Prediction of Conciliation Outcomes in Judicial Decisions at the Labor Court of Espírito Santo

1st Luís Câmara
DI - Department of Computer Science
UFES
Espirito Santo, Brazil
luis.camara@edu.ufes.br

***Abstract*—This paper compares methodologies for predicting conciliation in labor cases, integrating large language models (LLMs) and classical natural language processing (NLP) techniques with structured data analysis. The aim is to alleviate the burden on the judicial system and optimize resources through early conciliation predictions. Experiments indicate that LLM-based models outperformed traditional methods, achieving an 8.3% improvement in the Macro F1-Score. These results highlight the potential of LLMs to enhance efficiency in judicial practice.**

***Index Terms*—Conciliation Prediction, Labor Cases, Large Language Models, Natural Language Processing, Judicial Efficiency.**

I. Introduction

The Brazilian judiciary faces an increasing procedural overload. In the Labor Court, for instance, about 4.2 million new cases were initiated in 2023, representing a 28.7% increase compared to the previous year [1]. Conciliation serves as an effective release valve, with 12% of cases resolved through conciliation, reaching up to 36.5% in the Labor Court [2], [3]. Recent studies have shown that artificial intelligence (AI) techniques and natural language processing (NLP) can predict judicial outcomes with high accuracy (approximately 79% and 75% in distinct studies) [4]– [6]. These results encourage the use of such approaches to identify cases with a high probability of conciliation in advance.

This paper compares classical approaches and large language models (LLMs) for predicting conciliation outcomes in labor cases in Espírito Santo. Given a set of procedural data—including party information, case characteristics, and text documents—the objective is to predict whether a case will be resolved via settlement or by judicial decision.

A. Problem Definition

Due to the procedural overload in the Labor Court, it is essential to identify early whether a case will be resolved by conciliation or by a judicial sentence. This prediction assists in prioritizing and allocating judicial efforts, contributing to a faster and more efficient resolution of conflicts. In summary, the problem addressed is to predict, based on the data from a labor case, whether the outcome will be a conciliation or a judicial sentence.

II. Related Work

Predicting judicial outcomes using AI and NLP techniques has been widely investigated in both international and national contexts. Internationally, Aletras *et al.* [4] demonstrated that extracting textual features can yield robust predictions for decisions of the European Court of Human Rights, while Katz *et al.* [5] explored machine learning algorithms to analyze the behavior of the U.S. Supreme Court.

In Brazil, the study "Predicting Brazilian Court Decisions" by Silva *et al.* [7] adapted these techniques to local cases, and the development of Legal-BERT by Chalkidis *et al.* [8] highlighted the potential of pre-trained language models for extracting semantic information from legal texts.

III. Dataset Description

The dataset used in this work gathers essential information for predicting conciliation in labor cases in Espírito Santo, organized into three main categories:

- **Identification and Classification:** Data that organize the cases, such as a unique case identifier and the type or category of litigation.
- **Economic Characteristics and Party Data:** Information regarding the value of the case, the nature of the parties (public or private), and the number of claimants and defendants, reflecting the economic scale and complexity of the case.
- **Temporal and Outcome Data:** Information on filing and judgment dates, along with a binary variable indicating whether a conciliation was reached (1) or not (0). These data are fundamental for analyzing temporal patterns and outcomes.

For the analysis, only the following textual features were used: *Vara do Trabalho, Ramo de Atividade, Classe Processual, Cidade de Origem da Petição Inicial, OAB, Assuntos, RECDA PES FÍS OU JUR, Portador de Deficiência, Segredo de Justiça, RECDA Ativa-Inativa, Ente Público ou Privado, Indicador do Processo, Documentos das Reclamadas e Documentos dos Reclamantes*.

Additionally, to capture the temporal evolution of cases and potential trends related to judgment dates, the data were divided into sliding windows based on the judgment date [9]. Each window spans two years of data and advances every

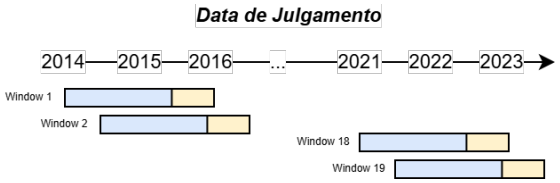

Fig. 1. Overlapping sliding windows strategy on *Data de Julgamento* with a total of 2 years of data and 6 months of displacement, with train (blue) and test (yellow) with 70%/30%

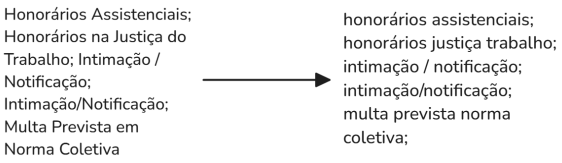

Fig. 2. Difference between the original text (left) and the pre-processed text (right).

six months. Within each window, the dataset was shuffled and splited into 70% for training and 30% for testing, such as Fig. 1, ensuring a robust and representative evaluation of predictive performance. The window size was chosen based on achieving a better data distribution.

