Comparative Study of Large Language Models Applied to the Classification of Accountability Documents

Pedro Vinnícius Bernhard[1], João Dallyson Sousa de Almeida[2], Anselmo Cardoso de Paiva[2], Geraldo Braz Junior[2], Renan Coelho de Oliveira[3], Luís Jorge Enrique Rivero Cabrejos[2] and Darlan Bruno Pontes Quintanilha[1]

[1]*Programa de Pós-Graduação em Ciência da Computação (PPGCC), Federal University of Maranhão - UFMA, Brazil*

[2]*Federal University of Maranhão - UFMA, Applied Computing Group - NCA/UFMA, Av. dos Portugues, SN, Campus Baganga, Baganga, CEP: 65085-584, São Luís, MA, Brazil*

[3]*Tribunal de Contas do Estado do Maranhão, Av. Carlos Cunha S/Nº Jaracaty, São Luís, 65076-820, MA, Brazil*

{*pedro.bernhard, jdallyson, paiva, geraldo, luisrivero, dquintanilha*}*@nca.ufma.br, rcoliveira@tcema.tc.br*

Keywords: Large Language Models, Natural Language Processing, Document Classification, Accountability, Public Accountability.

Abstract: Public account oversight is crucial, facilitated by electronic accountability systems. Through those systems, audited entities submit electronic documents related to government and management accounts, categorized according to regulatory guidelines. Accurate document classification is vital for adhering to court standards. Advanced technologies, including Large Language Models (LLMs), offer promise in optimizing this process. This study examines the use of LLMs to classify documents pertaining to annual accounts received by regulatory bodies. Three LLM models were examined: mBERT, XLM-RoBERTa and mT5. These LLMs were applied to a dataset of extracted texts specifically compiled for the research, based on documents provided by the Tribunal de Contas do Estado do Maranhão (TCE/MA), and evaluated based on the F1-score. The results strongly suggested that the XLM-RoBERTa model achieved an F1-score of $98.99\% \pm 0.12\%$, while mBERT achieved $98.65\% \pm 0.29\%$ and mT5 showed $98.71\% \pm 0.75\%$. These results highlight the effectiveness of LLMs in classifying accountability documents and contributing to advances in natural language processing. These approaches can potentially be exploited to improve automation and accuracy in document classifications.

1 INTRODUCTION

A Brazilian Court of Accounts oversees the auditing of administrators responsible for public finances at the state and municipal levels (LENZA, 2020). These administrators submit documents via an electronic Annual Accountability System, following Normative Instructions that categorize submissions (TCE/MA, 2023c,b,a). Typically, the entity's holder, technical manager, or an accredited third party manages this process.

Automating document classification can enhance efficiency and optimize human resources in accountability procedures (Stites et al., 2023). Extracting insights from large textual datasets is crucial for organizations, researchers, and professionals (Wan et al., 2019). In this context, Natural Language Processing (NLP) and neural networks have demonstrated significant potential (Khurana et al., 2023). An automated classification model can streamline workflows, improving resource allocation and processing speed (Stites et al., 2023).

Neural networks have excelled in solving complex problems because they can learn patterns and represent non-linear information. Combining this approach with natural language processing makes it possible to extract relevant characteristics from documents and use this information to classify them efficiently and accurately (Khurana et al., 2023).

Recently, there has been a considerable increase in public interest in artificial intelligence models for natural language processing, such as Large Language Models (LLMs) (Naveed et al., 2023). In natural language processing, text classification is an area with few works in Portuguese focusing on real, multi-page documents. The task is often applied to short, well-formatted texts, such as classifying short comments and summaries of academic articles or emails.

Bernhard, P. V., Sousa de Almeida, J. D., Cardoso de Paiva, A., Braz Junior, G., Coelho de Oliveira, R., Cabrejos, L. J. E. R. and Quintanilha, D. B. P.
Comparative Study of Large Language Models Applied to the Classification of Accountability Documents.
DOI: 10.5220/0013439800003929

In *Proceedings of the 27th International Conference on Enterprise Information Systems (ICEIS 2025) - Volume 1*, pages 944-951
ISBN: 978-989-758-749-8; ISSN: 2184-4992

This study seeks to assess and compare different Language Models (LLMs) concerning their effectiveness in categorizing annual accountability documents, particularly those received electronically. The objective is to identify the most suitable LLM for this task while also aiming to improve natural language processing systems for applications in the organization and classification of texts in Portuguese.

