# Intelligent Classification of Legislative Propositions

## A comparative analysis of machine learning models for proposition classification

**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.*** *Reescrever*

## 1. Introduction

A legislative proposition, as outlined in the internal regulations of *Câmara Legislativa do Distrito Federal* [Legislativa 2018], is any matter that requires discussion and approval in the institution. There are ten types of propositions, presented below in hierarchical order, with their original names in portuguese and their english translations:

- *Proposta de Emenda à Lei Orgânica (PELO)* - Proposal of Amendment to the Organic Law
- *Projeto de Lei Complementar (PLC)* - Complementary Law Project
- *Projeto de Lei (PL)* - Law Project
- *Projeto de Decreto Legislativo (PDL)* - Legislative Decree Project
- *Projeto de Resolução (PR)* - Resolution Project
- *Indicação (IND)* - Indication
- *Moção (MO)* - Motion
- *Requerimento (REQ)* - Request
- *Emenda* - Amendment
- *Recurso (REC)* - Appeal

A proposition must fulfill specific requirements to be considered valid and admissible. It must pertain to matters within the jurisdiction of Brazilian Federal District (*Distrito Federal*), in accordance with its Organic Law and the precepts of the Federal Constitution. Furthermore, the proposition must adhere to legislative techniques, contain the minimum number of subscribers (signatories), and be structured with the following elements: epigraph (a name that distinguishes one act from another, including number and date when applicable), indication of the author, summary (a synopsis of the content or purpose of the act, indication of the *Câmara Legislativa* as the legislative body, the text to be deliberated, a justification, the date and the signatures [Legislativa 2024].

In the *Câmara Legislativa do Distrito Federal* (CLDF) context, propositions are categorized based on their subject matter or theme. Typically, the author of a proposition is responsible for assigning it to one or more relevant themes. This thematic classification is crucial for optimizing legislative proposal analysis and processing. An effective classification streamlines the legislative process, aids in resource allocation and supports lawmakers in making informed decisions. Additionally, it enhances transparency and organization within the legislative workflow, ensuring accurate information retrieval and effective legislative management.

However, given the complexity and variety of legislative documents and due to ambiguous topics and the multidisciplinary nature of many propositions, manual classification methods often fall short, leading to inefficiencies and inaccuracies with a considerable proportion of propositions ending up classified under the non-specific category of ”other”. To illustrate this issue, data indicates that nearly 20% of all propositions are classified under the generic category of “other” as shown in the figure 1 presenting the percentage of propositions in each category relative to the total:

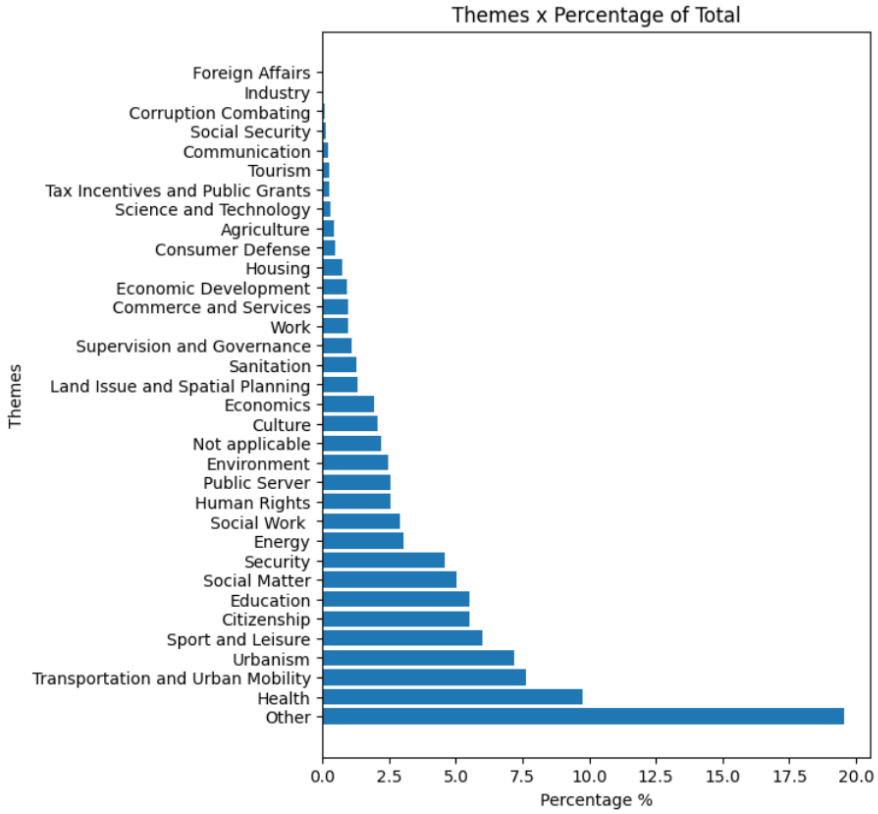


**Figura 1. Themes x Percentage of Total**

This excessive reliance on the “other” category underscores significant deficiencies in the current classification system. It hinders effective tracking, analysis, and transparency of legislative activities, making it difficult for both the public and lawmakers to understand and manage the legislative process.

## 2. Objective

This study aims to evaluate various machine learning models to determine which is most effective in suggesting more accurate thematic categories for legislative propositions. The goal is to identify themes that better align with the propositions, 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 propositions in order to encourage legislators to use the ”other”category only as a last resort.

## 3. Methodology

Our methodology adopts the Cross-Industry Standard Process for Data Mining (CRISP-DM) model as a reference citar alguma fonte:

### 3.1. Business Understanding

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 propositions and their metadata to align our goals with the requirements for effective classification.

### 3.2. Data Understanding

Then, 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.

### 3.3. Data Preparation

"Data Preparation" included several critical steps. First, we discarded data classified under the "other" and "not applicable" categories. Following this, we performed preprocessing tasks such as tokenization, normalization, stopword removal, and lemmatization. For vectorization, we employed three techniques: the Multilingual Sentence Embedding Model, based on the MiniLM architecture, which generates high-quality embeddings capturing the semantic meaning of the text; 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; and, finally, the BM25 Model, which is an extension of TF-IDF that incorporates term frequency saturation and document length normalization to improve retrieval performance by ranking documents based on their relevance to a given query.

### 3.4. Modeling

Incluir uma citacao por modelo

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.

### 3.5. Evaluation

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.

### 3.6. Deployment

Finally, in the "Deployment"phase, the results of the study could be used to develop an interface within the "*Processo Legislativo Eletrônico (PLE)*" system. This interface would suggest themes that best fit new legislative propositions, thereby enhancing the legislative classification process. However, this phase has not yet been executed, as it depends on the development of this interface within the system.

Parei aqui

## 4. 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.

## 5. 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 propositions, 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.

## 6. 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.

## Referências

Legislativa, C. (2024). Proposições, projetos, atos normativos (conceitos). Internet.

Legislativa, D. F. B. C. (2018). Regimento interno da câmara legislativa do distrito federal. Brasília: Câmara Legislativa do Distrito Federal.