Italian legislative text classification for Gazzetta Ufficiale

Marco Rovera[1,*], Alessio Palmero Aprosio[1], Francesco Greco[2], Mariano Lucchese[2], Sara Tonelli[1] and Antonio Antetomaso[2]

[1]Fondazione Bruno Kessler, via Sommarive 18, Trento, 38123, Italy
[2]Istituto Poligrafico e Zecca dello Stato, via Salaria 691, Roma, 00138, Italy

Abstract
This paper presents the current status of the project established between Istituto Poligrafico e Zecca dello Stato and Fondazione Bruno Kessler, aimed at creating a machine learning-based pipeline for the automatic classification of legislative acts in the Gazzetta Ufficiale, the official source of legislative information of the Italian state. The paper outlines the project's goals, introduces the annotated corpus of legislative acts used as dataset and presents the ongoing results for the task of sentence classification.

Keywords
Legal NLP, Text Classification, BERT, AI for Public Administration

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 organ plays a key role in Italian lawmaking, as for any legislative measure to enter into effect, its publication in this press organ 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*. One of its duties is also the index-based classification of all provisions that are published in the *Gazzetta*, a task that has so far been carried out manually by the Institute's staff.

As part of a more general effort towards digitisation of the Italian public administration, a collaboration has been established in 2021 between IPZS, the data provider, and Fondazione Bruno Kessler (FBK) as scientific partner and consultant. The aim of the partnership is two-fold: firstly, providing support for the adoption of state-of-the-art machine learning technologies in IPZS's workflow, with the specific goal of automatising the classification process of Italian legislative acts in the GU. Second, guiding the migration of IPZS towards the adoption of the EuroVoc[3] multilingual thesaurus, the standard framework in use in the European institutions.

This paper presents the current state of the project, focusing on the first of the above tasks, i.e. on the development of an automatic pipeline for the classification of GU documents. We first introduce the Italian thematic index and its features (Section 2), then we describe the textual dataset of legislative acts being used for automatic text classification (Section 3). Section 4 presents the first results and evaluation for the task, while in Section 5 we outline the future directions to be pursued within the scope of the project.

2. Resources

2.1. Gazzetta Ufficiale Index

As mentioned in the Introduction, each Italian legislative act entering the GU has been manually assigned one or more labels from a subject index to classify its semantic content. Such labels are then used by IPZS 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, which have been trained by IPZS staff.

Table 1
Structure of the GU Dictionary

	Layer	Number of unique labels
Main labels	`Voci Aperte`	781
Secondary labels	`Voci Chiuse`	1,019
	`Riferimenti`	149
	`Sommarietti`	71

Table 1 summarizes the structure of the thematic index. The resource is organized on four levels: Voci Aperte (Open Labels), Voci Chiuse (Closed Labels), Riferimenti

Ital-IA 2023: 3rd National Conference on Artificial Intelligence, organized by CINI, May 29–31, 2023, Pisa, Italy
*Corresponding author.
© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

[1]https://www.gazzettaufficiale.it/
[2]https://www.ipzs.it/ext/index.html
[3]https://op.europa.eu/en/web/eu-vocabularies/dataset/-/resource?uri=http://publications.europa.eu/resource/dataset/eurovoc

(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.

3. Dataset

The publication of acts in the GU is structured in six "Series": 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. 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. Table 2 reports the data distribution in the current dataset with respect to the different series in GU. For the time being, the project focuses on the General Series and the first three Special Series.

Table 2
Composition of the dataset.

Name	Institution/Topic	samples
General Series	Central/Periph. Adm.	363,989
Special Series 1	Constitutional Court	16,485
Special Series 2	European Community	26,926
Special Series 3	Regional Affairs	2,285
Special Series 4	Public exams	–
Special Series 5	State contracts	–

Given the high number of legislative acts contained in General Series compared to the other series, we select this dataset for our experiments, in particular all the legislative acts in the General Series published from 1988, when manual classification began, to 2021. Moreover, since each act consists of a title and a body, the dataset has been built up by extracting only the title of each act. This choice was made because titles seem to be expressive enough to concisely convey the main content of the legislative act. On the contrary, acts can vary a lot in terms of structure and length, sometimes exceeding the max length supported by BERT-like models. Therefore, we do not process the whole body for the moment, and we may leave this for future experiments. In terms of classification, each act 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. The General Series, we focus on, has a tagset of 877 unique labels and an average of 1.5 labels per sample. As in most of the datasets generated based on real-world data and over a long period of time, the distribution of labels is extremely unbalanced[4] and shows a Zipf-like trend, with very few labels counting tens of thousands of assignments and a long queue of labels with few or very few occurrencies.

4. Experiments and Evaluation

From a machine learning perspective, this text classification task is 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. The classifier must therefore learn both which and how many labels must be assigned to each datapoint. In order to provide a robust training and evaluation workflow, a stratified split approach has been chosen 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. Next, two sets of experiments have been conducted on the Serie Generale dataset, the first using FastText (static) embeddings [1] and the second by fine-tuning the Italian version of BERT transformer model [2], which was pre-trained on Wikipedia and other web corpora for a total of 81GB and 13,138,379,147 tokens.[5] The FastText model is the result of 100 training epochs with a learning rate of 5e-2, whereas the Italian BERT model has been fine-tuned for 6 epochs, with a learning rate of 5e-5 and batch size of 16. The results on the development and test sets are

Table 3
Performance results for automatic text classification on the Serie Generale (titles only).

Metric	FastText		BERT	
	dev	test	dev	test
Precision (micro)	0.827	0.791	0.851	0.816 (+0.024)
Recall (micro)	0.816	0.803	0.839	0.832 (+ 0.029)
F1 (micro)	0.822	0.797	0.850	0.824 (+0.027)
F1 (macro)	0.289	0.271	0.300	0.285 (+0.014)
F1 (weighted)	0.802	0.779	0.832	0.807 (+0.028)
Roc Auc	0.906	0.899	0.924	0.916 (+0.017)

[4] In the General Series, we observe an average of 568 occurrencies per label, with a median of 42 and a standard deviation of 3201.
[5] https://huggingface.co/dbmdz/bert-base-italian-xxl-uncased

showed in Table 3. The comparison between the two systems shows that pre-trained transformer models perform consistently better over all the considered metrics. However, this performance gain comes at a cost in terms of computation: with FastText, a single training session requires approximately 40 minutes, using 6 cores on a CPU, whereas the BERT-based model requires about 3 hours on a GPU. This trade-off is of particular importance with regard to the adoption of the systems described in the public administration workflow and in the perspective of its use in the medium to long term.

5. Conclusions and Future Work

This paper discussed the current state of the project established between Istituto Poligrafico e Zecca dello Stato and Fondazione Bruno Kessler for building an automatic text classification system for the Gazzetta Ufficiale. The paper first introduced the textual corpus of legal acts from the *Gazzetta*, then presented the evaluation of two automatic classification algorithms, the first based on static word embeddings, the second on pre-trained transformers. Currently, the technological solutions discussed above are being extended to the three Special Series. The next phase of the project involves the development of semi-automatic tools to support the migration of IPZS from the Italian subject index to the European multilingual standard EuroVoc.

References

[1] A. Joulin, E. Grave, P. Bojanowski, T. Mikolov, Bag of tricks for efficient text classification, arXiv preprint arXiv:1607.01759 (2016).

[2] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, 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), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423. doi:10.18653/v1/N19-1423.