Comparing Prompt-based LLMs, Fine-Tuning, and Classical Models for Legal Text Classification in Portuguese

Willgnner Ferreira Santos[1], **Arlindo Rodrigues Galvão Filho**[2], **Sávio Salvarino Teles de Oliveira**[1], **João Paulo Cavalcante Presa**[1]

[1]Institute of Informatics (INF) – Federal University of Goias (UFG)
Goiânia, Brazil

[2]Advanced Knowledge Center in Immersive Technologies (AKCIT)
Goiânia, Brazil

{willgnner_santos, joaopaulop}@discente.ufg.br

{arlindogalvao, savioteles}@ufg.br

Abstract. *This study compares fine-tuned Transformer models, LLMs with prompting, and traditional models in Portuguese legal text classification. A new Portuguese dataset was introduced for empirical evaluation. Four input formats were tested in prompting, complete text, summaries, centroids, and descriptions. Summaries achieved effective performance with reduced token usage. Similarly, KNN proved competitive in resource-limited scenarios. Input format and model capacity affected performance. The study discusses trade-offs between efficiency and interpretability. Guidelines are proposed for choosing effective strategies in legal NLP tasks.*

1. Introduction

Judicial systems face pressures stemming from increased case volumes and shortages of qualified personnel [Berman et al. 2021, Aguiar et al. 2021]. This scenario requires technological tools that can support the classification and organization of large volumes of legal documents. The identification of procedural elements can impact workflow efficiency. Artificial Intelligence (AI) has contributed to automating repetitive tasks in the legal field [Pandey and Malik 2022]. In countries such as the United States (USA) and European nations, AI-based systems are already employed in contract analysis, document categorization, and judicial decision prediction [Mills and Uebergang 2017, Kiesow Cortez and Maslej 2023, Nonato 2022]. Large Language Models (LLMs) [Moraes et al. 2024] have demonstrated a strong capability in handling semantic complexity in classification and information extraction tasks in legal texts.

While LLMs have achieved promising results in legal tasks, their performance depends on how they are applied in specific contexts. Strategies such as the use of prompt-guided LLMs, classical discriminative models, and fine-tuned architectures represent distinct approaches that can be compared regarding effectiveness, robustness, and computational feasibility. Prompt-based methods stand out for their ease of application, eliminating the need for retraining, which is beneficial in scenarios with limited resources [Elov et al. 2023]. On the other hand, Machine Learning (ML) techniques, such

as Naive Bayes with Term Frequency-Inverse Document Frequency (TF-IDF) vectors, offer an interpretable and low-cost baseline. Meanwhile, Bidirectional Encoder Representations from Transformer (BERT) models fine-tuned for the legal domain have the potential to capture deep linguistic nuances, but require greater computational investment.

This work performs a systematic comparison between these three approaches, evaluating their performance on the multiclass classification task of legal petitions in Portuguese, originating from a Brazilian Public Defender's Office [DPE-GO 2025], in which the shortage of professionals [ANADEP 2021] and the complexity of legal texts create challenges for automation. We introduce a new Portuguese dataset, with the aim of fostering future research in legal Natural Language Processing (NLP). To our knowledge, this is one of the first studies to compare in an integrated manner models based on prompting, traditional ML, and BERT fine-tuning for legal documents in the Portuguese language [Aguiar 2025].

2. Related Work

Language models have been trained on large volumes of unannotated text, exploiting self-supervised learning techniques. According to their architecture, these models can be grouped into three categories. The first encompasses encoder-based models, such as BERT, which follows the pre-training followed by fine-tuning paradigm for NLP tasks. During pre-training, masked language modeling prediction is used, and subsequently, the model is adapted with labeled data for specific tasks [Devlin et al. 2019, Liu et al. 2019a, Sun et al. 2020, Clark et al. 2019, Liu et al. 2019b].

The second category comprises decoder-only models, such as the Generative Pre-trained Transformer (GPT), which adopts an autoregressive approach to predict the next token in a text sequence [Radford et al. 2019]. Models like GPT-3, DeepSeek, Google Gemini [Rahman et al. 2025] formalize NLP tasks as text generation problems conditioned on an input prompt [Brown et al. 2020]. The third category involves architectures with combined encoder and decoder components, such as the Text-to-Text Transfer Transformer (T5), which unifies multiple tasks under a single text input and output structure [Raffel et al. 2023, Lewis et al. 2019, Xue et al. 2021].

