Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 3670–3674
Marseille, 20-25 June 2022

Identifying Draft Bills Impacting Existing Legislation: a Case Study on Romanian

Corina Ceaușu, Sergiu Nisioi
Faculty of Mathematics and Computer Science,
University of Bucharest
Academiei, 14, Bucharest
corina.ceausu16@gmail.com, sergiu.nisioi@unibuc.ro

Abstract

In our paper, we aim to build the necessary resources to investigate the automatic categorization of draft bills that have the potential to impact previous Romanian legislation. For this purpose, we collect a historical corpus of laws related to public procurement legislation and a corpus of draft bills that have been idenitifed by legal experts as impacting existing policies on public procurement. Our results show that legal articles can be identified with as much as 82% accuracy using a BERT-based classifier and 73% with xgboost and a bag-of-words representation.

Keywords: legal texts processing, Romanian corpora, text classification, text mining

1. Introduction

Contemporary democracies implement elaborate processes to draft, revise, and approve legislative proposals (bills) and several dozens of experts from different areas are employed to reason upon the impact of a new proposal (Fitsilis, 2021). Particularly, in Romania, according to the state Constitution, a new bill initiative may be proposed either by the Government, the Members of the Romanian Parliament, or citizens who have at least 10,000 signatures of popular support for the respective bill[1]. The Romanian Parliament is composed of the Chamber of Senate and the Chamber of Deputies where legal proposals are being debated, amended and voted upon.

The legislating process may be more elaborate, however, during the course of this paper we will assume the following steps: 1) a bill is first submitted to the Senate where it will be subjected to debate, amendments, and voting. If it passes the majority vote, 2) the bill will be sent to the Chamber of Deputies to go through the same process again. Finally, 3) if the bill passed the majority vote yet again, the Romanian President has the power to promulgate the bill or return it to the Parliament.

Both parliament's chambers have specialized committees of experts covering the majority of public policy interests in a state, these range from finance, transportation, human rights, health, and industry to agriculture, education, science, equality, and sports. To this date there are sixteen expert committees of the Senate and twenty five similar organizations at the Chamber of Deputies. Depending on the type of bill, several committees might be called in for a written opinion on the proposal. For example, a bill that regulates the financing of investments in tourism might require the advice from the Senate committees responsible for Industry, Finance, Public Administration, and Public Health. After being approved by the Senate, it will be passed to the Chamber of Deputies where yet again several committees will be asked for advice, e.g., The Committee for the Public Budget, for Work and Social Protection, and another one for Health.

In addition to these specialized committees, there are national institutions and agencies responsible for the legislation and audits in very particular aspects. One such example is The National Agency for Public Procurement (NAPP)[2] who is responsible to ensure the quality of public spending and to oversee the management of public investments. The institution itself is subordinated to the Government and has the ability to audit, propose, and update the legal apparatus related to public procurement. When a new bill is clearly aimed at changing the procurement legislation, the agency's advice is required at the Parliament.

Another responsibility of the agency is to identify legislative proposals that may impact or contradict, oftentimes unintentionally, the existing public procurement legislation. Bills may contain articles that implicitly contradict the legislation without having any specific mentions regarding "public procurement". In order to identify these proposals, several teams of experts are monitoring on a daily basis the draft bills and their amendments debated at each parliamentary committee. Given that draft bills may range from tens to several hundreds of pages, a considerable amount of human effort is required to keep track of legislative changes and constantly read all the articles, identify the parts that impact or contradict existing legislation, and create structured reports.

In our paper, we aim to create and release 1. a diachronic dataset comprising of Romanian public procurement legislation that has been published over time, 2. an annotated dataset of draft articles that have impacted the legislation at the time of being debated, and

[1] http://www.cdep.ro/pls/dic/site.page?id=124

[2] http://anap.gov.ro/web/en

3. to propose an automatic method of identifying articles that impact/ contradict public procurement legislation, existent up to a certain point in time. This effort will not be able to replace the expert human work needed to review the bills, however we believe that it has the potential to become the building block of a regulatory impact assessment tool (Leventis et al., 2021) that can help experts reduce the time to review a proposal and to facilitate the identification of contradicting or impactful draft articles.

