Natural Language Processing Application in Legislative Activity: a Case Study of Similar Amendments in the Brazilian Senate

Diany Pressato [1], Pedro L. C. de Andrade[1], Flávio R. Junior [2], Felipe A. Siqueira[1], Ellen Polliana R. Souza[1,2], Nádia F. F. da Silva[1,3], Márcio de S. Dias[1,4], and André C. P. L. F. de Carvalho[1]

[1]Institute of Mathematical Sciences and Computation, University of São Paulo (USP), São Paulo, Brazil
[1]{diany_press, pedroandrade, felipe.siqueira, andre}@usp.br
[2]Rural Federal University of Pernambuco, Pernambuco, Brazil
[2]{flavio.rocha, ellen.ramos}@ufrpe.br
[3]Federal University of Goiás, Goiás, Brazil, nadia.felix@ufg.br
[4]Federal University of Catalão, Goiás, Brazil, marciodias@ufcat.edu.br

Abstract

This paper presents an automated approach to organize and analyze legislative amendments documents by utilizing topic-based clustering and retrieval. The system allows legal consultants to associate amendments with predefined topics, improving efficiency in handling a large number of amendments. The study evaluates different retrieval methods based on BM25, a term-matching scoring function, and SBERT architectures, and finds that the BM25L approach performs best in relation to recall metric, particularly when considering the full content of the amendment documents, since an exact match is possible to occur. In addition, this work highlights the importance of preprocessing when employing BM25 methods, since our best results, when taking into account both recall scores and preprocessing computational time, were obtained when applying more preprocessing steps and with the adoption of the RLSP, a rule-based algorithm specifically developed for the Portuguese Language.

1 Introduction

The legislative process comprises the drafting, analysis, and voting of various types of bills. During this process, amendments can be proposed with the aim of modifying or enhancing the original text of the bill by adding, removing, or altering provisions. The proposed changes are subjected to evaluation for their admissibility and are subsequently discussed and voted upon by parliamentarians in both committees and plenary sessions.

As part of its daily activities, the staff of the Brazilian Senate and Chamber of Deputies collects and organizes amendments presented for specific bills. Similar amendments, those applying similar modifications to a law, must be discussed and voted simultaneously. In a short period of time, a large number of amendments can be presented and manually analyzed. Thus, automation tools to speed up the process and improve the service are essential.

In this paper, we present and evaluate an approach where a list of topics is provided for each amendment. In this way, the consultant can associate the amendment with one or more related topics to enhance the amendment approval analysis, since groping them in predefined topics helps the understanding of the proposed changes. For instance, one amendment might suggest a specific minimum age for retirement, while another might conflict by stipulating a different age threshold. By grouping these amendments under the same predefined category, the consultant is better equipped to comprehend these proposed changes and consequently formulate an assessment of the admissibility of these alterations.

We analyze the clustering of similar amendments into predefined topics related to the PEC 6/2019 from the Senate Committee on Constitution, Justice, and Citizenship report [1] [2]. Each topic is represented by a single word or by a small number of words. This research is conducted within the context of the *Ulysses Project* [3], an institutional framework comprising artificial intelligence initiatives aimed at enhancing transparency, fostering improved relations between the government and citizens, and providing complex analysis to support legislative activities.

This paper is organized as follows: Section 2 presents the major related studies. Section 3 details the methods used. Section 4 presents and discusses

[1]https://www12.senado.leg.br/noticias/arquivos/2019/08/27/relatorio

[2]Example of an amendment document: https://legis.senado.leg.br/sdleg-getter/documento?dm=7990869&disposition=inline

[3]https://www.camara.leg.br/noticias/548730-camara-lanca-ulysses-robo-digital-que-articula-dados-legislativos/

the obtained results, and details approaches evaluation. Section 5 brings the conclusion and highlights future works.

