labels in each group in *isolation* (i.e., zero labels are evaluated by considering only the *top-k zero labels*, ignoring the frequent and few ones), as well as an overall RP@K that considers all the top- k labels by the model. Since the overall RP@K is strongly influenced by the frequency of each label (i.e., the performance of frequent labels is weighted more), it does not provide a clear picture of how often the system is able to place not frequent labels in the final shortlist.

⁶<https://dumps.wikimedia.org/itwiki/20220701/>

Table 2

Evaluation results of the baseline and ZERO-BIGRU-LWAN, with FASTTEXT and WIKIPEDIA2VEC, testing with top-15 and Ratio Treshold Strategy. All results are expressed as RP@15 on the test set.

Method	Word Embeddings	Overall	Frequent	Few	Zero
Baseline-top15	/	0.444	0.413	0.456	0.450
ZERO-BIGRU-LWAN-top15	ft	0.751	0.833	0.589	0.050
ZERO-BIGRU-LWAN-top15	w2k	0.742	0.828	0.566	0.075
ZERO-BIGRU-LWAN-RTS	w2k	0.734	0.821	0.552	0.530
ZERO-BIGRU-LWAN-RTS	ft	0.738	0.823	0.565	0.525

Therefore we report the RP@K of each label group always evaluating against the final shortlist produced and shown to the domain expert by the system.

Baseline We compare the results obtained by ZERO-BIGRU-LWAN with baseline that first encodes both article text and label descriptors using TF-IDF. Then, it ranks each label based on the cosine similarity between the article and the label representation. The final rank is used to select the labels for the shortlist.

We chose this baseline since it is similar, in spirit, to what the attention layer does in ZERO-BIGRU-LWAN, namely “comparing” a label and an article by their semantic representation. Furthermore, it provides information about the difficulty of the task by showing how many classifications can be made by just retrieving labels with similar content.

5.1. System evaluation

Table 2 summarizes the results of our experiment of training the ZERO-BIGRU-LWAN classifier with WIKIPEDIA2VEC and FASTTEXT, creating the shortlist with both a top- k strategy and our proposed Ratio Threshold Strategy (RTS).

We report the RP@K obtained on the test set, using a shortlist of size $K=15$ as this number has been agreed with Domain Experts to be comfortable for a quick overview. As explained in Section 5, we report the RP@15 of each label group always considering the full final shortlist shown to the Domain Expert.

From the table, we can see how the baseline behaves using a top-15 strategy. Despite the results are not high, they clearly show that matching label descriptors with the text content is a good signal for this type of classification. Furthermore, we expected balanced results across label groups since the distinction of frequent/few/zero relates to training data and this baseline is non-learning.

ZERO-BIGRU-LWAN, both with WIKIPEDIA2VEC and FASTTEXT, as expected achieved good RP@15 performance. Even considering the group of labels with few training examples, it beats the baseline by exploiting other than label descriptor content the training data. On the other hand, we found that it struggles to put extremely rare labels (zero) in the final shortlist when used with a naive top-15 strategy. Driven by the results of [5], that reports good RP@5 when considering just zero labels on a similar dataset, we hypothesized that there may be a useful signal in the rank of just zero labels to create a better shortlist. For this reason, our custom

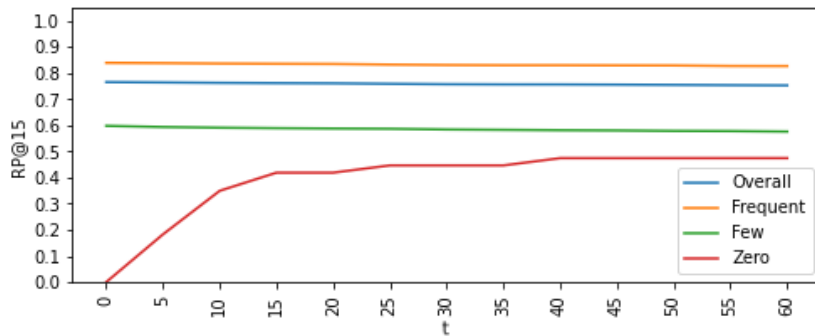


Figure 3: RP@15 of ZERO-BIGRU-LWAN with RTS fixing $m = 3$ while varying t (Ratio Threshold). We selected the point $t = 40$ as it maximizes the performance on the zero labels on the validation set, with minimal loss on other groups.

Ratio Threshold Strategy described in Section 4.3, attempts to replace unlikely frequent/few labels with the top- m zero ones, by boosting the probability of zero labels of a factor t .

