DOI: 10.53660/CLM-3943-24R37

Comparing CNN-LSTM and BERTimbau: An Analysis of AI Models in Legal Document Classification

Received: 21-07-2024 | Accepted: 25-08-2024 | Published: 31-08-2024

Lucas de França Carneiro Agra
ORCID: https://orcid.org/0009-0003-2963-5931
SENAI CIMATEC, Brasil
E-mail: lucascarneiroagra@hotmail.com
Guilherme Neves Acorsi
ORCID: https://orcid.org/0000-0000-0000-0000
SENAI CIMATEC, Brasil
E-mail: guilherme_zero@hotmail.com
Oberdan Rocha Pinheiro
ORCID: https://orcid.org/ 0000-0002-8904-520X
SENAI CIMATEC, Brasil
E-mail: oberdan.pinheiro@gmail.com

ABSTRACT

This study addresses the urgent need of the Attorney General's Office of the State of Bahia (PGE) to automate the classification of initial petitions, a challenge exacerbated by the lack of standardization in file naming. To tackle this issue, the work proposes the implementation of advanced Deep Learning models, aiming to overcome the limitations of the currently used approach based on regular expressions (Regex), which shows an average accuracy of 80%. The research compares the efficacy of a hybrid model, integrating Convolutional Neural Network (CNN) with Long Short-Term Memory (LSTM), and the BERTimbau model, with the goal of not only enhancing the precision in identifying these essential documents but also promoting procedural efficiency through automation. Preliminary results reveal that the CNN-LSTM model achieved an accuracy of 99.34%, while BERTimbau obtained 98.51%, demonstrating the great potential of both techniques in optimizing the judicial workflow in the digital era.

Keywords: Convolutional Neural Network; Long Short-Term Memory; BERTimbau; Judicial;

INTRODUCTION

The digital era has brought an exponential increase in the availability of data, especially in text formats, as is the case with initial petitions in electronic processes. The efficient classification of these texts represents a significant, but essential, challenge for the optimization of judicial and administrative processes. Deep Learning techniques have shown promise in text classification, given their ability to learn deep and rich data representations, especially in large volumes (MASTELLA, 2021).

This study aims to apply advanced Deep Learning models for the binary classification of initial petitions at the Attorney General's Office of the State of Bahia (PGE). The traditional approach of the PGE, based on regular expressions (Regex), presents an average accuracy of 80%, serving as a baseline for this project. However, this methodology faces challenges of maintenance and updating due to the continuous evolution of legal language, becoming ineffective and costly over time.

The main motivation for the development of Artificial Intelligence models, in this work, lies in the need to overcome the limitations of the Regex model. The challenge is compounded by the absence of a uniform standard in the naming of files, allowing lawyers to assign any name to documents. This variability introduces a significant obstacle to efficient classification, resulting in excessive time consumption by judicial servers in manually identifying such documents when the Regex model fails in its detection. Therefore, the implementation of the models aims not only to surpass the accuracy barriers imposed by the previous method but also to ensure unprecedented speed in the classification process, essential for the optimization of workflow in courts and the promotion of procedural efficiency.

In this work, we initially explore binary classification using a hybrid model of Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM), with Word2Vec embeddings trained on texts from the Attorney General's Office of the State. This model was chosen based on research that highlights the efficacy of CNNs and LSTMs in Natural Language Processing (NLP) tasks. CNNs are excellent at extracting local features in sequential data, while LSTMs capture long-term dependencies, essential for understanding contexts in texts. The integration of Word2Vec embeddings facilitates an enriched semantic representation of words, crucial for NLP. (P, MUKHERJEE, 2023).

Additionally, aiming to enhance the comparison and evaluation of models, we incorporated BERTimbau, a variant of BERT pre-trained specifically for Brazilian Portuguese, which has shown remarkable results in various NLP tasks (MAGALHÃES,

POZO, MACHADO, 2022). The inclusion of BERTimbau allows a direct comparative analysis between the hybrid CNN-LSTM model and a state-of-the-art approach in Deep Learning models, highlighting the potentialities and limitations of each technique in the context of initial petition classification.