2 Related works

(Smywiński-Pohl et al., 2021) describe three strategies to automatically detect amendments in legal texts by performing Named Entity Recognition (NER), treated as a token-classification problem.The BiRNN architecture was remarkable for achieving high values of F1 scores, up to 98.2%.

(Agnoloni et al., 2022) automates tasks to assist the Senate staff in identifying groups of amendments, that were annotated in groups according to their similarity in lexical structure, in order to schedule their simultaneous voting. The authors points Hierarchical Agglomerative Clustering (HAC) as the most appropriate approach.

The cited literature primarily focuses on legal amendments; however, none of these sources specifically address our problem. Only (Souza et al., 2021) encompass an information retrieval task for legal context, but with classical approaches like bag-of-words and BM25 variants.

Although (Agnoloni et al., 2022) bears the closest resemblance to our work, it primarily tackles an unsupervised clustering task and lexical similarity. In contrast, our research centers on the grouping of amendment documents based on topics provided by a specialist, with a focus on evaluating semantic similarity.

3 Methods

In Figure 1 is presented our approach for amendments recuperation (named as *Look for Amendments*). Our system retrieves the relevant amendment documents related to a specific topic from a predefined set of topics provided by a legislative consultant, based on the amendment documents accompanying a bill.

3.1 Corpus

Our dataset is composed of 269 legal amendment documents proposed to PEC 6/2019[4], and each document was labeled in a topic according to a legal consultant, with the dataset [5] comprising 28 topics, as can be seen in Figure 2 . PEC 6/2019, which

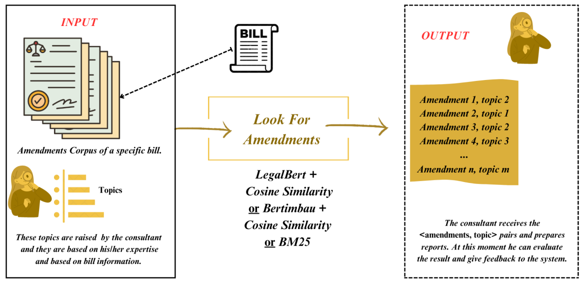

Figure 1: Our pipeline describing the input, the methods, and the output.

pertains to the Brazilian pension reform, was selected due to its extensive collection of proposed amendments, and the availability of topic annotations prepared by a consultant. Brazilian legislative texts have linguistic peculiarities and distinctive structures.We applied a preprocessing step to remove "noises" and used the NLTK (Bird and Loper, 2004) library to segment the legal documents in sentences.

3.2 Pipeline

Our pipeline operates by treating each amendment as a query and the topics as the retrieved elements. It incorporates the "Look for Amendments" component, which offers a choice between three methods: BM25L (Lv and Zhai, 2011), LegalBERT (Silva et al., 2021), and BERTimbau (Souza et al., 2020)).

For BERT models, we compute the cosine similarity between the embeddings of document contents and topics. Also, we have investigated different types of segmentation of our corpus, since each segment of the legal document contains different semantic meaning.

3.2.1 BM25 models and Variants

In their study, (Souza et al., 2021) have explored the preliminary search process for retrieving legal documents from the Brazilian Chamber of Deputies. They designed a pipeline where job requests acted as queries and bills served as the output, ranked based on their relevance to the query. Their pipeline includes the following preprocessing steps: converting text to lowercase, removing stopwords, accentuation, and punctuation. They applied two stemming algorithms: RSLP ("*Removedor de Sufixos da Língua Portuguesa*"), a rule-based algorithm specifically developed for Portuguese, and Savoy. The main purpose of stemming is to reduce the inflected words into its root form or stem. Thus, words can be mapped to the same concept,

[4] https://www.camara.leg.br/proposicoesWeb/fichadetramitacao?idProposicao=2192459

