Bidirectional Encoder Representations from Transformers in Brazilian News Classification of a Public Security Agency

Thiago R. Lobo[*], Karla S. Silva[*], Claudia A. Martins[*], Raphael S. R. Gomes[*] and Anderson C. S. Oliveira[†]
[*]Institute of Computing
[†]Department of Statistics
Federal University of Mato Grosso, Mato Grosso, Brazil 78060-900

***Abstract*—As part of the institution's 2024–2027 strategic plan, which includes the objective of understanding how the media portrays the organization to strengthen its public image, this work aims to investigate the application of a combination of Bidirectional Encoder Representations from Transformers (BERT) models, pre-trained in Portuguese, to classify and evaluate news headlines about a public security institution. This classification took place in three positions: neutral, negative, and positive. This study seeks to collaborate with the organization's strategic objectives by presenting the model with the highest metrics during the analysis. In this way, BERT-Large stands out, with an F1 Score of 91.62% in the model's general metrics and an equally high result compared to the others during the classification of headlines into classes.**

***Index Terms*—BERT, Sentiment Analysis, News Classification, Transformers**

I. Introduction

Perícia Oficial e Identificação Técnica (POLITEC) is a public agency responsible for producing expert reports and issuing identity cards for municipalities in the state of Mato Grosso, Brazil. During 2023, the Strategic Actions Center (SAC) prepared the institution's strategic plan for 2024-2027, containing 16 Strategic Objectives (SOs) with targets and indicators.

One of the SO (SO 02) aims to strengthen the agency's institutional image, seeking to understand and classify news published by the main state media outlets as part of the strategy to measure society's perception of the institution, identifying what has generated positive and negative repercussions in the media.

In order to help solve this problem, computational techniques for identifying textual data and classifying news stories are used to assist in this process. To this end, The area of NLP is explored so that text information can be extracted and analyzed in order to observe patterns present in the content.

These analyzed patterns allow entities to be found, such as places of operation and organizations involved, supporting the categorization of documents according to the content, or emotion expressed by the text.

In this way, text processing in the news classification process aims to evaluate the comments made about the institution, analyzing sentiments that can be classified as positive (they improve the organization's image), neutral (they only mention the institution's performance), or negative (they damage the organizational image).

This sentiment analysis can be carried out on different media, such as news, comments, descriptions, among others, enabling organizations to make assertive decisions about their processes, products, or services, helping to advance established strategies [1].

In an initial study, [2] explored the use of Machine Learning (ML) techniques for classifying news about Politec. Although the models achieved relatively high accuracy scores, they struggled with negative and positive headlines, reaching those values mainly due to the predominance of neutral samples in the dataset. This limited effectiveness highlighted the need for techniques capable of better capturing the contextual nuances of headlines.

Proposed in 2017, the Transformer architecture [3] represented a paradigm shift in the NLP field, replacing traditional mechanisms in Recurrent Neural Networks (RNN) with an approach based on Self Attention - capable of parallelizing the input and processing long texts more efficiently.

From this architecture onward, models such as the Bidirectional Encoder Representations from Transformers (BERT) [4] emerged, which achieved prominence in several classification, extraction, and semantic analysis tasks, being widely adopted in real applications, including sentiment analysis in news.

In this context, this work aims to apply, analyze, and compare pre-trained BERT models in the analysis of news headlines about POLITEC in the main media outlets of the state. Furthermore, to evaluate the generalization capacity of the developed models, StratifiedKFold Cross-Validation was used.

This paper is organized as follows: The next section identifies the Related Works. Section 3 describes the Materials and Methods used to process the classification of news headlines. Section 4 presents and analyzes the results, including the performance of the models, and the last section concludes the study.

II. Related Work

As the main step, an overview review was carried out, an approach that focuses on ensuring an overview of the existing literature, without following a systematic search, selection, or critical analysis protocol [5]. To this end, searches were carried out using Google Scholar search engine to identify works that addressed the use of sentiment analysis applied to the classification of news and news headlines, specifically in the Brazilian domain.

