Title is (Not) All You Need for EuroVoc Multi-Label Classification of European Laws

Lorenzo Bocchi[1,†], Alessio Palmero Aprosio[1,*,†]

[1] University of Trento, Italy

Abstract

Machine Learning and Artificial Intelligence approaches within Public Administration (PA) have grown significantly in recent years. Specifically, new guidelines from various governments recommend employing the EuroVoc thesaurus for the classification of documents issued by the PA. In this paper, we explore some methods to perform document classification in the legal domain, in order to mitigate the length limitation for input texts in BERT models. We first collect data from the European Union, already tagged with the aforementioned taxonomy. Then we reorder the sentences included in the text, with the aim of bringing the most informative part of the document in the first part of the text. Results show that the title and the context are both important, although the order of the text may not. Finally, we release on GitHub both the dataset and the source code used for the experiments.

Keywords

EuroVoc taxonomy, Sentence reordering, Text classification

1. Introduction

The presence of Machine Learning and Artificial Intelligence techniques has become almost ubiquitous in many fields, from hobbyist projects to industrial and government usage. Also inside the Italian Public Administration, there have been efforts to digitize and modernize the processes for more than a decade. In particular, some documents released by the Italian PA suggest the use of EuroVoc,[1] a multilingual thesaurus developed and maintained by the Publications Office of the European Union (EU), that covers a wide range of subjects (law, economics, environment, ...) organized hierarchically. Outside Italy, Portuguese [1] and Croatian [2] communities are making efforts to automatically perform tagging of official regulations using EuroVoc. In addition to that, in 2010 the EU organized in Luxembourg the Eurovoc Conference,[2] in order to facilitate the comprehension and use of the taxonomy.

The classification of a document with respect to the EuroVoc taxonomy has previously been addressed by several studies (see Section 2), since at present the classification of the documentation in the PA is carried out manually, a task that can be very expensive in the long run.

In this context, we concentrate our work on automatically assigning EuroVoc labels to a document, starting from the existing approaches in document and text classification, that use pretrained large language models followed by a fine-tuning phase on a specific task. Unfortunately, these families of language models have an intrinsic limit regarding the maximum number of words present in a text (usually 512). In the case of documents that can be quite large, like legal ones, it is important to try and make sure that the key information about a text is included in the chosen set of words. The previous research deals with this limit by concatenating the title with the raw text, and then clipping it to the limit.

In some countries (such as Italy, see [3]) the title is usually very well formulated and it is very important to correctly classify a document. On the contrary, the text of a law is usually very redundant, and the most representative text is often after a notable sequence of preambles.

Given these premises, we investigate how the previous approaches work on European laws and apply different strategies to create a summarized version of a text by reordering the sentences. The results show that in this specific case, both the title and the context are important, and that the best approach in regulations enacted by the European Parliament is to fill the 512-words limit with as much information as possible.

The paper is structured as follows: Section 2 will expose the related work; Section 3 describes the data; the approach and the experiments are described in Section 4; the results are then discussed in Section 5.

Finally, both the software and the dataset are available for download, as described in Section 6.

CLiC-it 2024: Tenth Italian Conference on Computational Linguistics, Dec 04 — 06, 2024, Pisa, Italy

[*] Corresponding author.
[†] These authors contributed equally.
lorenzo.bocchi@unitn.it (L. Bocchi); a.palmeroaprosio@unitn.it (A. Palmero Aprosio)
0000-0002-1484-0882 (A. Palmero Aprosio)

[1] https://bit.ly/eurovoc-ds
[2] https://bit.ly/eurovoc-conference

2. Related work

There have been a number of studies that explored the classification of European legislation with EuroVoc labels.

JRC EuroVoc Indexer [4] is a tool that allows the categorization of documents with EuroVoc classifiers in 22 languages. The data used is contained in an old dataset [5] with documents up to 2006. The algorithm used involves generating a collection of lemma frequencies and weights. These frequencies are associated with specific descriptors, referred to as associates or topic signatures in the paper. When classifying a new document, the algorithm selects the descriptors from the topic signatures that exhibit the highest similarity to the lemma frequency list of the new document.

The research described in [6] explored the usage of Recurrent Neural Networks on extreme multi-label classification datasets, including RCV1 [7], Amazon-13K [8], Wiki-30K and Wiki-500K [9], and an older EUR-Lex dataset from 2007 [10].