[5] https://www.diap.org.br/images/stories/emendas-pec-6-sointese-2.pdf

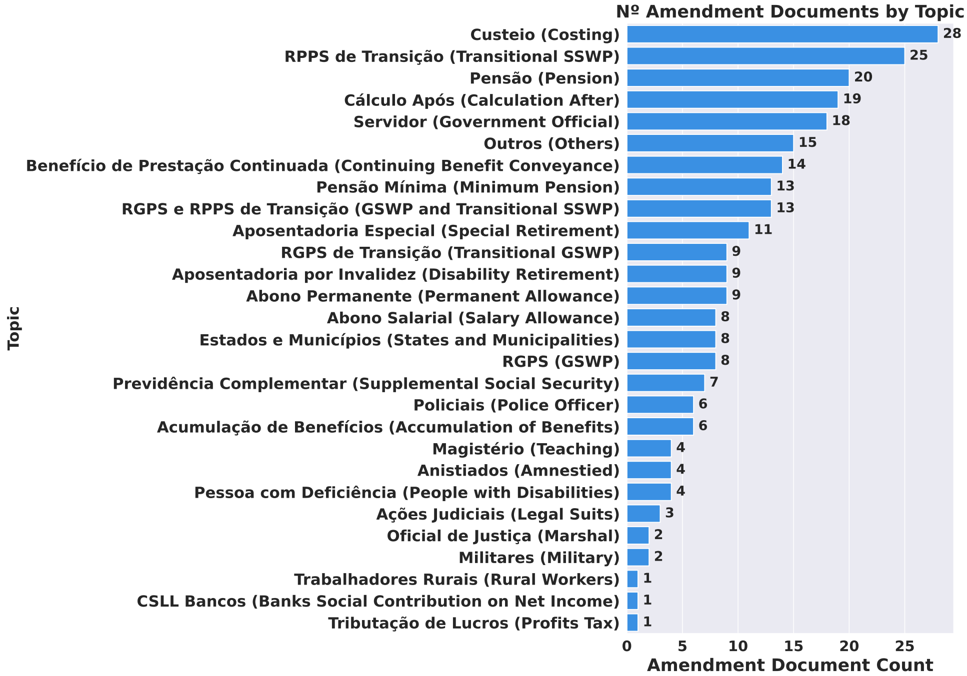

Figure 2: The number of amendment documents that are grouped per topic. "RPPS (Regime Próprio de Previdência Social)" stands for "Special Social Welfare Policy (SSWP)", and "RGPS (Regime Geral de Previdência Social)" means "General Social Welfare Policy (GSWP)" and "CSLL" denotes "Contribuição Social sobre o Lucro Líquido".

improving the process of Information Retrieval, regarding its ability to index documents and to reduce data dimensionality (Oliveira and C. Junior, 2018). RSLP algorithm was chosen because of its effectiveness in the retrieval of documents (Nogueira de Oliveira and Júnior, 2017; Oliveira and C. Junior, 2018; Flores et al., 2010; Flores and Moreira, 2016). Additionally, a language model based on N-grams was employed, with four different combinations of word N-grams evaluated. The authors utilized the BM25 scoring function, which follows a "bag of words" approach for legal domain. They also evaluated variants of BM25, including Okapi BM25, BM25L, and BM25+. The study's findings indicated that the BM25L variants performed better than other models in relation to recall metric, and that combining unigrams and bigrams demonstrated improved results for the BM25 scoring function. In contrast to our work, they do not experiment sentence models or word embeddings.

Our pipeline followed the one presented in (Souza et al., 2021), because it uses documents similar to this work and it was developed and evaluated for texts written in Portuguese. We selected the configurations which obtained the best results (see Table 1) and the BM25L (Souza et al., 2021) variant as information retrieval method because it presented the best results for the retrieval of legislative documents of our task. BM25L (Lv and Zhai, 2011) was built on the observation that Okapi BM25 penalizes more longer documents compared to shorter ones since it *shifts* the term frequency normalization formula to boost scores of very long documents.