1.1 Contributions

Research on using large language models (LLMs) for text classification has primarily focused on English, with limited resources and studies available for Portuguese, especially when dealing with long documents. Portuguese, as a Romance language with rich morphology and syntactic complexity, poses unique challenges for natural language processing (NLP). These challenges, combined with the relative scarcity of annotated datasets and tools for Portuguese compared to English, make this research particularly valuable.

From an academic perspective, this study bridges a significant gap in NLP research by exploring the effective use of LLMs for classifying long-form documents in Portuguese. This contribution broadens the applicability of state-of-the-art NLP techniques to less studied languages, facilitating advancements in linguistic resource utilization and model adaptation. Furthermore, the focus on the Brazilian context strengthens NLP research in a region with distinct linguistic and cultural characteristics, paving the way for applications ranging from sentiment analysis to information retrieval in Portuguese-speaking environments.

From a practical standpoint, this research enhances the efficiency of annual accountability procedures for governmental bodies and managers in Brazil. By automating the classification of documents submitted for accountability purposes, it reduces the workload associated with manual reviews and promotes transparency and reliability in public administration.

The remainder of this paper is organized into five sections. Section 2 reviews related work on document classification. Section 3 details the methodology, covering document acquisition, model selection, and evaluation processes. Section 4 presents and discusses the performance of the models based on key metrics. Finally, Section 5 offers concluding remarks and suggestions for future research.

2 RELATED WORK

Document classification has advanced significantly with the use of transformer-based models. Adhikari et al. (2019a) pioneered BERT for this task, proposing KD-LSTM$_{reg}$, a Knowledge Distillation LSTM model, which achieved an F1-score of 88.9% $\pm$ 0.5 on the Reuters dataset. This outperformed the previous state-of-the-art LSTM$_{reg}$ model by Adhikari et al. (2019b), which scored 87% $\pm$ 0.5%. However, these studies primarily focused on short documents, averaging 175 words.

For longer documents, Wan et al. (2019) proposed dividing documents into segments before classification, achieving an F1-score of up to 98.2% in multi-label classification. In the legal domain, Song et al. (2022) introduced POSTURE50K, a legal multi-label dataset, and a domain-specific pre-trained model with a label attention mechanism, achieving a micro F1-score of 81.2% and a macro F1-score of 27.6%.

Peña et al. (2023) explored multi-label classification of Spanish public documents using RoBERTa, training separate models for each class. The highest sensitivity achieved was 93.07% using an SVM classifier. While these works demonstrate progress, they differ from this study, which focuses on single-label classification of long Portuguese accountability documents using multilingual models. Despite the lack of direct comparability, this work achieves competitive results, surpassing previous metrics in document classification tasks.

3 MATERIALS AND METHOD

This section outlines the methodology for evaluating the performance of LLMs in this study, summarized in Figure 1. Each step is detailed below.

3.1 Document Acquisition

A total of 19,853 documents were collected from the Tribunal de Contas do Estado do Maranhão (TCE-MA) database via the e-PCA system. These PDF documents were converted to text using the pypdfium2[1] library. A sanitation phase removed corrupted files, duplicates, and documents with extraction issues to ensure dataset quality.