In [11] the authors explore the usage of different deep-learning architectures. Furthermore, the authors also released a dataset of 57,000 tagged documents from EUR-Lex.

There are also other monolingual studies on the topic, that mainly concentrate on Italian [12], Croatian [13], and Portuguese [1].

More recent works on multi-language classification on EuroVoc are described in Chalkidis et al. [14], Shaheen et al. [15], and Wang et al. [16].

3. Dataset

3.1. EUR-Lex

The primary source for European legislation is EUR-Lex[3], a web portal offering comprehensive access to EU legal documents. It is available in all 24 official languages of the European Union and is updated daily by its Publications Office. Most documents on EUR-Lex are manually categorized using EuroVoc concepts.

3.2. EuroVoc

EuroVoc's hierarchical structure is divided into three layers: Thesaurus Concept (TC), Micro Thesaurus (MT, previously known as the "sub-sector" level), and Domain (DO, previously known as the "main sector" level). Each layer contains descriptors for documents, covering a broad range of EU-related subjects such as law, economics, social affairs, and the environment, each at varying levels of detail. The TC level is the foundational layer where all key concepts reside, and documents on EUR-Lex are tagged with labels from this level. Each TC is linked to an MT, which is then part of a specific DO.

The version of EuroVoc used for our studies is 4.17, released on 31st January 2023, containing 7,382 TCs, 127 MTs, and 21 DOs.

3.3. Dataset collection

To collect the documents for our task, we built a set of tools written in Python that can be customized to obtain different subsets of the data (year, language, etc.). In total, after filtering out the documents not tagged with EuroVoc or not containing an easy accessible text (for instance, old documents only available as scanned PDFs), we collect around 1.1 million documents in four languages (English, Italian, Spanish, French).

As a subsequent task, we also removed labels that have been deprecated by the EuroVoc developers throughout the years.[4] Following previous work [11], we also remove labels having less than 10 examples.

Finally, by looking at the data, we see that the labelling became consistent starting from 2004, while many deprecated labels are still present in documents, especially previous to 2010. We therefore consider only documents published in the interval 2010-2022.

The final dataset will consist of 471,801 documents. On average, each law is labelled with 6 EuroVoc concepts. Table 1 shows some statistics about the dataset used.

4. Experiments

In this Section, we describe the experiments performed on the above-described data.

4.1. Data split

To keep our experiments consistent with previous similar approaches [17], we split the data into train, dev, and test sets with an approximate ratio of 80/10/10 in percentage, respectively.

In order to make the training reproducible and to avoid that a single random extraction could be too (un)lucky, we repeat the split using three different seeds and a pseudo-random number generator.

Each partition into train/dev/test is done using Iterative Stratification [18, 19], in order to preserve the concept balance.

Unless differently specified, all the results in the rest of the paper refer to the average of the values obtained by our experiments on the three splits.

[3] https://eur-lex.europa.eu/

[4] https://bit.ly/eurovoc-handbook

	English	Italian	Spanish	French
Total documents	195,236	177,952	178,444	183,068
Documents with text and EuroVoc labels	118,296	117,711	117,882	117,912
Number of EuroVoc labels used before filtering	6,098	6,088	6,098	6,088
Number of EuroVoc labels having less than 10 documents	2,070	2,077	2,070	2,070
Final number of labels	4,028	4,011	4,028	4,018
Removed documents	3	3	3	3

Table 1
Number of documents in English, Italian, Spanish, and French relative to the time interval 2010-2022.

4.2. Methodology

Our models are trained using BERT [20] and its derivatives.

The choice of the best pre-trained model is very important for the accuracy of the classification using the model obtained after fine-tuning. In particular, [21] shows that classification tasks over the legal domain obtain better performance when pre-trained on legal corpora. Nevertheless, in some preliminary experiments, we have tried BERT models pre-trained on various datasets (among them, legal ones of course), but not always the results award models built from legal texts.

Although the difference was not statistically significant, we decided to use these models anyway (from HuggingFace[5]):

- `legal-bert-base-uncased` [22], consisting of 12 GB of diverse English legal text from several fields (e.g., legislation, court cases, contracts) scraped from publicly available resources;
- `bert-base-italian-xxl-cased` [23], the main Italian BERT model, consisting of a recent Wikipedia dump and various texts from the OPUS corpora collection[6] and data from the Italian part of the OSCAR corpus;[7]
- `bert-base-spanish-wwm-cased` [24], also called BETO, is a BERT model trained on a big Spanish corpus[8] that consists of 3 billion words;
- `camembert-base` [25], a state-of-the-art language model for French based on the RoBERTa model [26].