The process involves stages of text preprocessing, feature extraction, and classification using the mentioned Deep Learning architectures, with model evaluation based on accuracy. This study not only contributes to the existing literature by presenting effective models for the binary classification of initial petitions but also compares the efficacy of the hybrid CNN-LSTM model with the advanced BERTimbau, illustrating the potential of both approaches for the modernization and automation of legal processes in the digital age.

METODOLOGY

The methodology of this project began with a detailed analysis of the problem and the available data. It was essential for the team to accurately identify the features that define an Initial Petition in contrast to other documents. This preliminary step allowed for the construction of a suitable database for the development of the study.

After analyzing and understanding the problem and the data that will be used in the project, the data collection began through the API of the Court of Justice of Bahia (TJBA) and the SQL Server database system of PGEnet. Data collection through the API was done by downloading complete cases with all their documents, while data acquisition via SQL is done by individual files based on their labels.

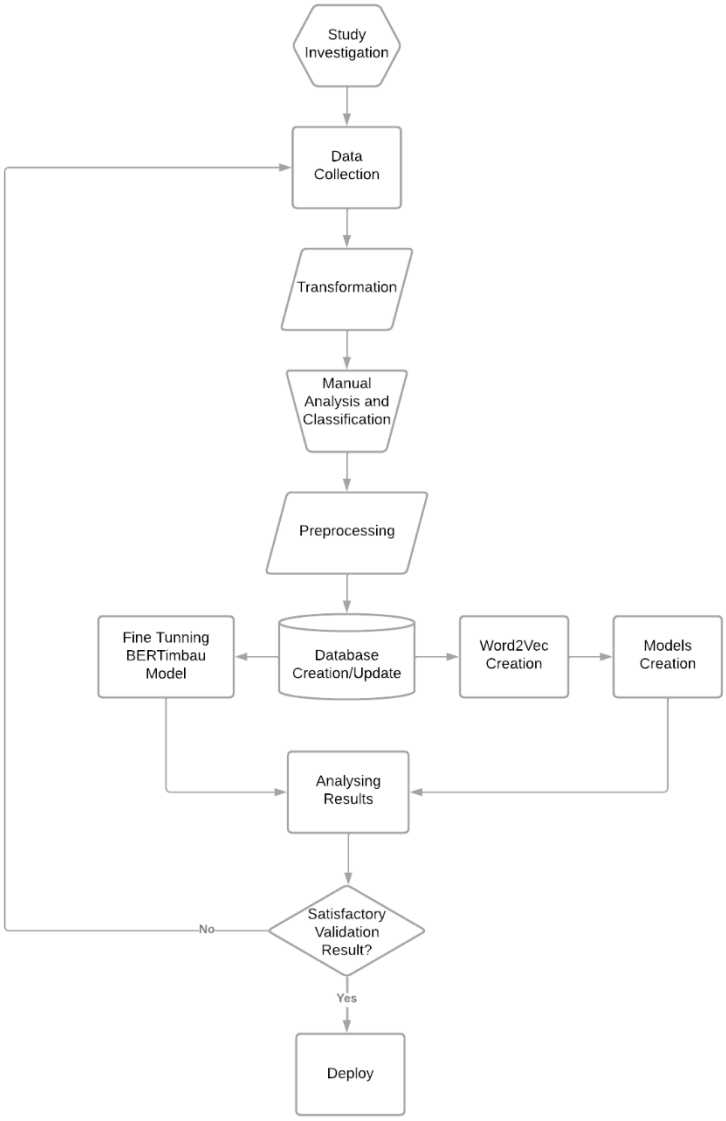
The data collected from SQL Server comes in binary ZIP format, requiring transformation to PDF and then to TXT. Meanwhile, the data from the API come in Base64 format, needing additional processing to convert them into HTML or PDF and then TXT.

