Intelligent Classification of Legislative Proposals

A Comparative Analysis of Machine Learning Models for Proposal Classification

Luca Peres Quinta da Guarda Ronie Paulucio Porfirio

Universidade de Brasília (UnB)

Programa de Pós-Graduação em Computação Aplicada

Course Mineração de Dados Massivos

Prof. Dr. Marcelo Ladeira and Pr. MSc. Gustavo Cordeiro

Contents

Legislative Proposal

- A **Legislative Proposal** is any matter subject to deliberation by the Legislative Chamber (CLDF Internal Regulations, art. 129).

Legislative Proposals can be of the following types:

- Proposta de emenda à Lei Orgânica;
- Projeto de lei complementar;
- Projeto de lei;
- Projeto de decreto legislativo;
- Projeto de resolução;
- Indicação;
- Moção;
- Requerimento;
- Emenda;
- Recursos;

Introduction

Business Understanding - Legislative Proposal's Themes

Legislative Proposals are classified into one or more themes:

- Agricultura
- Assistência Social
- Assunto Fundiário e Ordenamento Territorial
- Assunto Social
- Cidadania
- Ciência e Tecnologia
- Combate à Corrupção
- Comunicação
- Comércio e Serviços
- Cultura
- Defesa do Consumidor
- Desenvolvimento Econômico
- Desporto e Lazer
- Direitos Humanos
- Economia
- Educação
- Energia
- Fiscalização e Governança
- Habitação
- Incentivos Fiscais e Concessões Públicas
- Indústria
- Meio Ambiente
- Não se aplica
- Outro
- Previdência Social
- Relações Exteriores
- Saneamento
- Saúde
- Segurança
- Servidor Público
- Trabalho
- Transporte e Mobilidade Urbana
- Turismo
- Urbanismo

Total: 34 themes

Thematic Classification Benefits

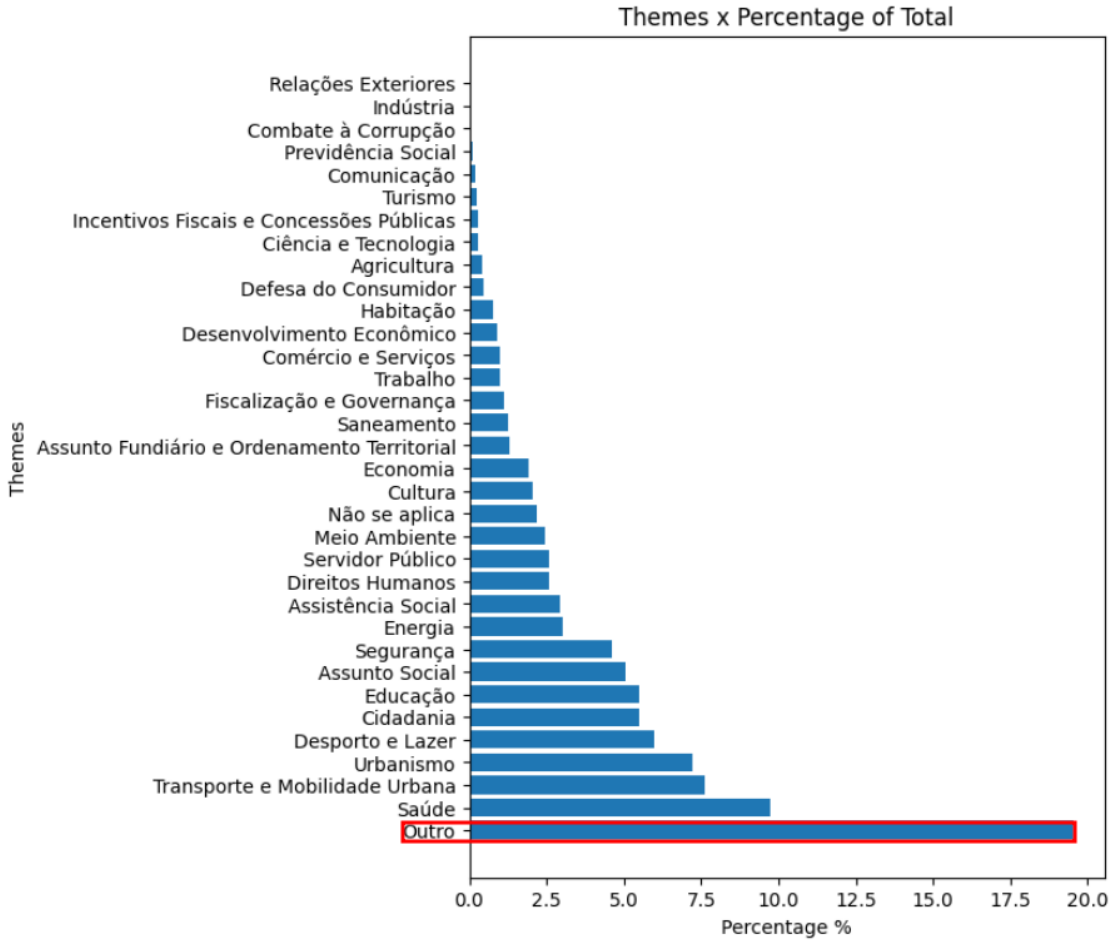
- Efficient classification of legislative proposals is crucial to **streamline their analysis** and processing within the legislative process helping to **determine which committees a proposal should go through**.
- By categorizing legislative proposals into relevant themes, lawmakers can streamline their analysis, **allocate resources efficiently**, and **make informed decisions**.
- This process **enhances transparency** and facilitates a more **organized legislative workflow**.
- Thematic classification plays an important role in maintaining **accurate information retrieval** and **ensuring effective legislative management**.