[1] https://pypi.org/project/pypdfium2/

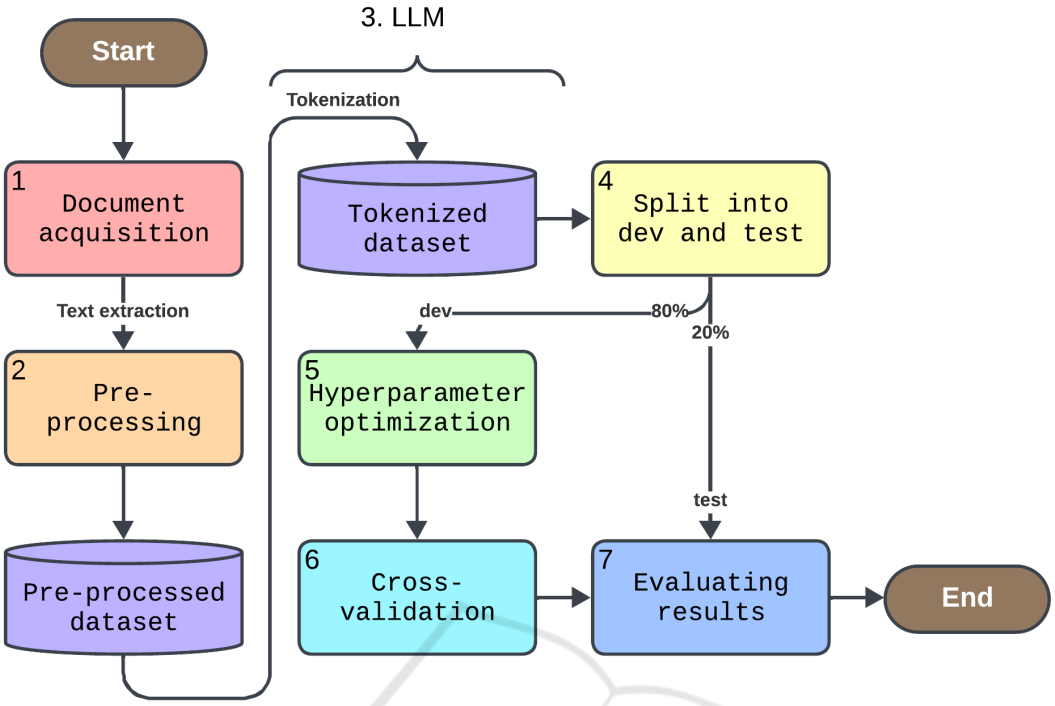

Figure 1: Methodology steps.

3.2 Preprocessing

Text extracted from PDFs underwent preprocessing to reduce noise from the conversion process. Algorithm 1 outlines the steps: removing irrelevant special tokens, repeated characters, and excessive blank or special characters. Parameters were empirically set to balance noise reduction and information preservation.

Input: Text to be preprocessed.
Output: Preprocessed text.
1 Remove special tokens with no relevance to the text;
2 Remove repeated characters;
3 Shorten blank character sequences;
4 Shorten special character sequences;

Algorithm 1: Text preprocessing.

Figure 2 illustrates the preprocessing impact on a budget balance sheet document. Noise, such as repeated dashes, was removed to ensure the model receives relevant information.

The final document types for classification comprised (DCASP) Budget balance sheet, (DCASP) Financial statement, (DCASP) Balance sheet, (DCASP) Statement of changes in equity, (DCASP) Statement of changes in assets, (DCASP) Cash flow statement,

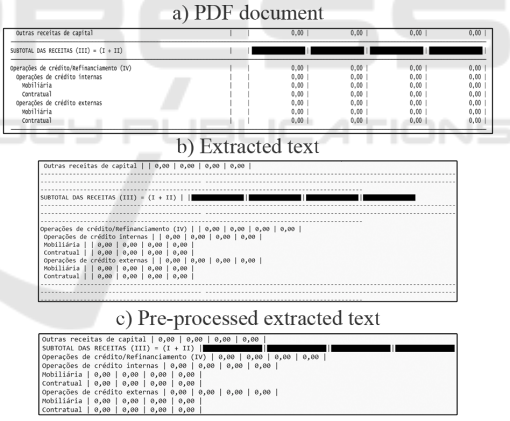

Figure 2: Budget balance sheet document. (a) PDF document. (b) Extracted text. (c) Preprocessed extracted text.

(DCASP) Explanatory notes, Audit report and certificate (including the opinion of the head of the internal control body), Detailed management report, Bank statements and reconciliations, and Letter to TCE/MA.

