Italian Legislative Text Classification for Gazzetta Ufficiale

Marco Rovera,[1] Alessio Palmero Aprosio,[1] Francesco Greco,[2]
Mariano Lucchese,[2] Sara Tonelli,[1] Antonio Antetomaso [2]
[1]Fondazione Bruno Kessler, Trento, Italy
[2]Istituto Poligrafico e Zecca dello Stato, Rome, Italy
{m.rovera, aprosio, satonelli}@fbk.eu
{f.greco, m.lucchese, a.antetomaso}@ipzs.it

Abstract

This work introduces a novel, extensive annotated corpus for multi-label legislative text classification in Italian, based on legal acts from the Gazzetta Ufficiale, the official source of legislative information of the Italian state. The annotated dataset, which we released to the community, comprises over 363,000 titles of legislative acts, spanning over 30 years from 1988 until 2022. Moreover, we evaluate four models for text classification on the dataset, demonstrating how using only the acts' titles can achieve top-level classification performance, with a micro F1-score of 0.87. Also, our analysis shows how Italian domain-adapted legal models do not outperform general-purpose models on the task. Models' performance can be checked by users via a demonstrator system provided in support of this work.

1 Introduction

The *Gazzetta Ufficiale*[1] (GU), in both its printed and digital editions, is the official source of the Italian Republic through which every legislative act issued by Italian central and peripheral institutions, like the Parliament, the Constitutional Court, the Ministeries, the regional administrations, among others, is brought to the attention of citizens. This official journal plays a key role in Italian lawmaking, as for any legislative measure to enter into effect, its publication in the GU is explicitly required by law. The Istituto Poligrafico e Zecca dello Stato[2] (IPZS), based in Rome, is in charge of editing and publishing the *Gazzetta*. A crucial step in the publication process is the assignment of each legislative measure to one or more labels from a subject index to concisely express its main semantic content and ease future search. However, so far this annotation has been carried out only manually, relying on a set of annotators that needed to be carefully trained to master around 1,600 labels available in the taxonomy. In order to support this task, reduce manual effort and ensure future consistency, we develop an automatic pipeline for the classification of GU documents. The task is particularly challenging due to the large number of labels, their different nature (i.e. mandatory vs. optional), the uneven representation of some categories and the different topics covered in the data. In our classification experiments we compare different transformer models for Italian, focusing in particular on performance differences between general-purpose and legal adapted models. We release the classification models for reproducibility, which can also be tested through our online demo.[3] We also make available a GitHub repository[4] with the annotated dataset used for our experiments, which consists of over 363k act titles from the General Series, manually annotated by domain experts.

2 Related Work

In the legal domain, text classification has an established tradition, both in the monolingual (Šarić et al., 2014; Papaloukas et al., 2021) and in the multi-lingual setting (Steinberger et al., 2006, 2012; Chalkidis et al., 2019; Avram et al., 2021; Chalkidis et al., 2021). Moreover, the large availability of legal data, produced by national and supranational public institutions, set the stage for the development of domain-adapted models (Chalkidis et al., 2020; Douka et al., 2021; Masala et al., 2021; Licari and Comandè, 2022). As for Italian, a multi-label classification system for bills has been proposed by De Angelis et al. (2022), based on Bi-GRU architecture using static word embeddings and employing a dataset of 28k legal document tagged with the TESEO thesaurus. In our work, conversely, we

[1] https://www.gazzettaufficiale.it
[2] https://www.ipzs.it/ext/index.html
[3] https://dh-server.fbk.eu/ipzs-ui-demo/
[4] https://github.com/dhfbk/gazzetta-ufficiale

Proceedings of the Natural Legal Language Processing Workshop 2023, pages 44–50

compare the performance of different BERT-based models (Devlin et al., 2019) for Italian, including domain-adapted ones (Licari and Comandè, 2022), on a large multi-label legal dataset covering 35 years of publication in the Gazzetta Ufficiale.