IV. Methodology and Experiments

This section describes the methodology used to predict conciliation by combining classical NLP techniques with LLM-based approaches. This combination leverages both the a priori knowledge embedded in manually engineered features and statistical models specific to the legal domain, as well as the capacity of pre-trained language models to capture complex semantic and contextual nuances in texts.

For comparison purposes, the classical approach was used as a baseline for the LLM models. Data pre-processing and splitting into windows were performed as described in Section IV.

A. Pre-processing and Window Division

Initially, textual pre-processing methods were applied to standardize the data and remove potential noise, ensuring higher quality for model training. The following steps were applied to all textual features:

1) Removal of entries with null feature values;
2) Removal of duplicates (using the case number);
3) Conversion of all letters to lowercase;
4) Tokenization of textual data;
5) Removal of stopwords [10].

As a result, standardized texts were obtained, as illustrated in Fig. 2.

Subsequently, the data were segmented into sliding time windows as described earlier. A total of 19 windows (each spanning two years and advancing by six months) were used, covering the period from 2014 to 2023. In each window, the cases were randomly divided into 70% for training and 30% for testing.

B. Classical Methodology (Baseline)

After pre-processing and window division, a target encoding technique with smoothing was applied to the categorical textual features. Many of these features have repeated values; thus, numerically categorizing them reduces the risk of overfitting, especially for low-frequency categories. For a given category c, the encoding is defined as:

$$\mathrm{TE}(c) = \frac{n_c\,\mu_c + \alpha\,\mu}{n_c + \alpha},$$

where:

- n_c is the number of occurrences of category c in the training set;
- μ_c is the average target value for category c;
- μ is the global average target value in the training set;
- α is the smoothing parameter.

For records containing multiple categories (e.g., values separated by a delimiter, as shown in Fig. 2), the encoding for a record is obtained as the average of the individual encodings:

$$\mathrm{TE}(x) = \frac{1}{k}\sum_{j=1}^{k}\mathrm{TE}(c_j),$$

where $\{c_1, c_2, \ldots, c_k\}$ is the set of categories present in entry x.

For each time window, three classification models (with default parameters) were trained using the transformed data:

- Logistic Regression,
- Random Forest,
- Gradient Boosting.

After training, the models were evaluated based on performance metrics such as accuracy, precision, recall, and F1-Score. The training and testing workflow is depicted in Fig. 3.

C. Methodologies with LLM

This subsection describes the methodological approach for text classification using LLMs. Two strategies were explored: (1) feature extraction using a pre-trained LLM followed by a traditional classifier, and (2) direct text classification using a fine-tuned LLM via the HuggingFace framework [11].

1) Feature Extraction with LLM: For this approach, embeddings were extracted using a pre-trained language model, followed by the application of a Random Forest classifier. After data preparation, the pre-trained BERTimbau model [12]—chosen for its specialization in the Portuguese language to capture linguistic nuances [13], this can be seen better in Fig 6, where we compare the attention heads with a english based BERT, where the BERTimbau model tokenized and got a cleaner context compared with BERT base—was used to convert texts into high-dimensional vector representations.

Initially, a text column was created by concatenating the selected categorical variables. The texts were then processed using a tokenizer that applies truncation and padding, ensuring a maximum input length of 512 tokens. This step is crucial for maintaining input consistency and preventing memory issues.