3.3 LLMs

Three LLMs were selected for their multilingual capabilities and performance on Portuguese tasks:

- **mBERT.** Uses WordPiece tokenization, known for strong contextual embeddings (Devlin et al., 2019).
- **XLM-RoBERTa.** Employs SentencePiece tokenization, excelling in cross-lingual tasks (Conneau et al., 2020).
- **mT5.** Utilizes SentencePiece tokenization, ideal for sequence-to-sequence tasks (Xue et al., 2021).

The three models – mBERT, XLM-RoBERTa, and mT5 – were selected based on their proven performance in multilingual NLP tasks, their effectiveness in handling Portuguese, and their accessibility for this study. Each model represents a distinct architecture (e.g., encoder-only, encoder-decoder) and tokenization approach, offering a diverse set of methods for comparison. While other models, such as GPT-based architectures, were considered, they were excluded due to computational constraints, licensing limitations, or their focus on tasks outside the scope of document classification.

To optimize the training and validation processes, all documents were preprocessed and tokenized before model training. The preprocessed and tokenized documents were saved to disk, reducing computational overhead by eliminating the need for tokenization during training.

Figure 3 illustrates the classification process. The text is first extracted from PDF documents, followed by preprocessing to remove noise. The preprocessed text is then tokenized, and the first 512 tokens are fed into the LLM. The LLM extracts latent information and outputs a vector representation of the document. This output is passed to a classification module, which consists of a Multilayer Perceptron (MLP). The MLP, combined with a softmax function, generates 11 probabilities, each corresponding to one of the final document classes. This step-by-step process forms the core of the document classification methodology in this study.

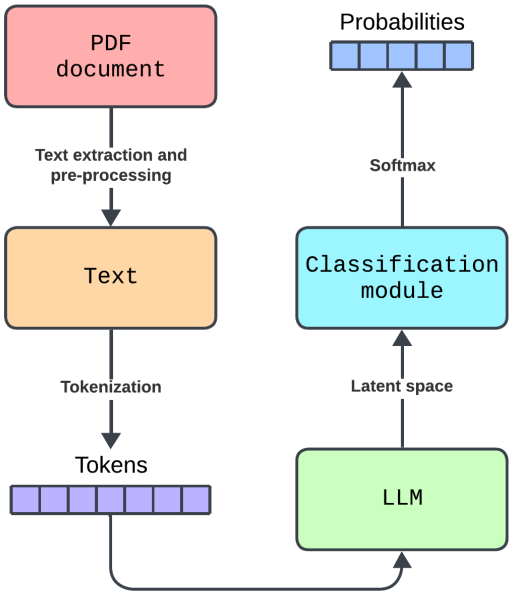

Figure 3: Model execution.

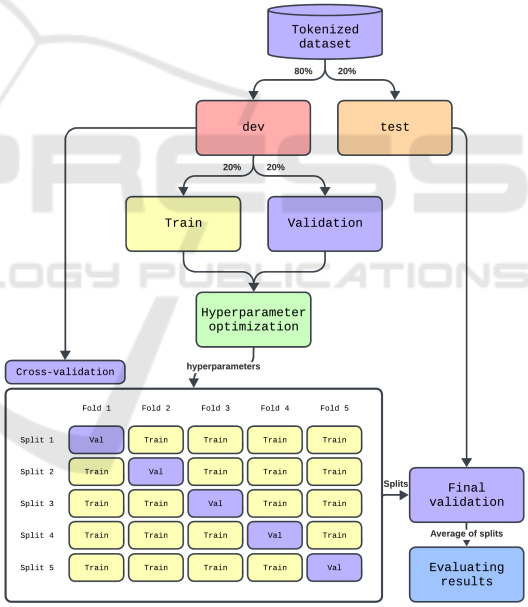

Figure 4: Data division.

3.4 Data Division

Tokenized documents were split into development (80%) and test (20%) sets, maintaining class proportions. The development set was used for hyperparameter optimization and cross-validation, while the test set was reserved for final validation. Figure 4 illustrates the data division process.