LLMs have been utilized in legal scenarios to classifying documents through different prompting strategies. These models demonstrate the ability to handle complex linguistic constructions, making them suitable for domains where semantic precision is important [Radford et al. 2019, Dai et al. 2019, Keskar et al. 2019]. Furthermore, recent prompting strategies, such as the Clue And Reasoning Prompting (CARP) method, have been proposed to address nuanced linguistic challenges in legal text classification, such as distinguishing between similar legal outcomes based on subtle argumentative differences, for instance, discerning between a granted and a partially granted habeas corpus [Sun et al. 2023]. Despite these advancements, traditional supervised learning approaches continue to play a role in legal NLP tasks, including contexts with limited resources or data. Several studies have explored such methods in legal settings.

In parallel to these innovations in LLMs, classical ML techniques have also remained and continue to be investigated in legal NLP research. For instance, Sil and Roy [Sil and Roy 2021] proposed a system based on multiple classical classifiers (including Multi-Layer Perceptron (MLP), k-Nearest Neighbors (KNN), and Support Vector

Machine (SVM)) for defendant identification in domestic violence cases in India, with promising results on datasets collected from state courts. Chen et al. [Chen et al. 2022] conducted a comparative study between random forest and deep learning models, demonstrating that, in certain legal contexts in the USA, models based on specific conceptual features can outperform deep neural networks when accompanied by efficient feature selection.

On the other hand, Trautmann [Trautmann 2023] introduces the concept of prompt chaining for extensive legal documents, utilizing intermediate summaries and contextual examples to guide language models toward more precise classifications. These studies illustrate the diversity of strategies currently in use for automated legal text processing, emphasizing the importance of comparative evaluations between prompt-based models, traditional ML techniques, and fine-tuning of Transformer architectures, as conducted in this work. Furthermore, the use of a Portuguese language dataset, still underexplored in the literature, expands the empirical scope in the field of legal NLP.

3. Methodology

The case study was based on a dataset of 3,458 legal documents in Portuguese, provided by the DPE-GO and manually labeled into 24 procedural categories. Due to restrictions imposed by the General Data Protection Law (LGPD) and institutional guidelines, the original data could not be released publicly. To support the experiments, a synthetic corpus was generated using the Sabiá-3.1 language model [Abonizio et al. 2024], preserving the structural characteristics of the original dataset while ensuring the absence of sensitive or identifiable information.

Table 1 presents columns indicating the total number of documents (**Total**), the training subset used in generating summaries and centroids (**Training Set**), the test subset employed in the classification task (**Test Set**), the average word count (**Avg. Words**), the standard deviation of words (**Std. Words**), the average character count (**Avg. Chars**), and the standard deviation of characters (**Std. Chars**). This structuring was adopted in approaches with prompt-based models, which operate on representations synthesized from the training set, while the test set remained reserved for inference. For experiments with ML algorithms and fine-tuning, the data were used as detailed in Subsections 3.3 and 3.4. The text length statistics were calculated from the complete set of original documents (**Total**), disregarding the subsets used for training and testing.

3.1. Data Preprocessing and Normalization

The documents were converted `.docx` format to ensure structural uniformity, given the diversity of original formats, such as `.odt`, `.pdf`, and `.html` [Modrušan et al. 2020]. In the text cleaning stage, non-textual elements, HyperText Markup Language (HTML) tags, and other artifacts were removed [Pichiyan et al. 2023]. The linguistic normalization process included case standardization (uppercase and lowercase), accent removal, lemmatization, and stopword removal, following best practices for improving the quality of text vectors used in NLP models [Palanivinayagam et al. 2023]. The resulting data were organized in tabular format, with unique identifiers and respective procedural categories. To meet legal data protection requirements, an anonymization process based on regular expressions was applied,

aimed at removing sensitive information, such as proper names, personal documents (Individual Taxpayer Registry (CPF), General Registry (RG)), addresses, and financial data. Anonymization quality was ensured via automated checks complemented by manual review. The final dataset, already anonymized, was stored in Comma Separated Values (.CSV) and `.XLSX` formats, to facilitate its integration with classification pipelines.