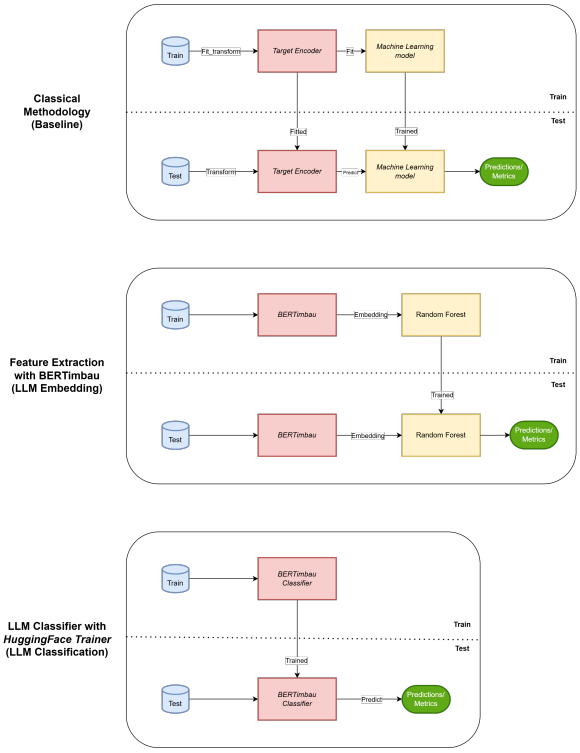

Fig. 3. Workflow diagram comparing the different methodologies

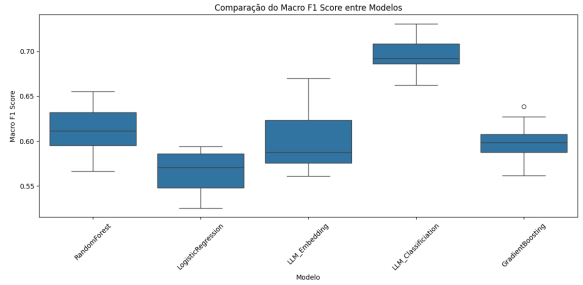

Fig. 4. Comparison of the Macro F1-Score across all time windows for each model.

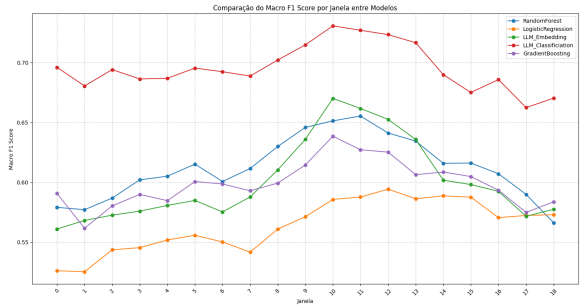

Fig. 5. Boxplot of F1-Macro scores for the evaluated models per time window.

Finally, the Random Forest classifier, which yielded the best results in the classical approach, was used.

The tokenized texts were fed into the BERT model to extract embeddings. The extraction was performed in batches, allowing for GPU usage when available, to optimize processing.

2) LLM Classification Using HuggingFace Trainer: In this approach, the HuggingFace Transformers framework was employed to fine-tune a text classification model end-to-end. The methodology integrates text tokenization, conversion into the HuggingFace Dataset format, and training of a classification model using the `Trainer` class.

After data preparation, an instance of a pre-trained BERTimbau classification model (via `AutoModelForSequenceClassification`) was loaded and configured for the desired number of labels (in this case, two classes: Conciliation or non-conciliation). Training arguments were then defined using the `TrainingArguments` class, adopting the same hyperparameters used in Silva *et al.* [7], where they predicted court decisions based on textual features.

A `Trainer` instance was configured with the model, training arguments, training and validation datasets, and a metric computation function (accuracy, precision, recall, and F1-Score). Training was conducted for each time window, and after training, the model was evaluated on the test set.

V. Results

The experimental results obtained from the predictive models described in Section V were applied to the dataset using sliding windows defined by judgment dates. Overall, the LLM-based classification model outperformed the classical methodologies.

Fig. 4 illustrates the Macro F1-Score across time windows for five approaches: Logistic Regression, Random Forest, LLM Embedding, LLM Classification, and Gradient Boosting. This figure provides a comparative view of the performance of each method over different temporal segments, highlighting trends and potential areas for improvement in predictive accuracy.

Traditional methods, such as Logistic Regression and Gradient Boosting, achieved F1-Scores between 0.60 and 0.65 across the evaluation windows. The LLM Embedding model yielded an F1-Score of approximately 0.67, while the LLM Classification model consistently achieved F1-Scores above 0.70. These metrics were obtained by evaluating the models on a standardized test set, allowing for a direct comparison of their performance.

The results indicate that classification approaches based on language models may offer improved prediction quality relative to traditional methods.

Additionally, Fig. 5 shows that the LLM Classification approach exhibited the highest range and smallest dispersion in F1-Score results compared to the other methodologies. These findings indicate that combining LLMs with direct classification methods is particularly promising for predicting settlements in labor cases, suggesting a valuable tool for supporting the judicial system.