Theme “others” is growing bigger

- Usually, the author of the proposal is responsible for classifying it into one or more themes.
- Unfortunately, due to various factors such as ambiguous topics, outdated categories, multidisciplinary nature, **many propositions end up being classified under the generic theme of “outro” (others)**.

The Problem

Understanding the problem



The chart shows that the number of proposals classified under the theme “outro” (other) represents almost 20% of the total.

The Problem

Problem definition

The problem is inadequate proposal classification

Inadequate classification hinders efficient tracking, analysis, and transparency of legislative activities, making it difficult for both society and lawmakers to understand and oversee the legislative process effectively.

Primary objective

- The primary objective of this study is to **compare different machine learning models** to determine which model is **most effective in suggesting more appropriate theme categories for legislative proposals**.
- The goal is to find **more suitable themes that better match these proposals**.

Objectives

Disclaimer

Attention!

- This study **does not aim to automate the classification process, replace human classification, or compare human classification with machine learning classification**.
- Instead, **the focus is on enhancing the existing categorization process by identifying the best model for suggesting themes for proposals currently classified under the generic label ”others.”**.

Literature review

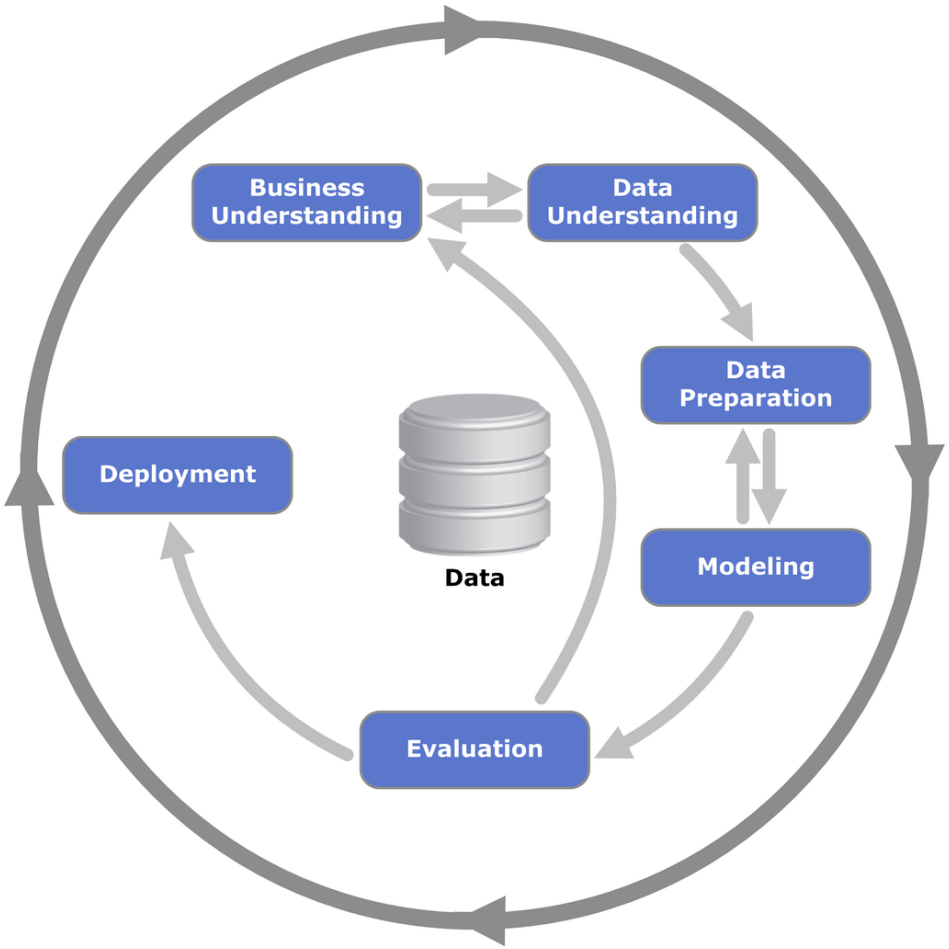
J. Andrade Junior, J. Cardoso-Silva, and L. Bezerra (2021)
Comparing Contextual Embeddings for Semantic Textual Similarity in Portuguese
Anais da X Brazilian Conference on Intelligent Systems

Alantari, J., Currim, S., Sanghvi, Y. et al. (2022)
An empirical comparison of machine learning methods for text-based sentiment analysis of online consumer reviews
International Journal of Research in Marketing - Pages 1-19

Shah, K., Patel, H., Sanghvi, D. et al. (2022)
A Comparative Analysis of Logistic Regression, Random Forest and KNN Models for the Text Classification
Augment Hum Res 5, 12

Methodology

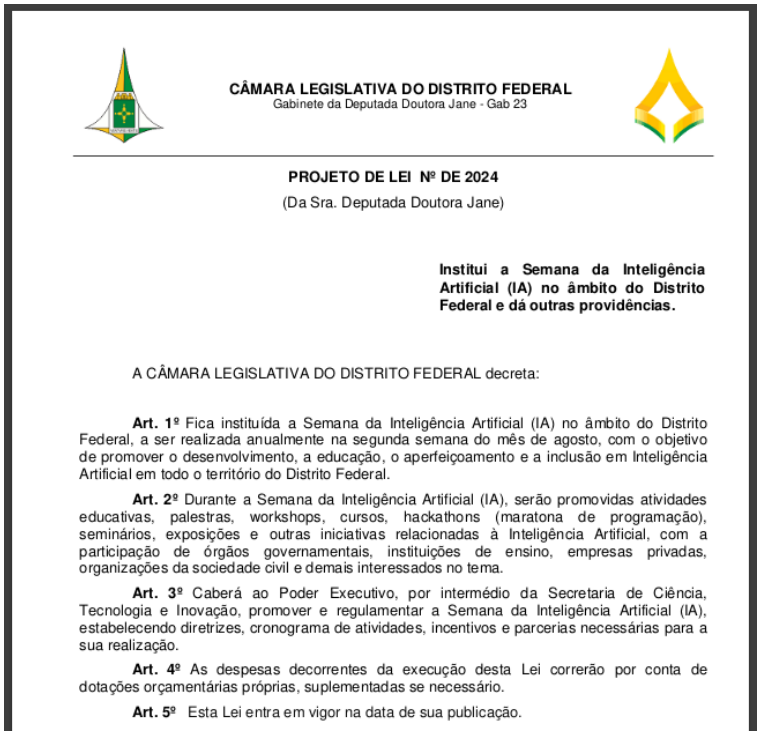
CRISP-DM's Methodology



We utilized the CRISP-DM methodology. Its stages will be explained in the following slides.

1 - Business Understanding

We have analyzed legislative documents to align our data mining objectives with legislative classification needs.