Table 1. Number of texts and text size statistics per legal category across full, evaluation, and classification sets

Category	Total	Training Set	Test Set	Avg. Words	Std. Words	Avg. Chars	Std. Chars
EXTINCTION-OF-PUNISHABILITY	500	100	400	447.5	483.8	3684.5	3972.5
APPEAL	441	90	351	1430.2	660.1	11807.7	5493.5
CHALLENGE	425	75	350	780.8	497.6	6414.2	4078.8
EMBARGOES	304	70	234	652.6	453.8	5322.2	3707.6
APELACAO	296	65	231	1029.5	554.2	8332.2	4514.0
CIVIL-REGISTRY	236	50	186	677.4	412.1	5443.3	1564.0
PAROLE	218	45	173	523.2	493.7	4411.3	3348.6
PARDON-COMMUTATION	170	40	130	502.6	368.2	5022.6	3682.3
ENFORCEMENT-OF-JUDGMENT	150	35	113	565.6	174.3	4542.4	1415.2
OFFICIAL-LETTERS	147	30	117	192.2	90.5	1653.2	940.5
DAMAGES	143	30	113	1879.6	835.7	15135.4	6731.7
PRE-ENFORCEMENT-OBJECTION	86	20	78	1303.5	619.0	10743.1	5286.6
TRANSFER-OF-EXECUTION	85	20	65	1022.4	493.1	2527.1	2231.4
ADVERSE-POSSESSION	42	20	22	1226.6	400.6	9649.6	3182.8
SENTENCE-UNIFICATION	35	15	24	486.8	412.2	3941.7	3223.2
HABEAS-CORPUS	28	15	21	1723.7	675.7	14468.1	5780.1
NEGATIVE-NOTIFICATION	26	15	11	102.4	97.6	885.1	456.7
CONDOMINIUM-DISSOLUTION	26	15	11	898.0	341.9	7883.1	2660.1
SENTENCE-REMISSION	25	15	10	369.2	299.4	3908.9	2409.6
SENTENCE-PROGRESSION	19	15	4	208.4	174.1	1788.4	1462.7
PAYMENT-INTO-COURT	16	10	8	386.7	298.5	8161.1	2779.7
COUNTER-ARGUMENT-TO-APPEAL	15	10	5	986.7	613.7	8030.2	5019.2
TEMPORARY-RELEASE	13	10	3	455.8	309.3	3867.1	2531.1
JUDICIAL-ORDER-OF-RELEASE	12	10	2	526.1	93.9	4251.9	697.0

3.2. Exploratory Analysis and Feature Engineering for Baseline Classification

To understand the intrinsic complexity of the legal document classification task and establishing a baseline using traditional supervised learning algorithms, an exploratory analysis of class separability was conducted, followed by the definition of structured features.

Linear separability between procedural categories was investigated through the projection of document vectors into a two-dimensional space. Each petition was represented as a sparse vector based on TF-IDF weights, which reflect the relative relevance of terms in the document compared to the rest of the corpus. For dimensionality reduction, the t-Distributed Stochastic Neighbor Embedding (t-SNE) algorithm was applied [Balamurali 2021], enabling the visualization of document distribution in a two-dimensional plane. Figure 1 illustrates the points corresponding to documents, color-coded according to their class. The analysis revealed overlap between different categories, with greater confusion observed among those with semantic or contextual similarity, sug-

gesting low linear separability. This result indicates potential limitations of methods based on simple vectorial representations for this task.