As preprocessing, both topic and amendments had their texts converted to lowercase and had stopwords, accentuation, and punctuation removed. The preprocessing techniques were performed using the Python NLTK. For the stopword removal, we used a Portuguese stopword list.

3.2.2 LegalBERT + cosine similarity

LegalBERT (Silva et al., 2021) is based on SBERT (Reimers and Gurevych, 2019) architecture, and was designed to be more adapted to the legal domain more effectively compared to general-purpose models. LegalBERT was trained on a large corpus of legal texts in Portuguese language, such as legislation and population comments about bills, to capture the unique patterns, terminology, and context specific to the legal domain. Once we have sentence embeddings computed, we compute the cosine similarity between amendment embeddings and topics embeddings to measure the semantic similarity of two texts. We consider the highest

configuration ID	preprocessing
0	stopword and accentuation removal
1	no preprocessing
5	lowercase + punctuation, accentuation, and stopword removal
8	lowercase + punctuation, accentuation, and stopword removal + stemming (RSLP)
21	lowercase + punctuation, accentuation, and stopword removal + stemming (Savoy) + unigram and bigram

Table 1: Subset of BM25 configurations from (Souza et al., 2021). We chose these configurations (configurations 0, 1, 5, 8 and 21) from (Souza et al., 2021) because they resulted in better recall scores when adapted to our task.

scoring pairs to associate the amendment and topic.

3.2.3 BERTimbau + cosine similarity

BERTimbau is an approach that replicates BERT's architecture to adapt it for the Portuguese language, outperforming previous models on various evaluation tasks in Portuguese. Once we have word embeddings computed, we compute the cosine similarity between amendment and topics embeddings in the same way we did for LegalBERT.

4 Results

To improve efficiency for the legal consultant, our system aims to retrieve relevant documents with high recall but without overwhelming the user with a large quantity of retrieved documents. To meet this objective and reduce manual analysis, we adopt Recall@28 as our evaluation metric, as we have 28 topics.

4.1 Results for BM25L configurations

The BM25L configurations we adopted are described in Table 1. Regarding the BM25L approach, the configurations 5, 8, and 21 had similar resulting recalls and performed better than the others. Configuration 5 is the fastest in relation to the previous 3 configurations, because it requires less preprocessing steps. We point out that the configuration 8 can be more advantageous than configurations 5 and 21 when taking into account both recall scores and preprocessing computational time. Configurations 0 and 1 had worse performances. (See Figure 3).

4.2 Results comparing our 3 methods (BM25L, LegalBERT and BERTimbau)

The type of segments adopted in this work are shown in Figure 4, in which each segment of the legal document is highlighted: *i) Main Text* has hierarchical and complex structure, referencing elements of a bill, such as legal articles, paragraphs, items, etc. *ii) Justification* or *Justificativa*, in Portuguese, is more similar to a natural language text, being less structured, offering the rationales behind the amendment proposal, and *iii) Full Content* considers the whole text of the amendment document, also including the *Main Text* and the *Justification*.

We choose the configuration 0 of BM25L (see Table 1) to make a fair comparison with the other BERT approaches, since the latter requires no preprocessing due to the fact that BERT models are trained on raw texts. Configuration 0 of BM25L only applies stopword and accentuation removal, while the others (5, 8 and 21) apply stemming. Although configuration 1 of BM25L requires no preprocessing, being more similar to BERT models in its text preprocessing step, it had the lowest performance in relation to the other BM25L configurations. Therefore, we argue that the configuration 0 is the most suitable for comparison with BERT models when considering both recall performance and text preprocessing.

In general, the BM25L approach surpasses the performance of LegalBERT and BERTimbau by obtaining higher recall values. In relation to the BM25L approach, using the Full Content segment of the amendment text had better recall values than adopting the other types of segments. For both LegalBERT and BERTimbau approaches, for fewer number of documents retrieved, the Justification segment presented better recall - interestingly, the Justification part of the amendment has its structure more similar to natural language. In our task, the LegalBERT approach was better than BERTimbau, possibly by capturing more the semantic structure of the legal text since it was adapted to this domain.