3 Gazzetta Ufficiale Subject Index

Each Italian legislative act entering the GU has to be manually assigned one or more thematic labels to classify its semantic content. Such labels, provided by the GU Subject Index, are then used by the data administrators for classification (in the printed edition) and for indexing and retrieval of Italian legislative acts (in the digital version). The manual labelling has been performed by a pool of expert annotators, trained by the data administration institution, over a time span of almost 35 years.

Layer	N. of unique labels
`Open Labels` (M)	781
`Closed Labels` (O)	1,019
`References` (O)	149
`Summaries` (O)	71

Table 1: Layered structure of the GU Subject Index. M = Mandatory, O = Optional.

Table 1 summarizes the structure of the Subject Index used for annotation. The resource is organized on four levels: *Voci Aperte* (Open Labels), *Voci Chiuse* (Closed Labels), *Riferimenti* (References) and *Sommarietti* (Summaries). Open Labels represent the main layer of the Subject Index and are considered mandatory, in the sense that each item to be labeled must receive at least one Open Label. Secondary labels are divided into three layers and are used as additional, optional refinement as needed. Closed Labels are specifiers that have the purpose of delimiting the meaning of the main open label. References and Summaries refer to thematic areas. As far as combinations are concerned, one or more Open Labels (the first will be the main one) or, alternatively, one Open Label and one or more labels from other layers can co-exist. Indeed, secondary labels can be used individually or in combination.

4 Dataset

The publication of acts in the GU is structured in six Series, depending on the regulatory body that issues the different types of acts: the *Serie Generale* (General Series) and five special Series. The General Series includes all acts like ordinary laws, presidential decrees, ministerial decrees and resolutions, as well as other regulatory acts from the central and peripheral state administrations. To this respect, it represents the most relevant Series in the GU and the backbone of the Italian legislative process. The five Special Series, on the other hand, focus on specific institutions or types of acts: Special Series 1 contains judgements and orders issued by the *Constitutional Court*, Special Series 2 refers to regulations and directives of the *European Community*, whereas Series 3 includes regulatory and administrative acts issued by *regional administrations*. Special Series 4 and 5 are dedicated to the publication of documents relating to *Public Exams* and *State Contracts*, respectively. Table 2 summarizes the structure of GU with the different document series. Given the high number of legislative acts contained in General Series compared to the other series, and its wide coverage in terms of annotation, this work focuses exclusively on this Series.

Name	Institution/Topic	samples
General Series	Central/ Periph. Adm.	364,123
Sp Series 1	Constitutional Court	16,485
Sp Series 2	European Community	26,926
Sp Series 3	Regional Affairs	2,285
Sp Series 4	Public exams	–
Sp Series 5	State contracts	–

Table 2: Document publication structure in the Gazzetta Ufficiale.

While all acts in GU include a title and the proper body of the law, we focus in this work only on the classification of titles. Indeed, if we consider the annotation procedure, in most cases annotators rely on the title of the act when assigning labels. Only in cases where the title is not sufficiently expressive, or where the act includes several measures of different nature, will the annotator refer to the body of the act. Also, although current large language models can deal with long documents (Beltagy et al., 2020), the use of titles alone allows us to minimize computational costs and avoid expensive preprocessing. In fact, by comparing different transformer-based models on the dataset, we intend to determine what the upper bound of performance is using only titles, which may be useful for future comparisons with document-level classifica-

tion systems.

Our dataset contains 364,123 labeled titles of legislative acts published in the General Series of the GU from 1988, when manual classification began, until early 2022. The temporal distribution of the documents in the dataset is depicted in Figure 1. According to the institutional annotation practices, each act (title) is manually annotated with at least one label chosen from the Open Labels set and, optionally, with one or more labels chosen among the Open Labels or from one of the other three layers (see Table 1). The General Series has a tagset of 1,587 unique labels and an average of 1.57 label assignments per document. The label distribution in the dataset exhibits a pronounced Zipf-like trend, with very few labels counting tens of thousands of assignments and a long queue of labels with few or very few occurrences.Quantitatively speaking, the dataset counts 7,464,114 tokens, with an average title length of 20.5 tokens.