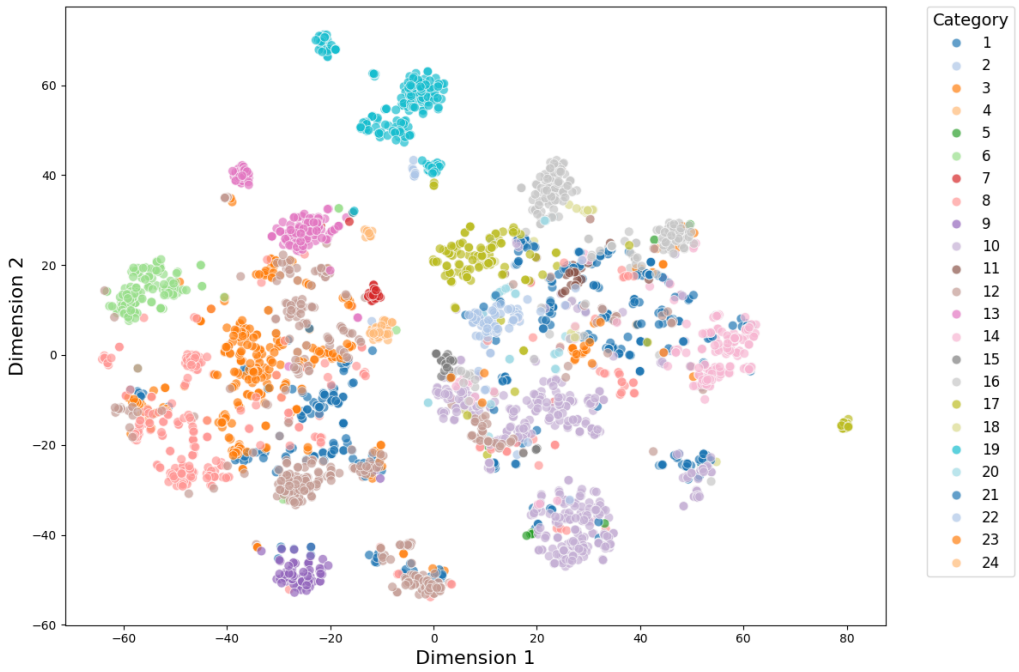

Figure 1. 2D t-SNE projection of document embeddings using TF-IDF. Colors represent different legal categories. Overlap between classes suggests low separability

For the construction of baseline classification models, including logistic regression, KNN, and Naive Bayes, two types of features were defined. The first corresponds to the TF-IDF representation of documents, configured with minimum and maximum frequency parameters to filter common or rare terms, and with support for unigrams and bigrams, allowing the capture of local dependencies between words. The second feature considered was the total document length, measured in number of words, with the hypothesis that variations in textual size may reflect differences between procedural classes, given the diversity of legal requirements and argumentative contexts.

These descriptors, by combining lexical and structural information, provided an interpretable and easily implementable foundation to feed the supervised models evaluated in this study.

3.3. Experimental Configuration Conventional ML Models

This work evaluated three supervised algorithms used in classification contexts, logistic regression [LaValley 2008], KNN [Peterson 2009], and Naive Bayes [Webb 2017]. These models were chosen due to their low computational complexity and adequate interpretability, compatible with the demands of the proposed problem. Experimental validation was conducted through stratified five-fold cross-validation, ensuring the maintenance of class proportions and providing consistent performance estimates, without requiring high-performance infrastructure.

The experiments involving traditional models included hyperparameter tuning for each algorithm. For Naive Bayes, the smoothing parameter α was tested across values

ranging from 0.01 to 10, along with the evaluation of whether to incorporate prior class probability adjustment. For K-Nearest Neighbors, the number of neighbors was varied between 3 and 19, using both uniform and distance-based weighting schemes, and considering three distance metrics, Euclidean, Manhattan, and cosine. The search algorithm and leaf size were maintained at their default configurations. In the case of logistic regression, an optimizer suitable for large datasets and regularized models was used, with L2 penalty, up to 1,000 training iterations, and a fixed regularization strength of 1.0. Multiclass classification was performed using through a one-vs-rest strategy.

3.4. Fine-Tuning Pretrained Transformers

This study fine-tuned three Transformer architectures from `HuggingFace` [Jain 2022], namely BERTimbau Large, BERTimbau Base [Souza et al. 2020], and XLM-RoBERTa [Goyal et al. 2021]. Models were evaluated on Portuguese legal text classification with 24 procedural categories. Tokenization used respective `AutoTokenizers` with 512-token maximum length, truncation, and padding. This limit complies with BERT architectural constraints, and although truncation may cause content loss in longer documents, prior analyses showed classification features appear in opening segments. Attention masks ensured consistent gradient propagation. The dataset was split using stratified sampling into 80% for training, 10% for validation, and 10% for testing, with a fixed random seed for reproducibility. Training used supervised learning with accuracy, precision, recall, and weighted F1-score monitoring. Early stopping applied after three epochs without validation loss improvement. Hyperparameters were standardized across models, including the AdamW optimizer with a learning rate of 1×10^{-5}, epsilon of 1×10^{-8}, weight decay of 0.2 (excluding normalization and bias terms), batch size of 32, linear learning rate decay, and gradient clipping at an L2-norm of 1.0. Training used Graphics Processing Unit (GPU) acceleration with Compute Unified Device Architecture (CUDA) support. Model parameters and evaluation metrics were serialized for reproducibility.

The codes implemented in this study, including the classification routines, experimental setup, and synthetic data, are available in the `GitHub` repository.[1]

3.5. Setup of Prompt-Based Models

The approaches evaluated in this stage combine text preprocessing, instruction formulation, and zero-shot classification through LLMs. Although initially referred to as textual representation strategies, they differ from classical techniques such as TF-IDF and embeddings, as they incorporate instructional elements and contextual inference guided by examples. For greater terminological clarity, we adopt the term prompt-based input strategies.