After the entire transformation process, extensive processing was necessary for label definition. Due to the absence of reliable tags in the data - for instance, documents named as “Initial Petition” did not always correspond to what their name indicates - it was necessary to manually assemble the database.

Initial petitions represent just one type of document among many in a legal case, making the class of interest ("Initial Petition") naturally imbalanced. This factor required additional effort in data collection, as it was necessary to access a large number of cases to obtain an adequate number of initial petition examples. Thus, the database was

structured into two categories: “Initial Petition” (positive class) and “Other Documents” (negative class).

Figure 1: Process Flowchart



Source: Authors.

After the manual separation of the data, an additional preprocessing step was carried out. This step consisted of removing stopwords, punctuation, and numbers, in addition to replacing personal data such as CPF (Social Security Number), email, and RG (Identity Card) with “cpf”, “email”, and “rg” using regex.

With the database clean and organized, the creation of a custom Word2Vec model for legal texts was created. This was necessary because the available pre-trained models did not adequately capture the specific terms found in the legal documents analyzed. (JANG, KIM, KIM 2019)

Finally, two machine learning models were developed for document classification: one using the BERTimbau architecture, which employs embeddings directly from BERT, and another based on a combination of CNN and LSTM layers, using the embeddings generated by the Word2Vec trained specifically for this research (BENTO, TEIVE, 2023).

The finalized database has a total of 42,019 texts, with 25,610 texts representing the negative class and 16,409 the positive class. For training the models, the data were divided as represented in Table 1 below.

Table 1: Database representation

Data	Other Documents	Initial Petitions	Total
Training	21769	13948	35717
Test	3841	2461	6302
Validation	13794	857	14651

Source: Authors

For the fine-tuning of BERTimbau, the initial 512 tokens were used, the maximum amount of tokens available by the model. Meanwhile, for the training of the CNN-LSTM model, 7000 Word2Vec embedding tokens were used.

After the development of the models, their validation was carried out, which aimed to replicate the production environment to test their effectiveness. For this, we selected the 1,000 most recent cases up to December 19, 2023, forming the basis of our decision-making analysis on the implementation of the models in a productive environment. The success criterion established for this validation phase was to achieve a minimum accuracy of 98.5%, ensuring the reliability and efficiency of the proposed models.

This new dataset consists of 14,651 texts, where 13,794 represent the negative class and 857 the positive class, as can be observed in Table 1 above. As mentioned earlier, the positive class is naturally imbalanced.

To achieve the desired results in the validation phase, it was necessary to iteratively repeat the methodological steps, with a continuous focus on enriching the dataset. This strategy involved the systematic expansion of the database, guided mainly by the error patterns identified in the models, culminating in the final version presented in Table 1. The complexity of this process was accentuated by the nature of the files in the negative class, which often share similarities with the Initial Petitions, requiring a meticulous approach for their distinction.

RESULTS AND DISCUSSIONS

The results obtained in this work were divided into testing and validation, where the test data were obtained from a subset of the training database, equivalent to 15% of it. This low value for the data split was chosen due to the extensive number of data available for validation.

Results of the test set can be observed in Table 2 below. It's important to emphasize that the metrics of recall, precision, and f1-score refer to the positive class.

In the dataset used for testing, both models, CNN-LSTM and BERTimbau, showed exceptionally high performance in classifying initial petitions. This is indicated by the Precision, Recall, and F1-Score metrics, all above 98%, which suggests a remarkable ability of these models to correctly identify initial petitions among other documents.

The CNN-LSTM model slightly outperformed BERTimbau in all the metrics considered. With a precision of 99.27%, recall of 99.39%, and F1-Score of 99.33%, compared to 99.12%, 99.21%, and 98.89%, respectively, for BERTimbau. This may indicate that the hybrid approach of combining CNN with LSTM might be more effective due to its ability to process 7000 tokens, compared to the 512 of the BERTimbau model.