CÂMARA LEGISLATIVA DO DISTRITO FEDERAL
Gabinete da Deputada Doutora Jane - Gab 23

PROJETO DE LEI Nº DE 2024

(Da Sra. Deputada Doutora Jane)

Institui a Semana da Inteligência Artificial (IA) no âmbito do Distrito Federal e dá outras providências.

A CÂMARA LEGISLATIVA DO DISTRITO FEDERAL decreta:

Art. 1º Fica instituída a Semana da Inteligência Artificial (IA) no âmbito do Distrito Federal, a ser realizada anualmente na segunda semana do mês de agosto, com o objetivo de promover o desenvolvimento, a educação, o aperfeiçoamento e a inclusão em Inteligência Artificial em todo o território do Distrito Federal.

Art. 2º Durante a Semana da Inteligência Artificial (IA), serão promovidas atividades educativas, palestras, workshops, cursos, hackathons (maratona de programação), seminários, exposições e outras iniciativas relacionadas à Inteligência Artificial, com a participação de órgãos governamentais, instituições de ensino, empresas privadas, organizações da sociedade civil e demais interessados no tema.

Art. 3º Caberá ao Poder Executivo, por intermédio da Secretaria de Ciência, Tecnologia e Inovação, promover e regulamentar a Semana da Inteligência Artificial (IA), estabelecendo diretrizes, cronograma de atividades, incentivos e parcerias necessárias para a sua realização.

Art. 4º As despesas decorrentes da execução desta Lei correrão por conta de dotações orçamentárias próprias, suplementadas se necessário.

Art. 5º Esta Lei entra em vigor na data de sua publicação.

(a) Proposal example

(b) Proposal's data

2 - Data Understanding

The dataset contains 22,267 summaries extracted from the "*Processo Legislativo Eletrônico (PLE)*" system, covering the period from 2021 to May 2024. Each summary is accompanied by its respective thematic classification.

3 - Data Preparation

1. **Preprocessing**: We discard data classified under “outro” and “não se aplica” themes. Then, we perform tokenization, normalization, stopword removal, and lemmatization processes.
2. **Vectorization**:
   - **Multilingual sentence embedding model** based on the MiniLM architecture, a lightweight and efficient BERT variant with 12 transformer layers, to produce high-quality embeddings that capture the semantic meaning of the text.
   - **TF-IDF (Term Frequency-Inverse Document Frequency) model** involves converting text into numerical vectors based on the frequency of terms within a document and across a collection of documents. This approach captures the importance of a term in a document relative to the entire dataset.

4 - Modeling

1. **DummyClassifier**: A **baseline model** that makes predictions using simple rules and establishes a baseline to **compare the performance of more complex models**.
2. **Support Vector Machine (SVM)**: A powerful model that **finds the hyperplane that best separates the classes in the feature space**. It is effective for high-dimensional spaces and when the number of dimensions exceeds the number of samples.
3. **Logistic Regression**: A linear model that **estimates the probability of a binary outcome based on input features**. It is simple and interpretable, good for linearly separable data and understanding feature importance.

4 - Modeling (cont...)

4. **XGBoost (Extreme Gradient Boosting)**: An optimized gradient boosting algorithm that **builds an ensemble of weak learners (typically decision trees) to improve model performance**. Known for high performance, speed, and scalability.
5. **Random Forest**: An **ensemble model that constructs multiple decision trees and aggregates their predictions**. Robust against overfitting, good for handling large datasets with higher dimensionality.
6. **K-Nearest Neighbors (KNN)**: A **non-parametric model that classifies a data point based on the majority class of its k nearest neighbors**. Simple and intuitive, effective for small datasets with low noise.

Why choose these models?

1. **Baseline Comparison**: DummyClassifier provides a benchmark to gauge the performance of more sophisticated models.
2. **Linear and Non-Linear Data**: Logistic Regression and SVM cover linear relationships, while Random Forest, XGBoost, and KNN handle non-linear data.
3. **Model Performance**: XGBoost and Random Forest are chosen for their strong predictive performance and ability to handle complex datasets.
4. **Interpretability**: Logistic Regression is valued for its simplicity and ease of interpretation.
5. **Versatility**: SVM, Random Forest, and XGBoost offer versatility across different types of data and problems.
6. **Scalability**: XGBoost is particularly chosen for its scalability and efficiency in handling large datasets.