4.3. Basic configurations

The basic configurations consist of using the sole title, the sole text, and the concatenation of the title and the text. Note that, apart from some rare outliers, title length is consistently less than 50 tokens.

4.4. Pre-processing

The text of the laws is preprocessed using spaCy,[9] a Natural Language Processing pipeline that can extract information from texts in 24 languages. In particular, we used it to perform sentence splitting part-of-speech tagging, and named-entities recognition, used to extract content words from the text and perform the selection of the sentences that are used in the task.

4.5. Summarization

Given that the input length for these BERT models is 512 tokens, while legislative texts are usually longer, summarizing the text by using the most important parts of it to make sure it fits in the input was seen as an important step to follow.

As underlined in the Introduction, the text of a law is usually very redundant, and its most representative part is often after a notable sequence of preambles.

Since the limit of 512 tokens is very strong if compared to the usual length of a legal document, we concentrate our summarization effort on reordering the sentences inside a single document so that the most informative part of the text can be brought to the beginning and therefore included in the first 512 tokens.

We use two different approaches to reach the goal: TF-IDF and centroid-based. In both cases, we perform training with the sole text reordered and the concatenation of the title and the above text.

4.5.1. TF-IDF

TF-IDF (Term Frequency-Inverse Document Frequency) is a widely used technique in information retrieval and text mining to quantify the importance of terms in a document within a larger collection of documents. It aims to highlight terms that are both frequent within a document and relatively rare in the overall collection, thus capturing their discriminative power.

The TF-IDF score of a term in a document is calculated by multiplying two factors: the term frequency (TF) and

[5] https://huggingface.co/
[6] http://opus.nlpl.eu/
[7] https://traces1.inria.fr/oscar/
[8] https://bit.ly/big-spanish-corpora
[9] https://spacy.io/

the inverse document frequency (IDF).

Let t be the term and d the document:

$$\begin{aligned}\mathrm{tf}(t,d) &= \frac{f_{t,d}}{\sum_{t'\in d} f_{t',d}}\\ \mathrm{idf}(t,D) &= \log\frac{N}{1+|\{d\in D: t\in d\}|}\end{aligned}$$

where $f_{t,d}$ is the frequency of term t in document d, and $N = |D|$ is the number of documents in the set D.

Beyond the usual TF-IDF, we also perform a label-based approach, that considers one document for each label, by concatenating all the texts belonging to the laws having that label.

Once all the documents have gone through this process, the TF-IDF matrix is calculated using TfidfVectorizer from the Python package scikit-learn[10] over the content words (see Section 4.4) of the texts.

After obtaining the TF-IDF matrix, the final step is to assign a score to each sentence. For each valid base form, its score is determined from the TF-IDF matrix by selecting the highest value within the corresponding column (which represents a word). These scores are then added to a list for each sentence. Once a sentence is processed, the maximum or average score is calculated ("max" and "mean" in the results). This calculated value becomes the sentence's score. The process is repeated for all sentences in every document.

4.5.2. Centroid

In this approach, described in [27], the centroid of the word vectors in the text is calculated, then a score is assigned to each sentence based on their cosine distance from the centroid. The closer a sentence is to the centroid, the higher the score it will receive. In our approach, we use fastText [28] for word embeddings.

The words used to compute the centroid are those that have been extracted as content words (see Section 4.4) and have a TF-IDF higher than a certain threshold t, which in this case was 0.3. The centroid is computed as the mean of the word embeddings of the previously selected words:

$$C = \frac{\sum_{w\in D_t} E[\mathrm{idx}(w)]}{|D_t|}$$

where D_t is the corpus of words with $\mathrm{tfidf}(w) > t$.

Each sentence in the document gets transformed into a unique embedding representation by averaging the sum of the embedding vectors of each word in the sentence:

$$S_j = \frac{\sum_{w\in S_j} E[\mathrm{idx}(w)]}{|S-j|}$$

where S_j is the j-th sentence in document D.