Despite the high values of precision, recall, and F1-Score, it is also important to observe the Loss metric, which provides a measure of how well the model is optimizing its performance during training. The CNN-LSTM model had a Loss of 0.0249, while BERTimbau recorded a slightly higher value of 0.0406. This suggests that the CNN-LSTM is not only predicting the classes more correctly but is also doing so with greater confidence.

These results are particularly relevant for the development of automatic legal document processing systems. The high precision and recall indicate that such models can be reliably used to automate the initial screening of documents, potentially saving significant time for legal professionals and improving the efficiency of legal processes..

Table 2: Test Results

Modelo	Precisão	Recall	F1-Score	Acurácia	Loss
CNN-LSTM	99,27%	99,39%	99,33%	99,47%	0,0249
BERTimbau	99,12%	99,21%	98,89%	99,12%	0,0406

Source: Authors

The results of the validation set can be observed in Table 3 below. And again, the metrics of recall, precision, and f1-score refer to the positive class.

For the validation data, as previously shown in Table 1, it can be seen that there is a large imbalance between the positive and negative classes. As mentioned before, this imbalance represents the real nature of the problem and had a significant impact on the metrics of recall and precision in the obtained results.

When analyzing the performance of the CNN-LSTM and BERTimbau models applied in a production environment, focusing on the accuracy and recall metrics which are crucial for this specific case, we observed remarkable results. Both models maintained high accuracy, with CNN-LSTM achieving 99.34% and BERTimbau 98.51%. This high accuracy suggests that, in general, the models continue to perform well in document classification, even under production conditions.

Furthermore, the extremely high recall, with CNN-LSTM reaching 99.42% and BERTimbau 99.65%, indicates that the models are effective in correctly identifying most of the initial petitions. This is especially important in the production context, where the goal is to capture as many initial petitions as possible.

However, the precision of the models showed a significant drop compared to the results obtained with balanced test data. The CNN-LSTM recorded a precision of 90.35%, and BERTimbau 79.89%, values considerably lower than the 99% observed during the testing phase. This decrease in precision suggests a higher occurrence of false positives, a result of the models' adaptation to an environment that more accurately reflects the variation and complexity of real data.

The F1-Score metric, which balances precision and recall, was also impacted by the reduction in precision. The CNN-LSTM model achieved an F1-Score of 94.67%, and BERTimbau, 88.68%, both lower than the values reached in testing.

Lastly, the slight increase in the Loss metric for both models indicates a higher cost of error during predictions in production.

This restructuring of the analysis highlights the importance of considering the real context of model application, emphasizing success in maintaining accuracy and high recall as key indicators of the practical implementation potential of these systems, despite the challenges presented by data imbalance in production.

Table 3: Validation Results

Modelo	Precisão	Recall	F1-Score	Acurácia	Loss
CNN-LSTM	90,35%	99,42%	94,67%	99,34%	0,064
BERTimbau	79,89%	99,65%	88,68%	98,51%	0,059

Source: Authors

The comparison between the CNN-LSTM and BERTimbau models reveals important nuances in the choice of machine learning solutions for legal document classification tasks. Despite the numerically superior results presented by the CNN-LSTM model, BERTimbau, based on the Transformers architecture, offers distinct advantages that make it an attractive option for long-term implementations.

Transformers are especially effective in capturing contexts and linguistic nuances due to their ability to process all parts of a text simultaneously, unlike the sequential approaches of Recurrent Neural Networks (RNNs) and LSTMs. This makes BERTimbau intrinsically more adaptable to linguistic variations and evolutions in legal language, ensuring its relevance and effectiveness over time.

The observed performance difference between the CNN-LSTM and BERTimbau models can be partially attributed to their respective token processing capabilities: while the CNN-LSTM can process up to 7000 tokens, BERTimbau is limited to 512 tokens. This significant difference in processing capacity may have direct implications on the effectiveness with which each model handles long and complex legal documents. Legal documents often contain extensive legal and argumentative narratives that can easily exceed the 512 token limit. The CNN-LSTM model, with its superior tokenization

capacity, potentially has an advantage in preserving more context and details of the document, which may contribute to its greater accuracy and overall performance.