Model selection considered three criteria, namely Application Programming Interface (API) accessibility, support for extended context windows suitable for extensive legal texts, and proven performance in multilingual tasks. Models from the Large Language Model Meta AI (LLaMA-3) family [Meta 2024] were used, along with smaller variants such as Mixtral-8x22B [Mixtral 8x22B 2024], Mistral-7B [Mixtral 8x7B 2024],

[1] `https://github.com/Willgnner-Santos/llm-legal-benchmark-pt`

and LLaMA-3.2-3B [Llama-3.2-3B 2024], all executed through the Together AI platform [Together AI 2025], ensuring experimental uniformity. The choice for prompt-based techniques are justified by their by their adaptation capability without requiring re-training, an aspect in contexts with computational resource constraints and scarcity of labeled data.

In the **original text-based strategy**, the model receives the complete document content, followed by a direct instruction to classify it into one of the predefined categories. A typical prompt follows the structure: *"Classify the following legal petition into one of the predefined categories. Return only the category."* The full list of categories was embedded in the prompt. The model's output, a predicted label, was then normalized for comparison and evaluation. Despite total preservation of semantic and contextual content [Wan et al. 2019], computational cost is high (average of 1,614 tokens) and accuracy may be impacted by noise in extensive texts. The **summary-based strategy** reduces documents to condensed versions of approximately 200 words [de Jesus Falcão et al. 2024], generated from centroid representations identified via TF-IDF and cosine similarity. During inference, the model receives a prompt containing category summaries and a new text to be classified. This approach is token-efficient but may lose nuances in complex cases.

In **centroid-based classification**, the summarization process is avoided. A representative document is selected for each category, also based on similarity measures, and inserted directly into the prompt. The model compares the input text with centroids to decide the appropriate category [Chai et al. 2020]. This technique reduces input variability but may be sensitive to internal class diversity. Finally, the **description-based strategy** utilizes semantic definitions of categories, written by public defenders and refined with GPT-4 support [Xu et al. 2022]. Descriptions are incorporated into the prompt as interpretive reference, and the model performs classification based on conceptual alignment between input text and defined criteria. While enabling conceptual transparency through explicit category definitions, this technique requires more extensive prompts and, therefore, greater processing capacity.

4. Results and Discussion

This section presents the analysis of results obtained with the different evaluated models, including prompt-based strategies, traditional supervised learning algorithms, and fine-tuned Transformer models. The objective is to understand the interactions between representational capacity, computational cost, and adaptability, considering the challenge of multiclass classification of legal documents. Table 2 summarizes the weighted average values of precision, recall, and F1-score for each configuration.

Fine-tuned Transformer models outperformed all other approaches, achieving the highest F1-scores due to their ability to capture deep linguistic patterns. The **BERTimbau Large** model achieved the highest F1-score in the analysis, with **94.70%**, evidencing the capacity of deep encoder architectures, pre-trained with large volumes of data in Portuguese, to capture linguistic nuances in the legal context. The **BERTimbau Base** model, although lighter, also obtained competitive results, surpassing all prompt-based models. The **XLM-RoBERTa**, despite being a robust and multilingual model, performed below the fine-tuned monolingual models. This suggests that generalization to multiple languages may reduce the capacity to model specific patterns of the Portuguese language

in scenarios with restricted fine-tuning data. Among the input strategies evaluated with LLMs, the **summary-based approach** presented the best results. The **LLaMA-3.1-8B** model achieved an F1-score of **88.52%** with this configuration, reaching performance comparable to comparable to that of fine-tuned models in scenarios with limited data. Text reduction to a condensed version appears to minimize noise without compromising semantic expressiveness, favoring more precise decisions.

Table 2. Performance comparison of different LLMs with input strategies, alongside traditional ML and fine-tuned Transformer models