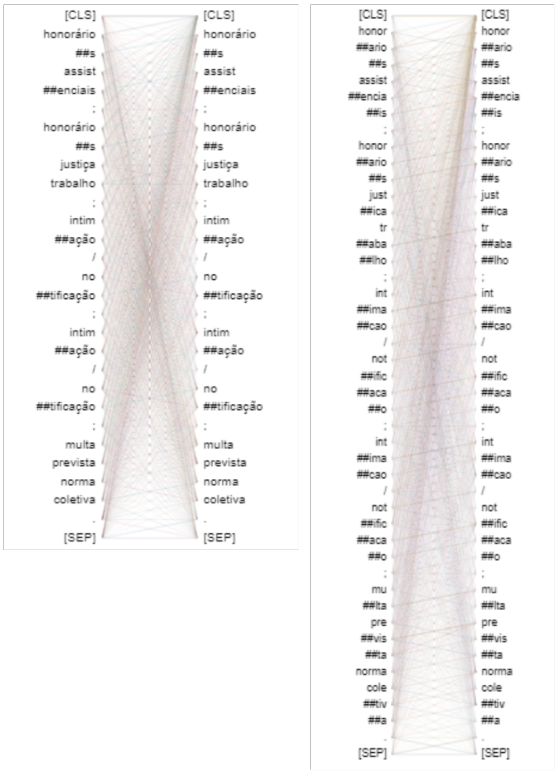

Fig. 6. Comparing attention head from layer 0, using preprocessed text from Fig 2. BERTimbau (left) and BERT base (right), darker colors representing more attention level, and colors representing the 12 different attention heads.

VI. Conclusions and Discussions

Our results indicate that models incorporating LLMs consistently outperform traditional methods, achieving F1-Scores above 0.70. This performance demonstrates the ability of LLMs to capture subtle semantic nuances, which are critical for effective conciliation prediction.

Despite their promising performance, LLMs introduce significant challenges. Their high computational cost and complex implementation requirements can restrict scalability. These challenges underscore the need for careful consideration when deploying such models in practical, real-world applications.

For future work, interpretability and optimization are important points to investigate. Techniques like Integrated Gradients can help elucidate model decisions, thereby increasing trust and transparency. Additionally, hyperparameter optimization and the development of hybrid architectures—combining traditional methods with deep learning—may further enhance the robustness, efficacy, and practical adoption of these models in judicial systems.

References

[1] Anamatra, "Justiça em Números," accessed: Mar. 17, 2025.

[2] Consultor Jurídico, "12% of 2023 Cases Were conciliated," accessed: Mar. 17, 2025.

[3] Anamatra, "Justiça em Números," accessed: Mar. 17, 2025. (Excerpt: "The conciliation tradition in the Labor Court, in the first instance, reaching 36.5%").

[4] N. Aletras, D. Tsarapatsanis, D. Preoţiuc-Pietro, and V. Lampos, "Predicting judicial decisions of the European Court of Human Rights: A natural language processing perspective," *PeerJ Comput. Sci.*, vol. 2, Art. e93, 2016.

[5] D. M. Katz, M. J. Bommarito, and J. Blackman, "Predicting the behavior of the Supreme Court of the United States: A general approach," *PLoS ONE*, vol. 9, no. 4, p. e92426, 2014.

[6] E. Jacob de Menezes-Neto and M. B. M. Clementino, "Using deep learning to predict outcomes of legal appeals better than human experts: A study with data from Brazilian federal courts," *PLoS ONE*, vol. 17, no. 7, p. e0272287, 2022.

[7] J. Silva, M. Oliveira, and P. Costa, "Predicting Brazilian court decisions," *J. Empir. Legal Stud.*, vol. 14, no. 3, pp. 345–378, 2017.

[8] I. Chalkidis, M. Fergadiotis, and I. Androutsopoulos, "Legal-BERT: The Muppets straight out of law school," *arXiv preprint*, arXiv:2010.02559, 2020.

[9] A. Bifet and R. Gavaldà, "Learning from time-changing data with adaptive windowing," in *Proc. 2007 SIAM Int. Conf. Data Mining*, 2007, pp. 443–448.

[10] NLTK Project, "NLTK Documentation," [Online]. Available: https://www.nltk.org. Accessed: Mar. 17, 2025.

[11] Hugging Face, "Hugging Face Transformers Documentation," accessed: Mar. 17, 2025, available at: https://huggingface.co/docs/transformers.

[12] F. Souza, R. Nogueira, and R. Lotufo, "BERTimbau: Pretrained BERT models for Brazilian Portuguese," in *Proc. 9th Brazilian Conf. Intelligent Systems (BRACIS)*, Rio Grande do Sul, Brazil, Oct. 2020 (to appear).

[13] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, 2019.