Intelligent Classification of Legislative Proposals

Luca Peres Quinta da Guarda[1], **Ronie Paulucio Porfirio**[1]

[1]Programa de Pós-Graduação em Computação Aplicada – Universidade de Brasília (UnB)
Campus Universitário Darcy Ribeiro, Brasília-DF – CEP 70910-900 – Brasília – DF – Brazil

lucapqg@gmail.com, ronie.porfirio@gmail.com

Abstract. *This paper presents a technique for the classification of legislative proposals, specifically focusing on reclassifying those initially categorized as "Others"into more appropriate subject categories. Leveraging machine learning (ML) and artificial intelligence (AI) methods, the technique analyzes and reclassifies legislative documents based on their content. Our approach aims to automate and make more cohesive the process of legislative proposal classification, facilitating better organization and retrieval of legislative information. Our results show promising improvements in classification accuracy. However, additional work is necessary to further refine and enhance this technique.*

1. Introduction

Legislative proposals, as outlined in the Internal Rules of the Legislative Chamber (CLDF, art. 129), cover a broad spectrum of materials for deliberation. These include amendments to the Organic Law, complementary and ordinary bills, legislative decrees, resolutions, indications, motions, requests, amendments, and appeals. Proper classification of these diverse documents is essential for efficient legislative processes, accurate information retrieval, and effective management.

Given the complexity and variety of legislative documents, traditional manual classification methods often fall short, leading to inefficiencies and inaccuracies. In the Câmara Legislativa do Distrito Federal (CLDF) context, proposals are categorized based on their subject matter or theme. However, due to factors such as ambiguous topics, outdated categories, and the multidisciplinary nature of many proposals, numerous documents are classified under the generic label "Others."

This broad classification impedes effective tracking, analysis, and transparency of legislative activities, complicating the ability of both the public and lawmakers to understand and manage the legislative process.

Thus, there is a pressing need for automated systems capable of accurately and rapidly classifying legislative proposals into appropriate categories. This paper addresses the challenges of proposal classification, identifies key issues, and proposes solutions to automate and enhance the accuracy and transparency of legislative categorization.

Our primary objective is to develop a model that can suggest more accurate categories for proposals currently classified as "Others,"thereby improving their classification accuracy and relevance.

Thematic classification is crucial for optimizing legislative proposal analysis and processing. An effective classification system streamlines the legislative process, aids in resource allocation, and supports lawmakers in making informed decisions. Furthermore,

it enhances transparency and organization within the legislative workflow, ensuring accurate information retrieval and effective legislative management, thus fostering a more structured legislative process.

2. The Problem

A significant issue is the increasing number of legislative proposals categorized under the broad theme of ”Others.”Typically, the author of a proposal is responsible for assigning it to one or more relevant themes. However, due to ambiguous topics, outdated categories, and the multidisciplinary nature of many proposals, a considerable proportion end up classified under the non-specific category of ”Others.”

To illustrate this issue, data shows that nearly 20% of the total proposals fall under this generic category. This excessive reliance on the ”Others”category highlights deficiencies in the current classification system.

3. Objectives

This study aims to evaluate various machine learning models to determine which is most effective in suggesting more accurate thematic categories for legislative proposals. The goal is to identify themes that better align with the proposals, thus enhancing the relevance and accuracy of their categorization.

It is important to clarify that this study does not seek to automate the entire classification process, replace human classifiers, or compare human classification with machine learning-based methods. Instead, the focus is on improving the existing categorization process by identifying the best machine learning model for suggesting themes for proposals currently labeled as ”Others.”

4. Methodology

In the initial phase of the methodology, known as Business Understanding, we analyzed legislative documents to ensure our data mining objectives were aligned with the needs of legislative classification. This involved reviewing various legislative proposals and their metadata to align our goals with the requirements for effective classification.