Model	Input Type / Strategy	Precision	Recall	F1-score
LLMs				
Llama-3.1-8B	Summaries	90.75%	91.31%	**88.52%**
	Centroids	70.12%	59.86%	54.54%
	Descriptions	92.94%	88.64%	89.98%
	Full Text	75.90%	49.48%	51.92%
Mixtral-8x22B	Summaries	83.19%	86.54%	83.88%
	Centroids	85.49%	89.90%	86.98%
	Descriptions	72.07%	52.49%	54.11%
	Full Text	84.06%	79.38%	80.31%
Mistral-7B	Summaries	72.45%	82.16%	75.90%
	Centroids	24.45%	31.45%	22.52%
	Descriptions	47.15%	29.53%	26.34%
	Full Text	56.46%	45.14%	41.86%
Mixtral-8x7B	Summaries	68.88%	79.26%	72.46%
	Centroids	54.78%	64.39%	57.02%
	Descriptions	72.07%	52.49%	54.11%
	Full Text	71.73%	48.79%	47.53%
Llama-3.2-3B	Summaries	72.38%	77.24%	71.19%
	Centroids	29.05%	21.31%	13.39%
	Descriptions	51.34%	30.62%	28.42%
	Full Text	41.75%	21.92%	16.65%
Traditional ML Models				
KNN	Structured Features	90.52%	84.03%	**85.74%**
Logistic Regression	Structured Features	86.63%	75.48%	78.97%
Naive Bayes	Structured Features	51.32%	47.59%	47.75%
Fine-Tuned Transformer Models				
BERTimbau Large	Fine-Tuning	94.23%	95.37%	**94.70%**
BERTimbau Base	Fine-Tuning	89.57%	92.77%	90.91%
XLM-RoBERTa	Fine-Tuning	71.14%	73.98%	70.47%

The **centroid strategy**, in turn, had competitive performance in higher-capacity models, such as **Mixtral-8x22B**, which achieved an F1-score of **86.98%**. This result suggests that when the architecture has sufficient capacity to process extensive examples,

the use of representative texts per class may be more advantageous than summarizations or abstract descriptions. The approach with **original text-based strategy** (Full Text), although preserving the original content of documents, showed lower performance across all models. This suggests that excessive document length and narrative variation can introduce noise into the classification process, affecting prediction accuracy, complex structures, and large discursive variations may introduce noise and hinder the identification of the correct category, even in large-scale LLMs.

The **description-based strategy** showed model-dependent variation, performing effectively in **LLaMA-3.1-8B** but poorly in smaller architectures such as **Mistral-7B**, **LLaMA-3.2-3B**, and **Mixtral-8x7B**. This confirms that conceptual strategies require semantic inference and contextual retention capabilities absent in smaller models. Summary-based approaches reduce inference cost, averaging 90–100 tokens per document versus over 1,600 tokens for centroid-based and full-text strategies. Classifying 2,600 documents costs less than US$1 with summaries compared to over US$14 for full-text inputs, demonstrating clear cost-effectiveness. Traditional models achieved results, with **KNN** with F1-score above 85% despite limited feature depth. **Logistic regression** maintained stable cross-fold performance with moderate precision-recall balance, while **Naive Bayes** performed worst due to conditional independence assumptions inadequate for legal contexts with high term co-occurrence.

Pearson correlation analysis reinforces **F1-score** as the primary metric, showing high correlations with recall (≈ 0.98) and precision (≈ 0.91), adequately balancing both factors. The precision-recall correlation (≈ 0.87) indicates result coherence across strategies. F1-score is justified as the central metric given the legal context, where classification errors impact document triage. Classification strategy selection depends on application context. Fine-tuned Transformers offer best performance with available labeled data and infrastructure. When fine-tuning is unfeasible, LLMs with summaries strike an effective balance between accuracy and inference cost. Centroid input suits higher-capacity models with representative class examples. Traditional models (KNN, logistic regression) remain useful for low-cost scenarios and interpretability requirements. Complete text usage harms performance due to noise and redundancy. Description-based strategies require conceptual inference capabilities, being suitable for explanatory tasks or smaller volumes.

5. Conclusion

This study conducted a comparative evaluation between three paradigms for multi-class classification of legal documents in Portuguese, namely traditional supervised models with structured vectors, fine-tuned Transformer architectures, and prompt-based approaches with LLMs. The results indicate that fine-tuning with models such as BERTimbau Large achieves superior performance, provided there are labeled data and training infrastructure. LLM-based strategies proved to be cost-effective and adaptable alternatives for scenarios lacking labeled data or infrastructure, with emphasis on summary-based input, which reconciled high performance and token economy, which is useful in contexts with computational constraints. On the other hand, complete text representations exhibited reduced classification performance, likely due to input length and redundancy, while description or centroid approaches depended on the capacity of the models used. Traditional models such as KNN and logistic regression demonstrated competitive F1-scores in constrained environments, supporting their applicability in operational systems, while of-

fering interpretability and consistent performance. As limitations, we highlight the use of generic LLMs without specific legal training, beyond class imbalance. Future work may investigate hybrid strategies, such as the integration of LLMs with conventional classifiers or the use of, Retrieval-augmented generation (RAG).