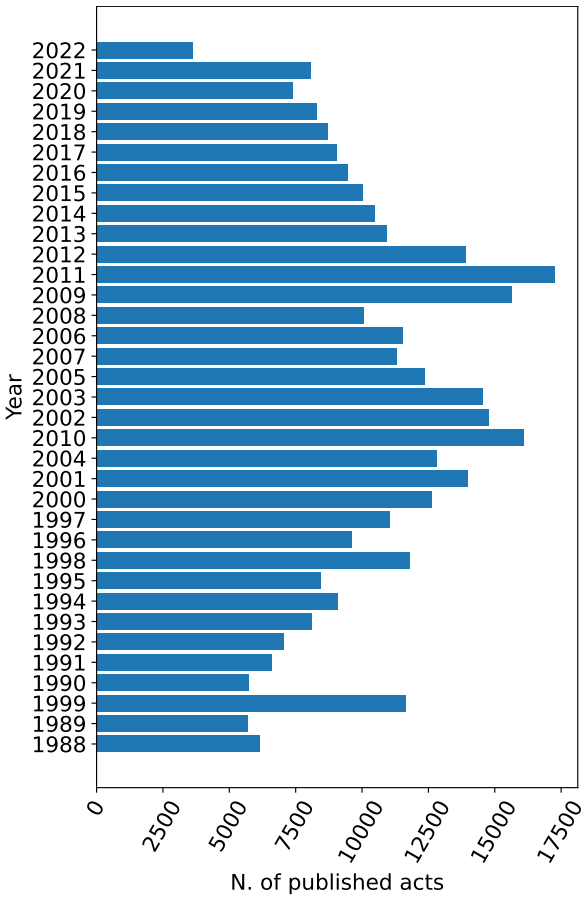

Figure 1: Acts by year

5 Legal Text Classification

We cast the text classification task as a multi-class, multi-label problem, i.e. each target document (the act's title in this case) is to be labeled with one or more classes from the Subject Index. We compare different transformer-based models, after applying the same preprocessing for all settings.

5.1 Preprocessing

Given the extremely skewed data distribution, a cutoff threshold of 10 label assignments is applied to the dataset, in order to allow a robust evaluation. The application of the threshold has little impact on the number of documents in the dataset (passing from 364,123 to 363,909) but the number of unique labels we take into account for the experiments decreases more significantly from 1,587 to 896.

5.2 Experimental Setup

The multi-label text classification experiments we run are based on the comparison of four different transformer-based models, and six different variants (cased/uncased versions of two models). The first two models are general-purpose BERT-base[5][6] and BERT-XXL[7][8] for Italian (in both cased and uncased variants), while the other two are Italian-Legal-BERT[9] and Italian-Legal-BERT-SC (Licari and Comandè, 2022). Italian-Legal-BERT has been obtained by adapting BERT-XXL to the legal domain by further training the model on a set of 235k documents from the National Jurisprudential Archive. Italian-Legal-BERT-SC, instead, has been created from scratch from the same set of legal documents.

In order to provide a robust training and evaluation workflow, we opt for a stratified approach for creating training, development and test sets, with a 60/20/20 split ratio. This results in a proportionally equal distribution of each label over the three splits. Also, we create 4 folds, one for hyper-parameter search and three for training and evaluation. We then use the Optuna toolkit (Akiba et al., 2019) for running a hyper-parameter search, by optimizing the learning rate, with a fixed batch size of 16. The best learning rate values (see Appendix A) have been used for training the four different models.