2. Data Collection

Historical Corpus of Public Procurement Legislation

In order to identify laws that impact existing public procurement legislation, first we created a corpus of laws that have been published over time with the aim to regulate public procurement. We call the corpus "historical" because a big majority of the laws collected have been changed in the meantime and they may not be valid in the present. The selection of the laws has been done semi-automatically in three stages:

1. a manual selection, according to expert guidance, of acts that encompass the currently active public procurement legislation
2. a web scrape on all the legal acts authored by the National Agency of Public Procurement
3. a scrape of all the historical legislation that contains the corresponding lemmatized words for "public procurement"; this part we further cleaned out manually by removing the proposals that were not strictly related to the procurement regulations

To obtain the scraped documents, we follow the same methodology used for building MARCELL Romanian legislative subcorpus (MARCELL Consortium, 2021), by crawling the public Romanian legislative portal[3] which provides free access to all the legal documents issued since 1881 (Váradi et al., 2020). Any HTML artifacts are removed to preserve only the textual form of the law. For each scraped document we give special attention to the date and year when the act was officially published. The historical date is to be used in the experiments to ensure the avoidance of future information leaking into the past, e.g., for the analysis of a draft bill from May, 2018 we only compare it against older documents, since more recent published legislation may already contain parts of the bill we are analyzing. Additionally, we create a tabular data file that stores various metadata covering the publishing year of each document, the legislation header that contains a summary of the main topics covered by the respective act, the source URL, and whether from a legal perspective it is considered primary legislation, i.e., derived from an EU treaty and passed by the main legislative bodies or secondary legislation, i.e., derived from the primary legislation.

	# words	# types	# articles
Public proc.	1,000,000	18,000	4,307
Impacting	39,000	4,500	124
Not-impacting	300,000	15,000	1,464

Table 1: Statistics regarding the size of each subcorpus. The first row is extracted from the historical corpus of public procurement legislation. The second and third rows indicate the manually annotated articles.

Several statistics regarding the number of articles and the size in words of this corpus are visible in the first row of Table 1.

Historical Corpus of Draft Bills

The National Agency for Public Procurement provides yearly reports with statistics in relation to the number of bills manually reviewed. The reports we use include past draft bills ranging from 2016 to 2019 that have been through various amendments at the Chamber of Senate and at the Chamber of Deputies for debate. The reports were converted into structured data by the following process:

1. parse the unstructured text to extract the proposal identification number that are given by the Senate (ids starting with letter "L") and the ones given by the Chamber of Deputies (ids starting with "PLX")
2. identify and build the URLs pointing to the proposal's web pages at the Senate[4] and at the Chamber of Deputies [5]
3. crawl all the variants of the bill, the additional documents, annexes, and written opinions that are part of the bill's folder from both chambers
4. the majority of the institutions of the Romanian Parliament work with scanned signed PDF files that have to be converted into text through OCR, therefore, for each PDF file, we applied the latest version of Tesseract OCR[6] for Romanian language; several preprocessing improvements have been applied with Tesseract using Twin Delayed Deep Deterministic Policy Gradient, as suggested by (Sporici et al., 2020)
5. we manually processed each expert report and annotated the actual articles and paragraphs from the bill that have been classified as impacting existing legislation

[3] http://legislatie.just.ro

[4] https://www.senat.ro

[5] http://www.cdep.ro

[6] https://github.com/tesseract-ocr/tesseract/releases/tag/5.0.0

6. each bill is segmented into articles using a simple regular expression heuristic and each article is labeled impacting or not-impacting

We manually re-check all the different forms of each bill to ensure the dataset is properly annotated. In total we selected 66 legislative proposals each with several impacting articles, in total summing up to 124 annotated articles.

Several bill from our corpus have already been approved and published as laws since the time of being reviewed, and we could use plain text form available on the Romanian legal portal. However, we choose to ignore the public documents and to input the initial forms of the documents that were debated at the time of being reviewed. This process involves re-crawling and applying OCR for each bill, to ensure that the model learns the patterns that are present in OCR-generated noisy data.

We are committed to openly publish our collected data and experiments[7].