5 - Evaluation

Our evaluation metrics include **Accuracy**, **Precision**, **Recall** and **F1 score**.

- **Accuracy** is **straightforward and easy to understand**.
- **Precision** and **recall** are more informative when dealing with **imbalanced classes** because they provide insights into the performance of the minority class, which accuracy might overlook.
- **F1 score** gives a **balance between precision and recall as a single metric** to **summarize** the model performance.

6 - Deployment

- The results might be used to create an interface in the "*Processo Legislativo Eletrônico (PLE)*" system to suggest themes that best fit new proposals.

Figure: Plenário da Câmara Legislativa do Distrito Federal

Experiments conducted

Understanding the experiments

The experiments

- In all experiments, a cross-validation with three stratified partitions is performed.
- The three partitions are randomly divided in a way that it allows training and testing each model for all four metrics, both for training and testing data.
- In addition to the four metrics, the training and inference times are also measured.
- For each metric, we get the average of the three validations.

Why? To check for overfitting or underfitting

- If there is **good performance in training but not in testing**, there is overfitting.
- If there is **poor performance in both training and testing**, there is underfitting.

Experiments conducted

Embedding - Cross Validation

Embedding Vectorizer

- Logistic Regression excels overall. It may not be the best in everything, but across the set of metrics, it is the best.

	fit_time	score_time	test_f1_w	train_f1_w	test_bal_a	train_bal_a	test_prc_w	train_prc_w	test_rec_w	train_rec_w
Model										
DummyClassifier	0.028611	0.013967	0.025819	0.025819	0.031250	0.031250	0.014462	0.014462	0.120258	0.120258
KNeighborsClassifier	0.040055	1.049414	0.307679	0.516615	0.179927	0.370583	0.323286	0.547667	0.311971	0.520135
LogisticRegression	3.597823	0.025262	0.252122	0.282075	0.217896	0.387424	0.356856	0.397779	0.234561	0.270780
RandomForestClassifier	225.171689	0.825105	0.299710	**0.807638**	0.163765	**0.865793**	0.328777	**0.835075**	0.327624	0.801934
SVM	42.684947	33.030448	0.295933	0.346580	**0.241816**	0.495388	**0.376351**	0.435997	0.276796	0.337201
XGBClassifier	253.781450	0.560534	**0.325231**	0.800150	0.178569	0.726839	0.330270	0.800984	**0.341866**	**0.802977**

Legends:

test_f1_w: test_f1_weighted
test_bal_a : test_balanced_accuracy
test_prc_w : test_precision_weighted
test_rec_w : test_recall_weighted

train_f1_w: train_f1_weighted
train_bal_a : train_balanced_accuracy
train_prc_w : train_precision_weighted
train_rec_w : train_recall_weighted

Table: Embedding - Cross Validation

Experiments conducted

TfidfVectorizer - Cross Validation

Tfidf Vectorizer

- Once again, Logistic Regression excels overall.

	fit_time	score_time	test_f1_w	train_f1_w	test_bal_a	train_bal_a	test_prc_w	train_prc_w	test_rec_w	train_rec_w
Model										
DummyClassifier	0.005016	0.013737	0.025819	0.025819	0.031250	0.031250	0.014462	0.014462	0.120258	0.120258
KNeighborsClassifier	0.004096	28.906825	0.323249	0.541327	0.181929	0.396655	0.446675	0.629905	0.299141	0.527901
LogisticRegression	3.725738	0.018837	0.467718	0.584556	**0.377732**	0.727172	0.506144	0.629863	0.458011	0.585697
RandomForestClassifier	267.923079	1.148535	0.435199	**0.806360**	0.272152	**0.864809**	0.445313	**0.834005**	0.449355	**0.801044**
SVM	44.651407	12.235956	**0.472803**	0.676972	0.319393	0.803307	**0.512542**	0.712173	0.456292	0.673389
XGBClassifier	102.740959	1.606697	0.449333	0.766095	0.277803	0.768935	0.449830	0.767603	**0.458809**	0.769460