Another point in favor of BERTimbau is the ease with which it can be updated or expanded to include new data or adapt to legal and linguistic changes. The nature of Transformer-based models allows for a relatively simple update with new training data, which can be crucial for maintaining the system's accuracy over time, especially in fields that are constantly evolving, like law.

FINAL CONSIDERATIONS

In the concluding remarks of this article, we highlight the significant advancement achieved in the classification of initial petitions, an important milestone in the automated processing of legal documents. The work detailed here represents a meticulous journey of research and development, which culminated in the substantial surpassing of the model based on regular expressions (regex) previously utilized. Through an iterative process of refinement and expansion of the dataset, based on careful error analysis and continuous model adaptation, we achieved a level of accuracy that not only met but in many cases, exceeded the expectations set in the validation phase.

Throughout this work, the choice of the BERTimbau model, grounded in the Transformers architecture, represented a turning point in classification effectiveness. This decision was not only based on its immediate performance but also on its promise of long-term sustainability. Transformers, as the underlying technology to BERT, offer an unprecedented ability to understand context and linguistic nuance, essential for adapting to evolutions in legal discourse and variations in documents. This strategic choice ensures that the model not only meets current needs but continues to deliver reliable results in the face of inevitable changes in the legal corpus and the language itself.

This study not only demonstrates the feasibility of surpassing regex-based methods in the classification of legal documents but also establishes a promising path for future investigations and practical applications at the intersection of artificial intelligence and law.

Reflecting on the work done, it is important to acknowledge that, while we have reached significant milestones, the field of automated legal document classification remains fertile ground for innovation. The continuous evolution of NLP techniques and deepening our understanding of the specificities of legal texts open new avenues for research and development. Thus, this article not only reports a significant advancement

in legal document processing technology but also serves as an invitation to the scientific community to further explore the untapped potential in this challenging and impactful area.

ACKNOWLEDGMENTS

We express our gratitude to Professor Oberdan Pinheiro, whose exemplary guidance was crucial for the realization of this work. His generosity in welcoming us into his Artificial Intelligence research team provided us with a unique opportunity for learning and professional growth.

We extend our sincere thanks to the coordinators of the PGE, Felipe Coelho and Carlos Alicio, whose support was essential for the development of this work. Thanks to their technical guidance, patience, and availability, it was possible to achieve the proposed objectives with excellence.

REFERENCES

MASTELLA, J. O. Uma metodologia usando ambientes paralelos para otimização da classificação de textos aplicada a documentos jurídicos. 2020. Dissertação (Mestrado em Ciência da Computação) – Escola Politécnica, Programa de Pós-Graduação em Ciência da Computação, Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre, 2020.

BENTO, F. M.; TEIVE, R. C. G. Classificação de documentos jurídicos utilizando a arquitetura transformer: uma análise comparativa com algoritmos tradicionais de Machine Learning e ChatGPT. Brazilian Journal of Development, Curitiba, v. 9, n. 6, p. 20208-20224, jun. 2023. DOI: 10.34117/bjdv9n6-97.

MAGALHÃES, D.; POZO, A.; MACHADO, S. Técnicas de Aprendizado de Máquinas Aplicadas à Classificação de Decisões Judiciais. Revista de Estudos Empíricos em Direito (Brazilian Journal of Empirical Legal Studies), v. 9, 2022. DOI: 10.19092/reed.v9.573.

JANG, B.; KIM, I.; KIM, J. W. Word2vec convolutional neural networks for classification of news articles and tweets. PLoS ONE, [S.l.], v. 14, n. 8, e0220976, ago. 2019. Disponível em: https://doi.org/10.1371/journal.pone.0220976.

P, SARATHA; MUKHERJEE, SASWAT. A novel approach for improving the accuracy using word embedding on deep neural networks for software requirements classification.

Research Article, Anna University Chennai, 31 mar. 2023. Disponível em: https://doi.org/10.21203/rs.3.rs-2742342/v1.