6. Acknowledgements

This work was supported by the National Institute of Science and Technology (INCT) in Responsible Artificial Intelligence for Computational Linguistics and Information Treatment and Dissemination (TILD-IAR) [grant number 408490/2024-1].

References

[Abonizio et al. 2024] Abonizio, H., Almeida, T. S., Laitz, T., Junior, R. M., Bonás, G. K., Nogueira, R., and Pires, R. (2024). Sabiá-3 technical report. *arXiv preprint arXiv:2410.12049*.

[Aguiar et al. 2021] Aguiar, A., Silveira, R., Pinheiro, V., Furtado, V., and Neto, J. A. (2021). Text classification in legal documents extracted from lawsuits in brazilian courts. In *Anais da X Brazilian Conference on Intelligent Systems*, Porto Alegre, RS, Brasil. SBC.

[Aguiar 2025] Aguiar, M. S. d. (2025). Comparative analysis of the performance of large language models in the classification of legal texts.

[ANADEP 2021] ANADEP (2021). Goiás is the second worst brazilian state in number of public defenders per inhabitant. `https://anadep.org.br/wtk/pagina/materia?id=49422`.

[Balamurali 2021] Balamurali, M. (2021). T-distributed stochastic neighbor embedding. In *Encyclopedia of mathematical geosciences*, pages 1–9. Springer.

[Berman et al. 2021] Berman, E. M., Bowman, J. S., West, J. P., and Van Wart, M. R. (2021). *Human resource management in public service: Paradoxes, processes, and problems*. Cq Press.

[Brown et al. 2020] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners.

[Chai et al. 2020] Chai, Y., Zhang, H., and Jin, S. (2020). Neural text classification by jointly learning to cluster and align.

[Chen et al. 2022] Chen, H., Wu, L., Chen, J., Lu, W., and Ding, J. (2022). A comparative study of automated legal text classification using random forests and deep learning. *Information Processing & Management*, 59(2):102798.

[Clark et al. 2019] Clark, K., Khandelwal, U., Levy, O., and Manning, C. D. (2019). What does bert look at? an analysis of bert's attention.

[Dai et al. 2019] Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q. V., and Salakhutdinov, R. (2019). Transformer-xl: Attentive language models beyond a fixed-length context.

[de Jesus Falcão et al. 2024] de Jesus Falcão, L. C. et al. (2024). Sumarização de texto em deep learning como etapa inicial para a construção de um modelo de recuperação da informação: análise do setor de mineração no brasil.

[Devlin et al. 2019] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding.

[DPE-GO 2025] DPE-GO (2025). Public defender's office of the state of goiás. http://www2.defensoria.go.def.br/.

[Elov et al. 2023] Elov, B., Khamroeva, S. M., and Xusainova, Z. (2023). The pipeline processing of nlp. In *E3S Web of Conferences*, volume 413, page 03011. EDP Sciences.

[Goyal et al. 2021] Goyal, N., Du, J., Ott, M., Anantharaman, G., and Conneau, A. (2021). Larger-scale transformers for multilingual masked language modeling. *CoRR*, abs/2105.00572.

[Jain 2022] Jain, S. M. (2022). Hugging face. In *Introduction to transformers for NLP: With the hugging face library and models to solve problems*, pages 51–67. Springer.

[Keskar et al. 2019] Keskar, N. S., McCann, B., Varshney, L. R., Xiong, C., and Socher, R. (2019). Ctrl: A conditional transformer language model for controllable generation.

[Kiesow Cortez and Maslej 2023] Kiesow Cortez, E. and Maslej, N. (2023). Adjudication of artificial intelligence and automated decision-making cases in europe and the usa. *European Journal of Risk Regulation*, 14(3):457–475.

[LaValley 2008] LaValley, M. P. (2008). Logistic regression. *Circulation*, 117(18):2395–2399.

[Lewis et al. 2019] Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension.

[Liu et al. 2019a] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019a). Roberta: A robustly optimized bert pretraining approach.

[Liu et al. 2019b] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019b). Roberta: A robustly optimized bert pretraining approach.

[Llama-3.2-3B 2024] Llama-3.2-3B (2024). Llama-3.2-3b technical report. https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct.

[Meta 2024] Meta (2024). Meta ai introduces llama 3.1: Advanced capabilities in language modeling. https://ai.meta.com/blog/meta-llama-3-1/?utm_source=twitter&utm_medium=organic_social&utm_content=video&utm_campaign=llama31&s=08.