Legends:

test_f1_w: test_f1_weighted
test_bal_a : test_balanced_accuracy
test_prc_w : test_precision_weighted
test_rec_w : test_recall_weighted

train_f1_w: train_f1_weighted
train_bal_a : train_balanced_accuracy
train_prc_w : train_precision_weighted
train_rec_w : train_recall_weighted

Table: TfidfVectorizer - Cross Validation

Experiments conducted

TfidfVectorizer and MultiOutputClassifier - Cross Validation

TfidfVectorizer and MultiOutputClassifier

- Once again, Logistic Regression proves to be the top-performing model.

	fit_time	score_time	test_f1_w	train_f1_w	test_bal_a	train_bal_a	test_prc_w	train_prc_w	test_rec_w	train_rec_w
Model										
DummyClassifier	0.004188	0.011938	0.025819	0.025819	0.031250	0.031250	0.014462	0.014462	0.120258	0.120258
KNeighborsClassifier	0.005413	29.213731	0.323249	0.541327	0.181929	0.396655	0.446675	0.629905	0.299141	0.527901
LogisticRegression	3.685860	0.019390	0.467718	0.584556	**0.377732**	0.727172	0.506144	0.629863	0.458011	0.585697
RandomForestClassifier	268.849362	1.154890	0.434992	**0.806396**	0.272071	**0.864844**	0.444953	**0.834183**	0.449048	**0.800982**
SVM	44.404848	12.142895	**0.472803**	0.676972	0.319393	0.803307	**0.512542**	0.712173	0.456292	0.673389
XGBClassifier	102.840638	1.657798	0.449333	0.766095	0.277803	0.768935	0.449830	0.767603	**0.458809**	0.769460

Legends:

test_f1_w: test_f1_weighted
test_bal_a : test_balanced_accuracy
test_prc_w : test_precision_weighted
test_rec_w : test_recall_weighted

train_f1_w: train_f1_weighted
train_bal_a : train_balanced_accuracy
train_prc_w : train_precision_weighted
train_rec_w : train_recall_weighted

Table: TfidfVectorizer and MultiOutputClassifier - Cross Validation

Results

Results

- Logistic Regression trained with TfidfVectorizer and MultiOutputClassifier had the best results.

Class	Precision	Recall	F1-Score	Support
0	0.11	0.08	0.10	24
1	0.45	0.43	0.44	128
2	0.48	0.33	0.39	78
3	0.07	0.16	0.10	94
4	0.18	0.34	0.24	120
5	0.50	0.14	0.21	44
6	0.00	0.00	0.00	6
7	0.38	0.12	0.19	24
8	0.23	0.21	0.22	43
9	0.49	0.38	0.43	110
10	0.50	0.13	0.21	76
11	0.19	0.11	0.14	66
12	0.68	0.69	0.69	254
13	0.23	0.16	0.19	154
14	0.48	0.57	0.52	65
15	0.59	0.64	0.62	208
16	0.83	0.75	0.79	146
17	0.52	0.27	0.35	90
18	0.44	0.25	0.32	56
19	0.27	0.10	0.15	30
20	0.00	0.00	0.00	9
21	0.47	0.45	0.46	108
22	0.60	0.20	0.30	15
23	1.00	0.20	0.33	5
24	0.68	0.45	0.54	80
25	0.53	0.80	0.63	259
26	0.48	0.57	0.52	159
27	0.53	0.43	0.48	130
28	0.33	0.14	0.20	90
29	0.60	0.66	0.63	293
30	0.20	0.07	0.10	30
31	0.42	0.48	0.45	264
Accuracy				0.47
Macro Avg				
	0.42	0.32	0.34	3258
Weighted Avg				
	0.48	0.47	0.46	3258

Table: Results

Conclusions

Accuracy Issues

- The classifications are not representative of the true distribution;
- There are insufficient data to meet the desired objectives;

As a recommendation model

- **Top 3 probabilities:** 0.7027
- **Top 4 probabilities:** 0.7613
- **Top 5 probabilities:** 0.8030

Future Work

Future Work

- Integrate a neural network model into the experiments;
- Refine class definitions and reclassify the data for improved accuracy;

Thank You