5 Conclusion and Future Work

In the preprocessing phase of the BM25 algorithm, it is crucial to apply case folding, keep punctuation and remove stopwords, especially in the legal domain. Neglecting any form of preprocessing has resulted in the poorest performances. Stemming reduces words to their base or root form, aiding in matching similar terms and improving retrieval accuracy. Therefore, in the context of the legal domain, incorporating both preprocessings and stem-

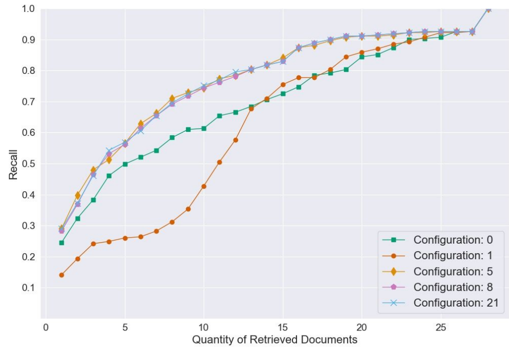

Figure 3: Resulting recall for different configurations of the BM25L considering the Full Content of the amendment documents.

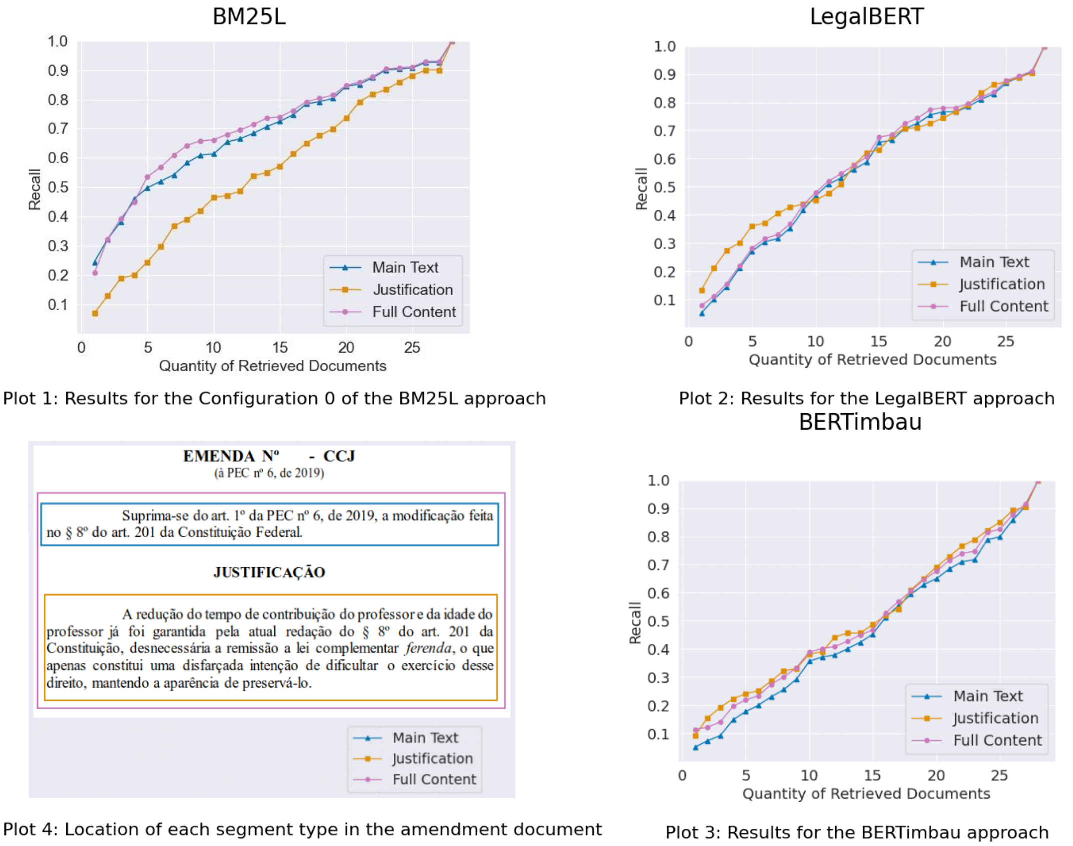

Figure 4: Comparison of our 3 approaches (BM25L, LegalBERT, and BERTimbau) considering each segmentation type (Main Text, Justification and Full Content). Note that only the configuration 0 of BM25L was used in order to perform a fair comparison with BERT models, as explained in subsection 4.2.

ming can significantly enhances the performance of the BM25L algorithm. BM25L shows stronger performance in relation to SBERT models allied with the cosine similarity. We argue that this happens because, in most cases, the words that describes a topic are also present throughout the amendment text, and an exact match is possible to occur. As limitations, our dataset can be considered small and no other data, annotated by a legal consultant, is available. As future work, it is possible to do a fine tuning on the amendment documents and use the embeddings of other models and assess their performance in our task and also to observe how the cited methods perform in a larger dataset, when available.

References

Tommaso Agnoloni, Carlo Marchetti, Roberto Battistoni, and Giuseppe Briotti. 2022. Clustering similar amendments at the Italian senate. In *Proceedings of the Workshop ParlaCLARIN III within the 13th Language Resources and Evaluation Conference*, pages 39–46, Marseille, France. European Language Resources Association.

Steven Bird and Edward Loper. 2004. NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain. Association for Computational Linguistics.