The Data Understanding phase involved working with a dataset of 22,267 summaries extracted from the “Processo Legislativo Eletrônico (PLE)” system, covering the period from 2021 to May 2024. Each summary in this dataset is accompanied by its respective thematic classification, providing the foundation for further analysis.

Data Preparation included several critical steps. First, we discarded data classified under the “others” and “not applicable” categories. Following this, we performed preprocessing tasks such as tokenization, normalization, stopword removal, and lemmatization. For vectorization, we employed two techniques: the Multilingual Sentence Embedding Model, based on the MiniLM architecture, which generates high-quality embeddings capturing the semantic meaning of the text, and the TF-IDF (Term Frequency-Inverse Document Frequency) Model, which converts text into numerical vectors by evaluating term frequency within documents and across the dataset, thereby highlighting term importance.

In the Modeling phase, we utilized several models to address the classification task. The DummyClassifier was used as a baseline model to provide a reference point

for comparing the performance of more complex models. The Support Vector Machine (SVM) was employed to find the hyperplane that best separates the classes in the feature space, proving effective in high-dimensional spaces. Logistic Regression, a linear model, was used to estimate the probability of a binary outcome based on input features, valued for its simplicity and interpretability. Additionally, XGBoost (Extreme Gradient Boosting) was chosen for its ability to build an ensemble of weak learners, such as decision trees, improving model performance through optimization. Random Forest, another ensemble model, constructs multiple decision trees and aggregates their predictions, demonstrating robustness against overfitting and effectiveness with larger datasets. Finally, K-Nearest Neighbors (KNN), a non-parametric model, classifies data points based on the majority class of their nearest neighbors, offering simplicity and effectiveness for smaller datasets.

The Evaluation phase focused on assessing model performance using various metrics. Accuracy was used as a straightforward measure of overall correctness. Precision and recall provided insights into performance, particularly with imbalanced classes, by focusing on the performance of minority classes. The F1 Score combined precision and recall into a single metric, summarizing model performance effectively.

Finally, in the Deployment phase, the results of the study could be used to create an interface within the "Processo Legislativo Eletrônico (PLE)" system. This interface would suggest themes that best fit new legislative proposals, thereby enhancing the legislative classification process.

5. Experiments conducted

In the experiments conducted, a cross-validation with three stratified partitions was performed for each model. The three partitions were randomly divided to allow for training and testing each model across all four metrics, both for training and testing data. Along with the four metrics, the training and inference times were also measured. For each metric, the average of the three validations was computed.

The primary aim of this approach was to assess overfitting and underfitting. Overfitting is indicated if there is good performance during training but poor performance during testing. Conversely, underfitting is suggested if there is poor performance in both training and testing phases.

For the embedding vectorizer experiments, Logistic Regression demonstrated overall excellence. Although it might not be the best in every individual metric, it consistently outperformed other models across the set of metrics.

Similarly, when using the Tfidf Vectorizer, Logistic Regression again emerged as the top performer. This model showed superior performance across the metrics compared to others.

In experiments involving the Tfidf Vectorizer combined with the MultiOutputClassifier, Logistic Regression continued to prove itself as the top-performing model. This combination did not alter the overall performance ranking of Logistic Regression, which remained the best-performing model across these experimental setups.

6. Conclusions

Logistic Regression, when trained with TfidfVectorizer and MultiOutputClassifier, achieved the best results. However, the analysis identified some accuracy issues. The classifications are not fully representative of the true distribution of legislative proposals, and the data volume is insufficient to meet desired objectives effectively.

Performance metrics indicated that:

- The top 3 probabilities achieved an accuracy of 0.7027.
- The top 4 probabilities reached an accuracy of 0.7613.
- he top 5 probabilities resulted in an accuracy of 0.8030.

These results reflect the model's ability to suggest relevant categories based on the probabilities of different themes.

7. Future work

Future research should focus on integrating neural network models to explore their potential for improving classification accuracy. Additionally, refining class definitions and reclassifying the data will be essential to address current limitations and enhance the effectiveness of the classification process.