3.5 Hyperparameter Optimization

The development set was further split (20% training, 20% validation) for hyperparameter optimization using the Tree-structured Parzen Estimator algorithm and the Optuna library from Akiba et al. (2019). The macro F1-score was optimized over 10 trials of 5 epochs each, tuning:

1. Learning rate (1×10^{-5} to 1×10^{-4}),
2. Weight decay (0.0 to 0.1),
3. Warm-up ratio (0.0 to 0.1).

AdamW optimizer parameters β_1 and β_2 were fixed based on Loshchilov and Hutter (2019) due to poor results during tuning.

3.6 Cross-Validation

Using the optimized hyperparameters and the development dataset (80% of the total data), stratified 5-fold cross-validation was performed. Each fold divided the development set into 80% training and 20% validation. Training ran for up to 15 epochs per fold, with early stopping triggered if the macro F1-score on the validation set did not improve for 5 consecutive epochs. The best-performing model weights from each fold, based on F1-score, were saved for further evaluation.

3.7 Evaluation of Results

Five model configurations from cross-validation were tested on the test set. Metrics included loss, accuracy, macro F1-score, ROC AUC, precision, and recall, chosen for their relevance in classification tasks with imbalanced data according to Opitz (2022). Results are presented as mean ± standard deviation.

4 RESULTS AND DISCUSSION

This section presents the experimental results and analysis of the models evaluated in this study. The dataset, preprocessing, and model performance are discussed, with a focus on key metrics and findings.

4.1 Dataset Analysis

From the original 19,853 documents retrieved from the TCE/MA database, 11,747 documents across 11 categories were retained after preprocessing. Table 1 shows the distribution of document types, with classes such as budget balance sheets, financial statements, and audit reports. The Fisher-Pearson skewness coefficient (0.023) indicates minimal class imbalance. Table 2 provides statistics on sequence counts, sizes, and document pages, while Table 3 lists the most frequent sequences, reflecting accountability-related terms. Figure 5 visualizes these sequences in a word cloud.

4.2 Model Performance

Hyperparameter optimization was performed using the TPE algorithm, with results shown in Figure 6. Table 4 lists the optimal hyperparameters for each model. Five-fold cross-validation was employed, and the best-performing weights from each fold were used for final evaluation.

Table 1: Document classes.

#	Name	Qt.	%
1	(DCASP) Budget balance sheet	1,225	10.43%
2	(DCASP) Financial statement	1,218	10.37%
3	(DCASP) Balance sheet	1,215	10.34%
4	(DCASP) Statement of changes in equity	1,206	10.27%
5	(DCASP) Statement of changes in assets	1,218	10.37%
6	(DCASP) Cash flow statement	963	8.20%
7	(DCASP) Explanatory notes	694	5.91%
8	Audit report and certificate, with opinion of the head of the internal control body	798	6.79%
9	Detailed management report	943	8.03%
10	Bank statements and reconciliations	1,135	9.66%
11	Letter to TCE/MA	1,132	9.64%
-	**Total**	**11,747**	**100%**

Table 2: Document statistics.

Statistic	Quantity of sequences per document	Size of a sequence	Quantity of pages per document
Lowest	14	1	1
Median	398	7	3
Greatest	1,218,778	128	9554
Mode	424	5	1
Average	7321.88	7.76	67.40
Standard deviation	36,412.81	3.38	337.91

Figures 7, 8, 9, 10, 11, and 12 show the training and validation metrics for loss, F1-score, accuracy, ROC AUC, precision, and recall, respectively. XLM-RoBERTa consistently outperformed mBERT and mT5 across most metrics, achieving the highest F1-score (99.21% ± 0.16%) and accuracy (99.22% ± 0.20%) on the validation set. mT5 achieved the highest ROC AUC (99.93% ± 0.04%), but the difference with XLM-RoBERTa was minimal.

Tables 5 and 6 summarize the validation and test set metrics. XLM-RoBERTa achieved the best results across most metrics, with an F1-score of 98.99% ± 0.12% on the test set. The confusion matrix in Figure 13 highlights that class 9 (Bank statements and reconciliations) had the highest error rate (2.64%), but overall performance remained strong.