Felipe N. Flores and Viviane P. Moreira. 2016. Assessing the impact of stemming accuracy on information retrieval – a multilingual perspective. *Information Processing & Management*, 52(5):840–854.

Felipe N. Flores, Viviane P. Moreira, and Carlos A. Heuser. 2010. Assessing the impact of stemming accuracy on information retrieval. In *Computational Processing of the Portuguese Language*, pages 11–20, Berlin, Heidelberg. Springer Berlin Heidelberg.

Yuanhua Lv and ChengXiang Zhai. 2011. When documents are very long, bm25 fails! pages 1103–1104.

Robert Nogueira de Oliveira and Methanias Júnior. 2017. Assessing the impact of stemming algorithms applied to judicial jurisprudence - an experimental analysis. pages 99–105.

Robert A. Oliveira and Methanias C. Junior. 2018. Experimental analysis of stemming on jurisprudential documents retrieval. *Information*, 9(2).

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Nádia Silva, Marília Silva, Fabíola Pereira, João Tarrega, João Beinotti, Márcio Fonseca, Francisco Andrade, and André Carvalho. 2021. Evaluating topic models in portuguese political comments about bills from brazil's chamber of deputies. In *Anais da X Brazilian Conference on Intelligent Systems*, Porto Alegre, RS, Brasil. SBC.

Aleksander Smywiński-Pohl, Mateusz Piech, Zbigniew Kaleta, and Krzysztof Wróbel. 2021. Automatic extraction of amendments from polish statutory law. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, ICAIL '21, page 225–229, New York, NY, USA. Association for Computing Machinery.

Ellen Souza, Douglas Vitório, Gyovana Moriyama, Luiz Santos, Lucas Martins, Mariana Souza, Márcio Fonseca, Nádia Félix, André Carlos Ponce de Leon Ferreira de Carvalho, Hidelberg O. Albuquerque, and Adriano L. I. Oliveira. 2021. An information retrieval pipeline for legislative documents from the brazilian chamber of deputies. In *Legal Knowledge and Information Systems*, pages 119–126. IOS Press.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. *BERTimbau: Pretrained BERT Models for Brazilian Portuguese*, pages 403–417.