3. Experimental Setup

Word Frequency Distributions

To begin, we carry an analysis of word frequency distributions among the two main classes: the *impacting* and *not-impacting* articles from the annotated legislation. We use the shifterator library[8] (Gallagher et al., 2021) to observe the Shannon Entropy shifts between words in the two classes. The Shannon Entropy for an entire set corpus is computed by the formula

$$H(P) = \sum_i p_i \log \frac{1}{p_i} \tag{1}$$

where p_i is the probability of word i in a corpus approximated by the relative frequency. The information Entropy function is correlated with the idea that low frequency words are more surprising to be seen in a text (Gallagher et al., 2021).

To compare the frequency shifts between the two corpora, we compute the difference between the information entropy of the word i in the impacting corpus $i^{(2)}$ and the word i in the not-impacting corpus $i^{(1)}$

$$\delta H_i = p_i^{(2)} \log \frac{1}{p_i^{(2)}} - p_i^{(1)} \log \frac{1}{p_i^{(1)}}. \tag{2}$$

In Figure 1, the contribution of each word i is ranked accordingly: on the right-hand side of the image the contribution of words that appear in articles impacting procurement legislation. On the left-hand side of the figure, with dark blue, are the words that have higher scores and are more specific to the not-impacting legislation.

Unsurprisingly, these preliminary results indicate that impacting legislation appears to have words that are