We chose $m = 3$ to limit the maximum amount of zero labels in the shortlist to the average number of labels per article (see Table 1) and found $t = 40$ as the first point that maximizes zero performance on the validation set (see Figure 3).

The last two rows of the Table 2 shows ZERO-BIGRU-LWAN using RTS with $k = 15$, $m = 3$, and $t = 40$. As can be seen, the proposed strategy is able to greatly increase the performance on the group of zero labels with minimal loss for the frequent and few groups.

Despite WIKIPEDIA2VEC and FASTTEXT performance being similar, we chose FASTTEXT for our final system since it is slightly better at handling frequent and few labels that represent the majority of TESEO.

5.2. Reliability attention score as explanation

As described in Section 4.1, the attention scores calculated by the model can be used to provide an explanation to the domain experts. The underlying assumption is that there is a causal relationship between the predicted label and the n tokens with higher attention scores.

To test such an assumption, we performed the following experiment. First, we randomly selected 6000 law articles from our dataset (about 30% of the total). For each of these we made

Table 3

Attention table.

n	% drop in probability (best n)	% drop in probability (random n)	avg positions in rank (best n)	avg positions in rank (random n)	% stable labels
1	53%	2%	-91	-11	89%
3	58%	0.5%	-132	-9	81%
5	57%	0.4%	-146	-12	76%
10	56%	0.6%	-166	-12	69%

predictions using the model (case A). Then, for each law article, we randomly selected a label from the top-5 predictions. We repeated the model’s predictions by changing the text of the law article: by removing the best n tokens according to the attention mechanism (case B), by removing n random tokens (case C).

Finally, we compared case A to cases B and C, by computing, respectively:

- how much the model-estimated probability changed for the selected label;
- the difference in the shortlist position of the selected label.

The results of this experiment are shown in table 3. Removing the best n tokens according to attention has a significantly greater impact on the metrics described above than removing n random tokens. For example, with $n = 5$ the model-estimated probability for the selected labels drops by about 57% on average, whereas by removing 5 random tokens the drop is 0.4%. For the ranked list, by removing the best 5 tokens according to attention, labels lose on average 146 positions, while by removing 5 random tokens the label loses 12.

It is also interesting to note that the removal of n tokens from the text has a rather circumscribed impact on the label being analyzed. In fact, the last column of table 3 shows the percentage of labels in common between the shortlist obtained in case A and that obtained in case B: for example, by removing the top 5 tokens according to attention, the newly obtained rank contains on average 76% of the labels it also contained previously.

In conclusion, label prediction by the model appears to be strongly correlated with tokens for which there is a higher attention score. Therefore, we can conclude that highlighting the best n tokens in the text of the article according to attention, in relation to a label, is equivalent to providing an explanation.

5.3. The importance of upgrading Word Embeddings

Although most bills that are written aim to regulate “classic” aspects of our lives, such as criminal law, religion, etc., there are special cases in which a new law may refer to a concept or aspect of life that had never been observed in the past; this new law, therefore, may be written with the use of new words that have never been used previously. For example, think of a technological advancement that needs to be regulated, such as the blockchain, or the outbreak of a pandemic caused by a new disease, such as COVID-19: the word “*blockchain*” first appears in an Italian bill in early 2020, while the Wikipedia page for it, in the Italian version, dates back to March 2016; similarly, before 2020 the word “*covid-19*” did not exist. For this reason, it is crucial to use a frequently updated external knowledge-base for such a classification system. The ability to periodically retrain the word embeddings used is of crucial importance to keep track of changes that, while rare, may occur in our society.

An interesting example that certifies the importance of using word embeddings trained on an up-to-date knowledge-base is given in the table4, in which the text of a law article discussing the Long Covid syndrome is shown. The 5 highlighted words are those with the highest attention scores (in round brackets) for predicting the label `VACCINAZIONI OBBLIGATORIE` (i.e., “mandatory vaccination”) using the ZERO-BIGRU-LWAN model, respectively, with FASTTEXT word embeddings trained on a 2022 Wikipedia dump (on the left) and a 2017 dump (on the right).

Art. 1. 1. Al fine di garantire **la(0.04)** presa in carico delle persone **affette(0.06)** da **sindrome(0.03)** Long **COVID(0.44)**, condizione clinica caratterizzata dal mancato ritorno da parte del paziente affetto da **COVID(0.18)** -19 allo stato di salute precedente l'infezione acuta, le regioni e le province autonome di Trento e di Bolzano istituiscono, presso le aziende sanitarie, appositi centri.