Table 3: Most frequent sequences (size ≥ 3).

Sequence	Quantity	Sequence	Quantity
saldo	1,912,656	enviada	472,248
conta	1,053,841	aplicação	431,267
valor	768,908	atual	429,039
com	741,795	municipal	416,304
extrato	569,844	mês	366,361
ted	569,396	cota	346,847
banco	563,153	por	326,218
transferência	548,584	transf	318,534
para	492,926	referente	307,035
anterior	472,681	ano	305,332

Figure 5: Word cloud of most frequent sequences.

Table 4: Optimal hyperparameters.

Model	Learning rate	Weight decay	Warm up ratio
mBERT	4.95×10^{-5}	0.063	0.045
XLM-RoBERTa	5.42×10^{-5}	0.097	0.081
mT5	7.46×10^{-5}	0.034	0.028

In conclusion, XLM-RoBERTa demonstrated superior performance, achieving the highest F1-score and accuracy, while mT5 excelled in ROC AUC. All models performed exceptionally well, with mBERT achieving an F1-score of 98.65% ± 0.29% on the test set, underscoring their effectiveness in document classification.

5 CONCLUSIONS

Given the above, this study proposed an in-depth analysis of the effectiveness and performance of advanced language models, known as Large Language Models (LLMs), in the specific task of classifying documents related to the Accountability of managers linked to the Tribunal de Contas do Estado do Maranhão. The proposal sought to evaluate how these models, which are trained on a large scale to understand and generate natural text, can optimize and improve the automation of the process of analyzing and categorizing accounting and financial documents submitted to the court.

After analyzing the data and implementing the strategies outlined in the research project, the study identified that the XLM-RoBERTa model is the most suitable for the document classification task since this model achieved an F1-score of 98.99% ± 0.12% on the test dataset.

When validating and testing the models discussed in this research, it was found that all the models showed favorable results in the metrics explored. This highlights how good LLMs are for classifying documents.

Throughout this research, the first specific objective of the research was achieved. The collection and organization of this information provided a solid basis for subsequent analyses, allowing for a systematic approach to applying LLMs for classifying these documents. The availability of a representative data set proved crucial to achieving the objectives set out in this research, giving validity and solidity to the proposed evaluation processes.

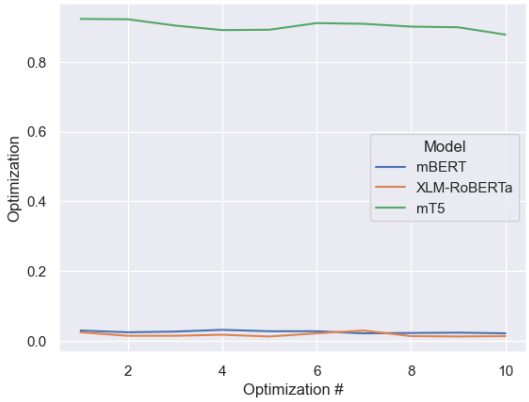

Figure 6: Hyperparameter optimization.

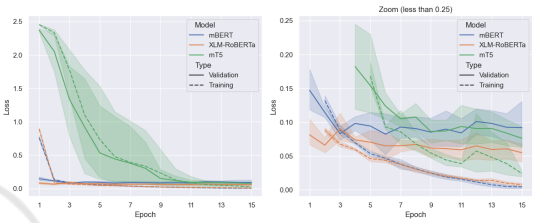

Figure 7: Loss over epochs.

Continuing with the other specific objectives of the research, the outlined goal of using these models in the document categorization process was achieved, demonstrating the ability of these advanced technologies to deal with the complexity inherent in the data contained in the documents analyzed.

Experiments to evaluate the performance of each LLM in classifying documents yielded significant results. The analysis covered several metrics, including accuracy, F1-score, ROC AUC, precision, and sensitivity, providing a broad and accurate assessment of each model's performance.