[5] https://huggingface.co/dbmdz/bert-base-italian-uncased

[6] https://huggingface.co/dbmdz/bert-base-italian-cased

[7] https://huggingface.co/dbmdz/bert-base-italian-xxl-uncased

[8] https://huggingface.co/dbmdz/bert-base-italian-xxl-cased

[9] https://huggingface.co/dlicari/Italian-Legal-BERT

Model	Micro				Macro		Weighted	
	Thr	P	R	F1	Thr	F1	Thr	F1
`bert-base-ita-uncased`	0.47	0.889	0.856	0.872	0.10	0.464	0.33	0.863
`bert-base-ita-cased`	0.46	0.888	0.859	**0.873**	0.11	0.462	0.31	**0.864**
`bert-base-ita-xxl-uncased`	0.47	0.888	0.858	**0.873**	0.09	**0.471**	0.33	**0.864**
`bert-base-ita-xxl-cased`	0.45	0.886	0.858	0.872	0.09	0.455	0.31	0.862
`italian-legal-bert`	0.48	0.892	0.847	0.869	0.12	0.416	0.33	0.857
`italian-legal-bert-sc`	0.46	0.889	0.854	0.871	0.10	0.444	0.29	0.861

Table 3: Performance scores on test data for the task of multi-label text classification on the dataset. Results have been computed on the test sets of three folds.

6 Evaluation and Error Analysis

Performance results, computed on the test sets of the three folds, are reported in Table 3. As model predictions are provided with a confidence score, we use predictions on the development set to find the confidence values which maximize the F1-score, and then apply these (threshold) values on test data. As revealed by the results, with the exception of macro-averaged scores, which are more sensitive to performance variations in classes with few examples, the different models yield a similar performance. In particular, domain adapted models show a comparable, or even slightly lower, performance with respect to their general purpose counterparts, although `italian-legal-bert` has been created by adapting `bert-base-ita-xxl` to the legal domain. This could depend on the fact that our dataset is likely to contain more terms on the regulated domain than legal terminology.

Although not directly comparable due to the use of different datasets/annotation schemas, the performance of our models is in line with current state-of-the-art approaches to legal text classification (Chalkidis et al., 2019; Avram et al., 2021), which however consider the full text of legal acts.

In order to better understand the source of error in our models, we perform an extensive error analysis on the test sets used for evaluation. Since in a multi-label setting it is not always possible to exactly map expected labels with predictions, due to possible many-to-many relations, we create a subset of wrong predictions where one-to-one correspondences are observed and use it as an approximation for the analysis. By manually inspecting the data, we observe that all models struggle with the same set of gold-predicted label mismatches. For example, only four pairs of labels are responsible, alone, for 10.5% of the overall error. These pairs are: *a)* AGRICULTURE vs FOOD AND BEVERAGE (4.54%), *b)* FINANCE ADMINISTRATION vs. ECONOMY AND FINANCE ADMINISTRATION (2.62%), *c)* UNIVERSITY vs. PUBLIC EDUCATION (1.73%), and *d)* PUBLIC HEALTH vs. DRUGS (1.69%).

		Predicted			
		O	C	R	S
Gold	O	70.57	4.8	0.14	0.21
	C	8.75	10.80	0.00	0.36
	R	0.18	0.05	0.08	0.00
	S	0.66	0.23	0.00	3.07

Table 4: Percentage of errors, on single-label assignments, mapped by layer (see Section 3). On the diagonal, errors within the same layer. O = Open Labels, C = Closed Labels, R = References, S = Summaries.

It is worth noting that many of such mismatch types are related to pairs exhibiting strong semantic ties, like overlap (*a*), possible equivalence (*b*) subsumption (*c*, *d*). This pattern, enabled by the very structure of the Subject Index, is pervasive throughout the dataset and might be one of the reasons for a partially inconsistent use of the scheme over the years by different annotators, of which we have found more than one evidence in the gold data. At the other end of the distribution, we observe that 32% of the overall error is explained by a long tail of mismatches occurring 2 times or less. Furthermore, in order to assess the model's ability to learn the hierarchical structure of the Subject Index, we analyze gold-prediction errors by mapping the respective labels' layers. As shown in Table 4, 84.5% of the errors happen between labels in the same layer, showing that, even in case of error, the system tends to be consistent with the hierarchical structure of the annotation schema.