[Mills and Uebergang 2017] Mills, M. and Uebergang, J. (2017). Artificial intelligence in law: An overview.

[Mixtral 8x22B 2024] Mixtral 8x22B (2024). Mixtral-8x22b technical report. https://huggingface.co/mistralai/Mixtral-8x22B-Instruct-v0.1.

[Mixtral 8x7B 2024] Mixtral 8x7B (2024). Mixtral-8x7b technical report. https://huggingface.co/mistralai/Mixtral-8x7B-Instruct-v0.1.

[Modrušan et al. 2020] Modrušan, N., Rabuzin, K., and Mrsic, L. (2020). Improving public sector efficiency using advanced text mining in the procurement process. In *Proceedings of the 12th International Conference on e-Business (ICE-B)*, pages 200–206. SCITEPRESS.

[Moraes et al. 2024] Moraes, L. d. C., Silvério, I. C., Marques, R. A. S., Anaia, B. d. C., de Paula, D. F., de Faria, M. C. S., Cleveston, I., Correia, A. d. S., and Freitag, R. M. K. (2024). Análise de ambiguidade linguística em modelos de linguagem de grande escala (llms). *arXiv preprint arXiv:2404.16653*.

[Nonato 2022] Nonato, L. G. (2022). O cenário regulatório da inteligência artificial.

[Palanivinayagam et al. 2023] Palanivinayagam, A., El-Bayeh, C. Z., and Damaševičius, R. (2023). Twenty years of machine-learning-based text classification: A systematic review. *Algorithms*, 16(5).

[Pandey and Malik 2022] Pandey, D. and Malik, N. S. (2022). Artificial intelligence, automation, and the legal system. In *Legal Analytics*, pages 1–10. Chapman and Hall/CRC.

[Peterson 2009] Peterson, L. E. (2009). K-nearest neighbor. *Scholarpedia*, 4(2):1883.

[Pichiyan et al. 2023] Pichiyan, V., Muthulingam, S., G, S., Nalajala, S., Ch, A., and Das, M. N. (2023). Web scraping using natural language processing: Exploiting unstructured text for data extraction and analysis. *Procedia Computer Science*, 230:193–202. 3rd International Conference on Evolutionary Computing and Mobile Sustainable Networks (ICECMSN 2023).

[Radford et al. 2019] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners.

[Raffel et al. 2023] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2023). Exploring the limits of transfer learning with a unified text-to-text transformer.

[Rahman et al. 2025] Rahman, A., Mahir, S. H., Tashrif, M. T. A., Aishi, A. A., Karim, M. A., Kundu, D., Debnath, T., Moududi, M. A. A., and Eidmum, M. (2025). Comparative analysis based on deepseek, chatgpt, and google gemini: Features, techniques, performance, future prospects. *arXiv preprint arXiv:2503.04783*.

[Sil and Roy 2021] Sil, R. and Roy, A. (2021). Machine learning approach for automated legal text classification. *International Journal of Computer Information Systems and Industrial Management Applications*, 13:10–10.

[Souza et al. 2020] Souza, F., Nogueira, R., and Lotufo, R. (2020). Bertimbau: pretrained bert models for brazilian portuguese. In *Brazilian conference on intelligent systems*, pages 403–417. Springer.

[Sun et al. 2020] Sun, C., Qiu, X., Xu, Y., and Huang, X. (2020). How to fine-tune bert for text classification?

[Sun et al. 2023] Sun, X., Li, X., Li, J., Wu, F., Guo, S., Zhang, T., and Wang, G. (2023). Text classification via large language models.

[Together AI 2025] Together AI (2025). Together ai: Open foundation models api and cloud platform. Accessed: 2025-06-15.

[Trautmann 2023] Trautmann, D. (2023). Large language model prompt chaining for long legal document classification. *arXiv preprint arXiv:2308.04138*.

[Wan et al. 2019] Wan, L., Papageorgiou, G., Seddon, M., and Bernardoni, M. (2019). Long-length legal document classification.

[Webb 2017] Webb, G. I. (2017). Naïve bayes. In *Encyclopedia of machine learning and data mining*, pages 895–896. Springer.

[Xu et al. 2022] Xu, S., Zhang, C., and Hong, D. (2022). Bert-based nlp techniques for classification and severity modeling in basic warranty data study. *Insurance: Mathematics and Economics*, 107:57–67.

[Xue et al. 2021] Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., and Raffel, C. (2021). mt5: A massively multilingual pre-trained text-to-text transformer.