Art. 1. 1. Al fine di garantire la presa in carico delle persone **affette(0.14)** da sindrome **Long(0.13)** COVID, condizione clinica caratterizzata dal mancato ritorno da parte del paziente affetto da COVID-19 allo stato di salute precedente l' **infezione(0.14)** **acuta(0.05)**, le regioni e le province autonome di Trento e di Bolzano istituiscono, presso le aziende **sanitarie(0.10)**, appositi centri.

Figure 4: Comparison between top-5 attention tokens for the label VACCINAZIONI OBBLIGATORIE using ZERO-BIGRU-LWAN trained with fastText 2022 (left) and fastText 2017 (right).

As can be seen, the model with the most recent knowledge-base “knows” the concept “covid” and manages to tie it to the label VACCINAZIONI OBBLIGATORIE, while the other model does not. In fact, the first model succeeds in suggesting the label VACCINAZIONI OBBLIGATORIE in the shortlist, while the other model does not.

6. Future Work

In the future we want to make the tool more affordable for non-practitioners, so that they can also use it to better understand the content of a law text. On a technical side, instead, we would like to:

- include further baseline approaches in the experiments;
- exploit incremental training of word embeddings [17] as an alternative to retraining on all the data, as a possible way to optimize this phase;
- take into account the labels hierarchy to improve the classification, as suggested by [4];

Finally, we want to extend the work so that it can also classify a Bill from the title alone, as this is another interesting use case for the Italian Senate. In addition, we want to integrate EUROVOC in our solution, meeting the standards dictated by the European Union.

Acknowledgement

We would like to thank Manuela Ruisi and Patrizia Toti, the domain experts of the Italian Senate who helped us in the realization of this project, as well as we thank all the annotators who have been involved over the years.

References

- [1] Andrei-Marius Avram, Vasile Pais, and Dan Ioan Tufis. 2021. PyEuroVoc: A Tool for Multilingual Legal Document Classification with EuroVoc Descriptors. In Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021).
- [2] Bartolini, R., Lenci, A., Montemagni, S., Pirrelli, V. and Soria, C., 2004, October. Automatic classification and analysis of provisions in italian legal texts: a case study. In OTM Confed-erated International Conferences" On the Move to Meaningful Internet Systems" (pages 593-604.

- [3] Bojanowski, P., Grave, E., Joulin, A. and Mikolov, T., 2017. Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5, pages 135-146.
- [4] Chalkidis, I., Fergadiotis, M., Kotitsas, S., Malakasiotis, P., Aletras, N. and Androutsopoulos, I., 2020, November. An Empirical Study on Large-Scale Multi-Label Text Classification Including Few and Zero-Shot Labels. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7503-7515.
- [5] Ilias Chalkidis, Emmanouil Fergadiotis, Prodromos Malakasiotis, and Ion Androutsopoulos. 2019. Large-Scale Multi-Label Text Classification on EU Legislation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6314–6322.
- [6] Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The Muppets straight out of Law School. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904.
- [7] Chalkidis Ilias and Dimitrios Kampas, 2019. Deep learning in law: early adaptation and legal word embeddings trained on large corpora. *Artificial Intelligence and Law*, 27(2), pages 171-198.
- [8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT. 2019*, pages 4171-4186
- [9] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, Jeff Dean, 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.
- [10] Christos Papaloukas, Ilias Chalkidis, Konstantinos Athinaios, Despina-Athanasia Pantazi, Manolis Koubarakis, 2021. Multi-granular Legal Topic Classification on Greek Legislation. *arXiv preprint arXiv:2109.15298*.
- [11] Andrea Tagarelli, Andrea Simeri, 2021. Unsupervised law article mining based on deep pre-trained language representation models with application to the Italian civil code. *Artificial Intelligence and Law*, pages 1-57.
- [12] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734.
- [13] Mike Schuster, Kuldip K. Paliwal, 1997. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11), pages 2673-2681.
- [14] Ikuya Yamada, Akari Asai, Jin Sakuma, Hiroyuki Shindo, Hideaki Takeda, Yoshiyasu Takefuji, Yuji Matsumoto, 2020. Wikipedia2Vec: An Efficient Toolkit for Learning and Visualizing the Embeddings of Words and Entities from Wikipedia. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 23-30.
- [15] Ikuya Yamada, Hiroyuki Shindo, Hideaki Takeda, Yoshiyasu Takefuji, 2016. Joint Learning of the Embedding of Words and Entities for Named Entity Disambiguation. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 250-259.

- [16] Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020. How Does NLP Benefit Legal System: A Summary of Legal Artificial Intelligence. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 5218–5230, Online. Association for Computational Linguistics.
- [17] Nobuhiro Kaji and Hayato Kobayashi. 2017. Incremental Skip-gram Model with Negative Sampling. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 363–371.
- [18] Diederik P. Kingma and Jimmy Ba, 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.