7 Conclusions

This paper first introduces a novel annotated corpus of titles with manually assigned subject labels, which are part of Gazzetta Ufficiale, the official Italian collection of legislative acts. We use this dataset to build and compare four multi-class classifiers for the legal domain, showing that specialised transformer-based models do not outperform general-purpose BERT models for Italian. In this first set of experiments we focused only on acts' titles, given that they yield promising results with a limited computational cost. However, in the future we plan to compare our results with those obtained by processing the full text of the acts.

Our analyses and experiments have highlighted some inconsistencies in the data, mainly due to the complex nature of the taxonomy and the large number of labels. As a next step, we will propose few improvements to make labels more consistent and speed up manual annotation. Furthermore, it would be interesting to measure the gains in terms of time and effort when human annotators are given access to our system as a tool to support manual labelling.

References

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Andrei-Marius Avram, Vasile Păiș, 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)*, pages 92–101.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv:2004.05150*.

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.

Ilias Chalkidis, Manos Fergadiotis, and Ion Androutsopoulos. 2021. Multieurlex-a multi-lingual and multi-label legal document classification dataset for zero-shot cross-lingual transfer. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6974–6996.

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.

Andrea De Angelis, Vincenzo di Cicco, Giovanni Lalle, Carlo Marchetti, and Paolo Merialdo. 2022. Multi-label classification of bills from the italian senate. In *Proceedings of 1st Workshop on AI for Public Administration co-located with 21st International Conference of the Italian Association for Artificial Intelligence (AIxIA 2022)*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Stella Douka, Hadi Abdine, Michalis Vazirgiannis, Rajaa El Hamdani, and David Restrepo Amariles. 2021. Juribert: A masked-language model adaptation for french legal text. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 95–101.

Daniele Licari and Giovanni Comandè. 2022. ITALIAN-LEGAL-BERT: A Pre-trained Transformer Language Model for Italian Law. In *Companion Proceedings of the 23rd International Conference on Knowledge Engineering and Knowledge Management*, volume 3256 of *CEUR Workshop Proceedings*, Bozen-Bolzano, Italy. CEUR. ISSN: 1613-0073.

Mihai Masala, Radu Cristian Alexandru Iacob, Ana Sabina Uban, Marina Cidota, Horia Velicu, Traian Rebedea, and Marius Popescu. 2021. jurbert: A romanian bert model for legal judgement prediction. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 86–94.

Christos Papaloukas, Ilias Chalkidis, Konstantinos Athinaios, Despina Pantazi, and Manolis Koubarakis. 2021. Multi-granular legal topic classification on greek legislation. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 63–75.

Frane Šarić, Bojana Dalbelo Bašić, Marie-Francine Moens, and Jan Šnajder. 2014. Multi-label classification of croatian legal documents using eurovoc thesaurus. In *Proceedings of SPLeT-Semantic processing of legal texts: Legal resources and access to law workshop*. ELRA.

Ralf Steinberger, Mohamed Ebrahim, and Marco Turchi. 2012. Jrc eurovoc indexer jex–a freely available multi-label categorisation tool. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC*, pages 798–805. European Language Resources Association (ELRA).

Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaz Erjavec, Dan Tufis, and Dániel Varga. 2006. The jrc-acquis: A multilingual aligned parallel corpus with 20+ languages. *arXiv preprint cs/0609058*.

A Appendix

Model	Learning Rate	Batch Size	Best Epoch	Weight Decay	Max Length
`bert-base-ita-uncased`	3.673e-5	16	10	0.01	128
`bert-base-ita-cased`	3.673e-5	16	10	0.01	128
`bert-base-ita-xxl-uncased`	3.519e-5	16	10	0.01	128
`bert-base-ita-xxl-cased`	3.519e-5	16	10	0.01	128
`italian-legal-bert`	4.544e-5	16	10	0.01	128
`italian-legal-bert-sc`	2.826e-5	16	10	0.01	128

Table 5: Hyper-parameters used for training. Learning rates have been optimized using Optuna (Akiba et al., 2019).