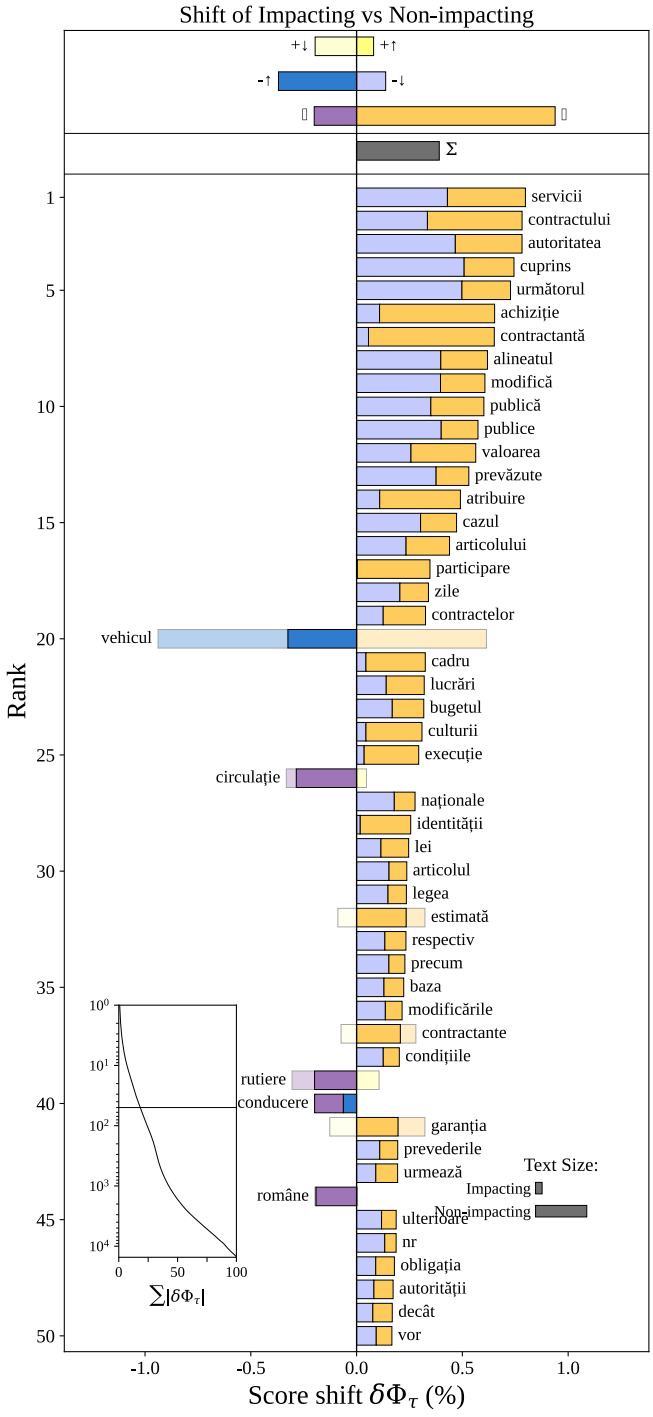

Figure 1: Shanon Entropy word shifts between impacting and not-impacting legislation. With orange the words that are specific to impacting legislation compared to their scores in the not-impacting legislation. The majority of words specific to impacting articles cover the financial semantic field, translated in this order from top to bottom: services, contract, authority, content, next, acquisition, contracting, paragraph, modify, publish, public, value, foreseen, attributed, case, article, participation, days, contracts.

[7] https://github.com/senisioi/rolegal

[8] https://shifterator.readthedocs.io/

	K-fold	K-fold std	LoBo	LoBo std
BERT	**0.86**	0.03	**0.82**	0.113
Nearest Neigh	0.51	0.01	0.5	0.01
XGBoost	0.74	0.071	0.73	0.227
Logistic Regression	0.69	0.046	0.68	0.212
Random Forests	0.58	0.04	0.57	0.174

Table 2: Balanced accuracy results and standard deviation for 5-fold cross-validation and Leave One Bill Out (LoBo) cross validation. The later is using the articles of each bill as a test set and training on the articles of the remaining bills.

more specific to public procurement semantic field, such as: services (*servicii*), contract (*contractului*), authority (*autoritatea*), purchase (*achiziție*), engaged in a contract (*contractanta*), and other words strongly related to financial information, budget, and deadlines.

Detecting Impacting Articles

We experiment and compare several approaches for identifying articles that have the potential to impact existing legislation:

1. fine-tune a BERT-based (Devlin et al., 2019) model on the task of impacting article classification
2. compare each paragraph with existing public procurement articles and label the articles based on the three nearest neighbours
3. train a traditional binary classifier with tf-idf bag of words representation to identify articles impacting vs. not-impacting public procurement legislation

For the first approach, we pre-train a BERT (Devlin et al., 2019) model on the entire Romanian legal corpus (Váradi et al., 2020) for 24 hours on 4 NVidia Tesla SXM2 with 16GB each. We follow the guidelines in (Izsak et al., 2021) and set a maximum sentence length of 512 tokens (since the majority of sentences our legal documents rarely exceed this limit), and a batch size of 4096, with a mini batch per GPU of 128. The weights are pre-initialized from the *bert-base-romanian-cased-v1* model (Dumitrescu et al., 2020) available in the huggingface registry. We use budgeted training (Li et al., 2019) by synchronizing the learning rate to decrease over the entire time window and we use the deep speed (Rasley et al., 2020) library and train with *fp16*. More details regarding the model parameters are available on the huggingface hub, where we released the trained weights[9].

For the second method, we simply compute a 3 nearest neighbour approach using the Jaccard coefficient between the test article and the set of all public procurement documents including the training set and a random subset of legislative texts from the MARCELL corpus (MARCELL Consortium, 2021). To efficiently approximate the Jaccard coefficient, we create a MinHash (Broder, 1997) index across the entire dataset.

The traditional binary classifiers are trained on tf-idf bag of words representations using word unigrams and bigrams and stop words removal. The results reported in Table 2 include the random forests and logistic regression implementations from scikit-learn (Pedregosa et al., 2011) with liblinear back-end (Fan et al., 2008), penalty l_2 and balanced class weights. Last but not least, we use the same document representation with Gradient Boosting Trees booster from XGBoost (Chen and Guestrin, 2016) library.

Given the strong imbalance in our data, we evaluate each model and approach using balanced accuracy metric by doing a classical K-fold cross validation and using a custom Leave One Bill Out (LoBo) cross validation, where we iterate bill by bill and all the articles comprising a bill at a certain point are part of the test set and the remaining articles part of the training set.

The best results in terms of balanced accuracy have been obtained using the BERT-based morel, followed by the tf-idf bag of words representation combined with XGBoost.