After obtaining the embedding for the sentence, its score is computed as the cosine similarity between the centroid and the embedding:

$$\mathrm{sim}(C, S_j) = 1 - \frac{C^T \cdot S_j}{||C|| \times ||S_j||}$$

By using the previously described approach, every text was converted into a list of ranked sentences, each with its own score.

4.6. Random

Because of the obtained results (see Section 4.7), we also added two configurations that used a random ordering of the sentences (one concatenated with the title, the other one containing only the randomly ordered text).

4.7. Evaluation

The evaluation of our experiments is performed by using the F1 score, macro-averaged so that each label has the same weight (this metric awards models that perform better on less-represented labels). Since we are dealing with a multi-label classification task, we have to choose between considering always the same number K of results ($P@K$, $R@K$, $F1@K$) or keeping only the labels whose confidence is higher than a particular threshold (usually between 0 and 1). In our experiments, we chose the second approach, since the number of concepts in each document of the dataset is not constant. Given the evaluation performed on the development set, we set that threshold to 0.5.

4.8. Results

Table 2 shows the results of the different configurations in the four languages. The first column contains the description of the experiment, while columns TC, MT, and DO show the result in terms of Thesaurus Concept (TC), Micro Thesaurus (MT), and Domain (DO), as described in Section 3.

5. Discussion

Results show that the best performances are reached when the title is included in the text (see the rows without "not") with the exception brought by the simple use of the text without reordering. An interesting outcome is that the experiment using title+random obtains very good results when compared to the best configurations.

On the contrary, using random text without the title, or using the sole title results in a decrease in global performance.

[10] https://scikit-learn.org

	English			Italian			Spanish			French		
	TC	MT	DO	TC	MT	DO	TC	MT	DO	TC	MT	DO
basic	0.484	0.729	0.812	0.450	0.709	0.798	0.493	0.732	0.818	0.383	0.666	0.775
basic-not	0.474	0.722	0.808	0.453	0.710	0.799	0.483	0.726	0.811	0.370	0.655	0.765
centroid	0.468	0.720	0.806	0.454	0.710	0.799	0.479	0.719	0.810	0.372	0.658	0.764
centroid-not	0.426	0.692	0.784	0.405	0.673	0.774	0.430	0.687	0.784	0.335	0.627	0.745
title-only	0.432	0.682	0.772	0.407	0.665	0.758	0.444	0.684	0.771	0.320	0.600	0.716
tfidf-max-doc	0.476	0.724	0.811	0.427	0.693	0.788	0.459	0.711	0.804	0.345	0.642	0.754
tfidf-max-lab	0.477	0.728	0.812	0.459	0.711	0.802	0.483	0.724	0.813	0.378	0.660	0.767
tfidf-mean-doc	0.479	0.726	0.812	0.427	0.693	0.786	0.484	0.726	0.812	0.381	0.663	0.774
tfidf-mean-lab	0.481	0.726	0.813	0.428	0.693	0.788	0.485	0.726	0.813	0.338	0.633	0.749
tfidf-max-doc-not	0.427	0.692	0.787	0.379	0.657	0.763	0.422	0.682	0.786	0.301	0.607	0.726
tfidf-max-lab-not	0.433	0.696	0.791	0.411	0.678	0.779	0.425	0.685	0.782	0.298	0.608	0.728
tfidf-mean-doc-not	0.433	0.696	0.790	0.415	0.682	0.781	0.442	0.700	0.796	0.332	0.626	0.742
tfidf-mean-lab-not	0.436	0.697	0.792	0.388	0.667	0.771	0.428	0.684	0.784	0.296	0.598	0.723
random	0.472	0.722	0.808	0.423	0.692	0.787	0.482	0.723	0.807	0.372	0.652	0.767
random-not	0.429	0.693	0.788	0.398	0.671	0.774	0.439	0.693	0.778	0.318	0.611	0.724

Table 2
Results of our experiments (macro F_1).

By looking at the statistical significance,[11] we find out that we can split, more or less, the experiments into two big groups: the ones that in the English part of the table have a DO F_1 above 0.80 and the remaining ones that are below 0.79. The exception is the "title-only" configuration, which obtains lower accuracy in all languages and contrasts with the results obtained in a similar previous work applied to Italian laws [3], where the use of the sole title results in an increase in performance with respect to the concatenation between title and text.