The analysis of the results obtained, the central target of this research, corroborates the achievement of the established objectives. This crucial stage validated the approaches adopted, contributing to an understanding of the role and potential of LLMs in document analysis and categorization. This analytical process represents not only a conclusive closure to this research but also opens doors for future investigations and the practical application of this knowledge in the wider context of document management and natural

Table 5: Validation set metrics.

Model	Loss	F1-score	Accuracy	ROC AUC	Precision	Recall
mBERT	7.31% ± 2.41%	98.81% ± 0.35%	98.89% ± 0.35%	99.88% ± 0.06%	98.83% ± 0.34%	98.80% ± 0.36%
mT5	8.51% ± 2.46%	98.63% ± 0.72%	98.71% ± 0.61%	99.93% ± 0.04%	98.73% ± 0.56%	98.58% ± 0.82%
XLM-RoBERTa	5.48% ± 1.64%	99.21% ± 0.16%	99.22% ± 0.20%	99.91% ± 0.07%	99.24% ± 0.16%	99.18% ± 0.17%

Table 6: Test set metrics.

Model	Loss	F1-score	Accuracy	ROC AUC	Precision	Recall
mBERT	9.08% ± 1.80%	98.65% ± 0.29%	98.69% ± 0.29%	99.90% ± 0.05%	98.67% ± 0.27%	98.65% ± 0.30%
mT5	8.12% ± 2.13%	98.71% ± 0.75%	98.75% ± 0.67%	99.90% ± 0.01%	98.75% ± 0.67%	98.70% ± 0.80%
XLM-RoBERTa	6.53% ± 0.77%	98.99% ± 0.12%	98.99% ± 0.12%	99.94% ± 0.02%	98.98% ± 0.12%	99.01% ± 0.12%

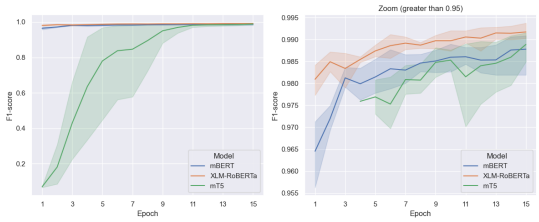

Figure 8: F1-scores over epochs.

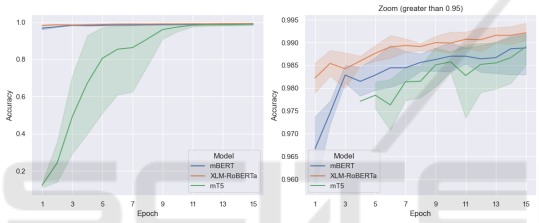

Figure 9: Accuracies over epochs.

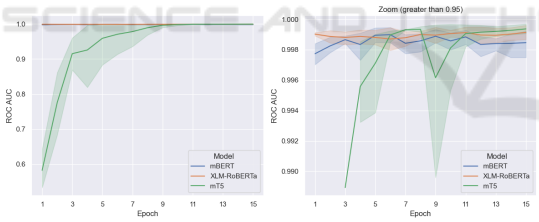

Figure 10: ROC AUC over epochs.

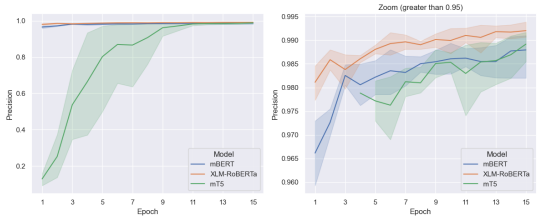

Figure 11: Precisions over epochs.

language processing technology.

The research is crucial to understanding the potential of these models in the area of auditing and inspection, contributing to the efficiency and effectiveness of the procedures for evaluating accountability in public management.

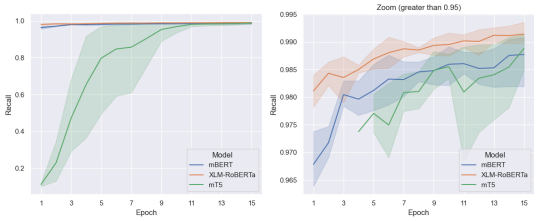

Figure 12: Recalls over epochs.