The models do not have access to a knowledge base of linked legal documents and neither to the full context for predicting an article, which inherently brings several limitations of such approaches. From the analysis of our results, we observed that:

- we could not observe improvements by removing, tokenizing or handling named entities on the legal domain (Păiș et al., 2021)
- misclassifications may occur because an article may contain references to previous laws
- misclassifications may occur because the relation between a contradicting article and an existing public procurement laws is not obvious only from the lexical occurrences
- wrong labels can be attributed to articles that contain words related to public procurement e.g., articles stating the need to follow the existing policies may be mislabeled as impacting
- given the small amount of data and the large models (either in terms of feature representation or in

[9] https://huggingface.co/snisioi/bert-legal-romanian-cased-v1

terms of hyper-parameters), the model may tend to over-fit sample of impacting articles, potentially finding patterns where there are not

4. Conclusions

Our main results cover the release of a novel dataset for Romanian containing legislative proposals that have contradicted existing legislation at the time of their debate in the Romanian Parliament. We provide an analysis of the data and experiment with several classification and information retrieval models in order to evaluate the ability of automated tools to help experts in the process of legislative screening. Our results indicate certain limitations that we can attribute to the relatively small data size of bills impacting legislation and hint towards potential improvements. We hope that our results will be useful for future work on automatic legislative screening and that it can facilitate the identification of impacting bills beyond the scope of public procurement legislation.

For future work, our aims are to extend the amount of labeled data and to bring forward an analysis of the lexico-grammatical and feature-related reasons for which a certain article is labeled as impacting or not by our models.

5. Bibliographical References

Broder, A. Z. (1997). On the resemblance and containment of documents. In *Proceedings. Compression and Complexity of SEQUENCES 1997 (Cat. No. 97TB100171)*, pages 21–29. IEEE.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (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, June. Association for Computational Linguistics.

Dumitrescu, S. D., Avram, A.-M., and Pyysalo, S. (2020). The birth of romanian bert. *arXiv preprint arXiv:2009.08712*.

Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., and Lin, C.-J. (2008). Liblinear: A library for large linear classification. *the Journal of machine Learning research*, 9:1871–1874.

Fitsilis, F. (2021). Artificial intelligence (ai) in parliaments – preliminary analysis of the eduskunta experiment. *The Journal of Legislative Studies*, 27(4):621–633.

Gallagher, R. J., Frank, M. R., Mitchell, L., Schwartz, A. J., Reagan, A. J., Danforth, C. M., and Dodds, P. S. (2021). Generalized word shift graphs: a method for visualizing and explaining pairwise comparisons between texts. *EPJ Data Science*, 10(1):4.

Izsak, P., Berchansky, M., and Levy, O. (2021). How to train BERT with an academic budget. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10644–10652, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Leventis, S., Fitsilis, F., and Anastasiou, V. (2021). Diversification of legislation editing open software (leos) using software agents—transforming parliamentary control of the hellenic parliament into big open legal data. *Big Data and Cognitive Computing*, 5(3).

Li, M., Yumer, E., and Ramanan, D. (2019). Budgeted training: Rethinking deep neural network training under resource constraints. In *International Conference on Learning Representations*.

Păiș, V., Mitrofan, M., Gasan, C. L., Coneschi, V., and Ianov, A. (2021). Named entity recognition in the romanian legal domain. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 9–18.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Rasley, J., Rajbhandari, S., Ruwase, O., and He, Y. (2020). Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3505–3506.

Sporici, D., Cușnir, E., and Boiangiu, C.-A. (2020). Improving the accuracy of tesseract 4.0 ocr engine using convolution-based preprocessing. *Symmetry*, 12(5):715.

Váradi, T., Koeva, S., Yamalov, M., Tadić, M., Sass, B., Nitoń, B., Ogrodniczuk, M., Pkezik, P., Barbu Mititelu, V., Ion, R., Irimia, E., Mitrofan, M., Păiș, V., Tufiș, D., Garabík, R., Krek, S., Repar, A., Rihtar, M., and Brank, J. (2020). The MARCELL legislative corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3761–3768, Marseille, France, May. European Language Resources Association.

6. Language Resource References

MARCELL Consortium. (2021). *Multilingual Resources for CEF.AT in the legal domain (MARCELL)*. MARCELL Project, distributed via ELRA, MARCELL resources, 2.0.