By listing the documents where EuroVoc labels are not extracted correctly, it seems that in the European legislation it is quite common to find very generic titles. For instance, the title of the document with ID "CELEX:32011Q0624(01)" is "Rules of procedure for the appeal committee (Regulation (EU) No 182/2011)", from which is very hard to extract relevant information about the topic. One can find other similar documents, such as "Action brought on 2 March 2011 — Attey v Council", title of law with ID "CELEX:62011TN0118".

In general, our experiments show that the classification of European laws obtains the best performance on BERT when all the possible tokens are filled, possibly using the title and some parts of the text. The high accuracy obtained in the experiments performed by randomly reordering the sentences demonstrates that the context is important per se, even when no particular strategies are used to select it.

French results bring significantly lower accuracy: this is not expected and is probably due to the choice of the BERT pre-trained model.

6. Release

The source code for all the experiments (from the retrieval of the documents to the training of the models), the data downloaded from EUR-Lex, and the models are available on the project Github page.[12]

7. Conclusions and Future Work

In this paper, we presented some approaches to perform document classification on long documents, by reordering their sentences before the fine-tuning phase. The best results are obtained when all the 512 tokens allowed in the BERT paradigm are filled, possibly including the title of the law.

In the future, we want to extend this approach to other languages, trying to understand whether the same reordering algorithm leads to some improvement in the classification task. We will also investigate other summarization approaches, or new architectures that rely on Local, Sparse, and Global attention [29] so that longer texts (up to 16K tokens) can be used to train the model.

[11] To calculate statistical significance, a one-tailed t-test with a significance level of .05 was applied to the scores of the five runs, with the null hypothesis that no difference is observed, and the alternative hypothesis that the score obtained with the summarized text is significantly greater than the one with the normal text.

[12] https://github.com/bocchilorenzo/AutoEuroVoc

References

[1] D. Caled, M. Won, B. Martins, M. J. Silva, A hierarchical label network for multi-label eurovoc classification of legislative contents, in: Digital Libraries for Open Knowledge: 23rd International Conference on Theory and Practice of Digital Libraries, TPDL 2019, Oslo, Norway, September 9-12, 2019, Proceedings, Springer-Verlag, Berlin, Heidelberg, 2019, p. 238–252. URL: https://doi.org/10.1007/978-3-030-30760-8_21. doi:10.1007/978-3-030-30760-8_21.

[2] T. D. Prekpalaj, The role of key words and the use of the multilingual eurovoc thesaurus when searching for legal regulations of the republic of croatia - research results, in: 2021 44th International Convention on Information, Communication and Electronic Technology (MIPRO), 2021, pp. 1470–1475. doi:10.23919/MIPRO52101.2021.9597043.

[3] M. Rovera, A. P. Aprosio, F. Greco, M. Lucchese, S. Tonelli, A. Antetomaso, Italian legislative text classification for gazzetta ufficiale (2022).

[4] R. Steinberger, M. Ebrahim, M. Turchi, Jrc eurovoc indexer jex-a freely available multi-label categorisation tool, arXiv preprint arXiv:1309.5223 (2013).

[5] R. Steinberger, B. Pouliquen, A. Widiger, C. Ignat, T. Erjavec, D. Tufiş, D. Varga, The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages, in: Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06), European Language Resources Association (ELRA), Genoa, Italy, 2006. URL: http://www.lrec-conf.org/proceedings/lrec2006/pdf/340_pdf.pdf.

[6] R. You, Z. Zhang, Z. Wang, S. Dai, H. Mamitsuka, S. Zhu, Attentionxml: Label tree-based attention-aware deep model for high-performance extreme multi-label text classification, Advances in Neural Information Processing Systems 32 (2019).

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

[8] J. McAuley, J. Leskovec, Hidden factors and hidden topics: Understanding rating dimensions with review text, in: Proceedings of the 7th ACM Conference on Recommender Systems, RecSys '13, Association for Computing Machinery, New York, NY, USA, 2013, p. 165–172. URL: https://doi.org/10.1145/2507157.2507163. doi:10.1145/2507157.2507163.

[9] A. Zubiaga, Enhancing navigation on wikipedia with social tags, arXiv preprint arXiv:1202.5469 (2012).

[10] E. Loza Mencía, J. Fürnkranz, Efficient multilabel classification algorithms for large-scale problems in the legal domain, 2010. URL: http://dx.doi.org/10.1007/978-3-642-12837-0_11. doi:10.1007/978-3-642-12837-0_11.