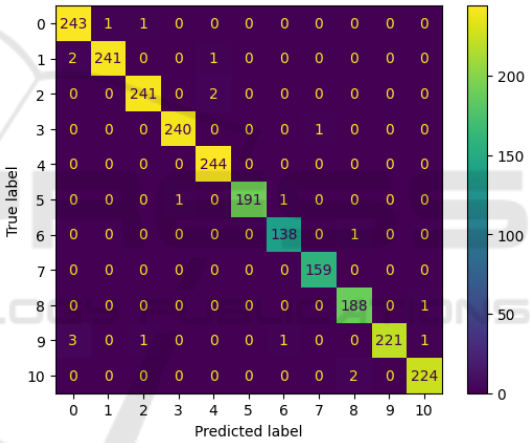

Figure 13: XLM-RoBERTa confusion matrix.

For further research, this study suggests several approaches to improve the understanding and application of LLMs in document classification. Firstly, we recommend evaluating more robust and recent LLMs with billions of parameters to explore the potential of these models on an even broader scale. Furthermore, contemporary techniques such as Document Image Classification (DIC) should be incorporated, which goes beyond textual analysis by classifying document pages as images, thus broadening the analysis perspectives. In addition, the research suggests considering not only the textual content but also the structure of the text, including elements such as location and style, as criteria for classifying documents. This more comprehensive approach seeks to enrich the understanding of the performance and capabilities of LLMs in more diverse and challenging contexts.

ACKNOWLEDGMENTS

The authors acknowledge the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Brazil - Finance Code 001, Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Brazil, and Fundação de Amparo à Pesquisa Desenvolvimento Científico e Tecnológico do Maranhão (FAPEMA) (Brazil) and Tribunal de Contas do Estado do Maranhão (TCE-MA) for the financial support.

During the preparation of this work the authors used ChatGPT in order to enhance the flow of the text and DeepL as a translation assistant to improve fluency. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

REFERENCES

Adhikari, A., Ram, A., Tang, R., and Lin, J. (2019a). Docbert: Bert for document classification.

Adhikari, A., Ram, A., Tang, R., and Lin, J. (2019b). Rethinking complex neural network architectures for document classification. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4046–4051, Minneapolis, Minnesota. Association for Computational Linguistics.

Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Burstein, J., Doran, C., and Solorio, T., editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Khurana, D., Koli, A., Khatter, K., and Singh, S. (2023). Natural language processing: state of the art, current trends and challenges. *Multimedia Tools and Applications*, 82(3):3713–3744.

LENZA, P. (2020). *Direito constitucional esquematizado*. Saraiva, São Paulo, 15. ed. rev. atual. ampl edition.

Loshchilov, I. and Hutter, F. (2019). Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N., and Mian, A. (2023). A comprehensive overview of large language models.

Opitz, J. (2022). From bias and prevalence to macro f1, kappa, and mcc: A structured overview of metrics for multi-class evaluation.

Peña, A., Morales, A., Fierrez, J., Serna, I., Ortega-Garcia, J., Puente, I., Córdova, J., and Córdova, G. (2023). Leveraging large language models for topic classification in the domain of public affairs.

Song, D., Vold, A., Madan, K., and Schilder, F. (2022). Multi-label legal document classification: A deep learning-based approach with label-attention and domain-specific pre-training. *Inf. Syst.*, 106(C).

Stites, M. C., Howell, B. C., and Baxley, P. A. (2023). Assessing the impact of automated document classification decisions on human decision-making. Technical report, Sandia National Lab.(SNL-NM), Albuquerque, NM (United States).

TCE/MA (2023a). e-pca - sistema de prestação de contas anual eletrônica.

TCE/MA (2023b). InstruÇÃo normativa tce/ma nº 52, de 25 de outubro de 2017.

TCE/MA (2023c). Sistema de prestação de contas anual eletrônica (epca) já está disponível aos usuários.

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

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. In Toutanova, K., Rumshisky, A., Zettlemoyer, L., Hakkani-Tur, D., Beltagy, I., Bethard, S., Cotterell, R., Chakraborty, T., and Zhou, Y., editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.