[11] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, I. Androutsopoulos, Large-scale multi-label text classification on eu legislation, arXiv preprint arXiv:1906.02192 (2019).

[12] G. Boella, L. Di Caro, L. Lesmo, D. Rispoli, Multi-label classification of legislative text into eurovoc, Legal Knowledge and Information Systems: JURIX 2012: the Twenty-Fifth Annual Conference 250 (2013) 21. doi:10.3233/978-1-61499-167-0-21.

[13] F. Saric, B. D. Basic, M.-F. Moens, J. Šnajder, Multi-label classification of croatian legal documents using eurovoc thesaurus, 2014.

[14] I. Chalkidis, M. Fergadiotis, I. Androutsopoulos, 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, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 6974–6996. URL: https://aclanthology.org/2021.emnlp-main.559. doi:10.18653/v1/2021.emnlp-main.559.

[15] Z. Shaheen, G. Wohlgenannt, E. Filtz, Large scale legal text classification using transformer models, 2020. arXiv:2010.12871.

[16] L. Wang, Y. W. Teh, M. A. Al-Garadi, Adopting the multi-answer questioning task with an auxiliary metric for extreme multi-label text classification utilizing the label hierarchy, 2023. arXiv:2303.01064.

[17] A. Avram, V. F. Pais, D. Tufis, Pyeurovoc: A tool for multilingual legal document classification with eurovoc descriptors, CoRR abs/2108.01139 (2021). URL: https://arxiv.org/abs/2108.01139. arXiv:2108.01139.

[18] K. Sechidis, G. Tsoumakas, I. Vlahavas, On the stratification of multi-label data, Machine Learning and Knowledge Discovery in Databases (2011) 145–158.

[19] P. Szymański, T. Kajdanowicz, A network perspective on stratification of multi-label data, in: L. Torgo, B. Krawczyk, P. Branco, N. Moniz (Eds.), Proceedings of the First International Workshop on Learning with Imbalanced Domains: Theory and Applications, volume 74 of *Proceedings of Machine Learning Research*, PMLR, ECML-PKDD, Skopje, Macedonia, 2017, pp. 22–35.

[20] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, CoRR abs/1810.04805 (2018). URL: http://arxiv.org/abs/

1810.04805. arXiv:1810.04805.

[21] I. Chalkidis, E. Fergadiotis, P. Malakasiotis, N. Aletras, I. Androutsopoulos, Extreme multi-label legal text classification: A case study in EU legislation, in: Proceedings of the Natural Legal Language Processing Workshop 2019, Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 78–87. URL: https://aclanthology.org/W19-2209. doi:10.18653/v1/W19-2209.

[22] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, I. Androutsopoulos, LEGAL-BERT: The muppets straight out of law school, in: Findings of the Association for Computational Linguistics: EMNLP 2020, Association for Computational Linguistics, Online, 2020, pp. 2898–2904. URL: https://aclanthology.org/2020.findings-emnlp.261. doi:10.18653/v1/2020.findings-emnlp.261.

[23] S. Schweter, Italian bert and electra models, 2020. URL: https://doi.org/10.5281/zenodo.4263142. doi:10.5281/zenodo.4263142.

[24] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, J. Pérez, Spanish pre-trained bert model and evaluation data, in: PML4DC at ICLR 2020, 2020.

[25] L. Martin, B. Muller, P. J. O. Suárez, Y. Dupont, L. Romary, É. V. de la Clergerie, D. Seddah, B. Sagot, Camembert: a tasty french language model, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020.

[26] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized BERT pretraining approach, CoRR abs/1907.11692 (2019). URL: http://arxiv.org/abs/1907.11692. arXiv:1907.11692.

[27] G. Rossiello, P. Basile, G. Semeraro, Centroid-based text summarization through compositionality of word embeddings, in: Proceedings of the multiling 2017 workshop on summarization and summary evaluation across source types and genres, 2017, pp. 12–21.

[28] P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, Enriching word vectors with subword information, Transactions of the association for computational linguistics 5 (2017) 135–146.

[29] C. Condevaux, S. Harispe, Lsg attention: Extrapolation of pretrained transformers to long sequences, in: Pacific-Asia Conference on Knowledge Discovery and Data Mining, Springer, 2023, pp. 443–454.