RNNs and their variations, such as Long-Short Term Memory (LSTM), Bi-LSTM, and Gated Recurrent Unit (GRU), have been widely used in sentiment analysis. [6] classified news comments in Bengali using both traditional ML algorithms (Naive Bayes, KNN, Random Forest, Decision Tree, SVM, Logistic Regression) and RNNs, achieving 95% accuracy with RNNs. Similarly, [7] focused on financial news, employing LSTM networks to capture long-term dependencies, reaching 91.5% accuracy and 92% precision.

However, with the advent of BERT models [4], derived from Transformer architectures [3], sentiment analysis experienced significant advances, especially in capturing subtle textual nuances.

For example, [8] applied sentiment analysis through the BERT model to classify harmful news of a fake news detection datasets from Kaggle. The proposed model makes it possible to train the database and identify the linguistic procedures and sentiments of dangerous media, which can proliferate misleading news and generate social challenges. The proposed BERT model achieved 97.36% of F1 Score in harmful news detection, outperforming all XLM-RoBERTa and BART large metrics.

[9] also carried out a classification of headlines in URDU using the BERT model (fine-tuned), but combined the features of this architecture with Logistic Regression (LR) and a Multilayer Perceptron (MLP), showing an average accuracy of 95%.

[10] classified news in real time using the Natural Language Processing (NLP) transformer Bi-directional Encoding Representational Transformers (BERT), obtaining an accuracy of 91%. [11] compared Deep Learning (DL) models - biLSTM and biGru - with the BERT model for news classification, using comparable constraints between the two techniques, in which the BERT model obtained 98.1% accuracy, outperforming the others.

[12] proposed a deep learning system for sentiment analysis of financial news, providing an interface for accessing and analyzing multiple sources. The study compared RNN, CNN, and Transformer architectures, showing that BERT achieved the best performance with 90% accuracy, surpassing RNN (80%) and CNN (83%).

In their work, [13] carried out sentiment analysis on news headlines using four types of transformers: BERT, RoBERTa, DistilBERT, and XLNet. They were used to classify the news as positive, neutral, and negative, as proposed in this study. During the tests, the model that performed best was BERT, achieving an F1 Score of 93.6%.

Soomro et al. [14] investigated sentiment analysis of Sindhi news headlines, comparing approaches based on Machine Learning, Deep Learning, and Transformer models, highlighting the advantages of transformer-based architectures for capturing contextual nuances. Similarly, Khatun and Khan [15] addressed the identification of counterfeit news in Bangla using BERT, showing the relevance of transformer models for journalistic content analysis. These studies reinforce the adaptability of transformer-based models to different languages and journalistic domains.

Furthermore, scientific research conducted in Brazil has investigated the use of Sentiment Analysis in various domains. The study by [16] applies the concepts of sentiment analysis to predict stock prices based on financial market news. In turn, [17]–[19] uses text corpora from social networks, such as X/Twitter and Reddit, to understand how certain topics are discussed on these platforms. Finally, [20] explores the use of language models, such as ChatGPT, as a tool for Sentiment Analysis tasks.

Despite recent advances, no studies were identified applying BERT-based approaches to sentiment analysis of news in government domains, particularly related to public security institutions, showing that this area remains underexplored in Brazil.

Although recurrent architectures such as RNNs, LSTMs, and BiLSTMs have achieved competitive results in NLP fields, their reliance on fixed embeddings and limited ability to capture long-range dependencies reduce their effectiveness in unbalanced and domain-specific datasets. In contrast, BERT-based models use contextual embeddings pre-trained on large corpora, allowing a more refined understanding of semantic nuances.

This feature is especially relevant in sentiment analysis of news headlines about public security, where words with negative connotations do not always indicate negative institutional contexts. Thus, this study explores BERT-based architectures on a manually built, domain-specific dataset, advancing sentiment analysis in Portuguese for public security institutions.

III. Materials and Methods

In order to identify the best model for analyzing sentiment in headlines, this study proposed a methodology based on the use of different pre-trained BERT models, using StratifiedKFold Cross-Validation to statistically evaluate the results obtained, as illustrated in Fig. 1.

The methodological process began with data collection via Web Scraping, followed by textual pre-processing and statistical analysis of the data to understand the main characteristics of each class. In the next stage, pre-trained BERT models were used to obtain contextual representations of the texts, using its Transformer-based architecture, along with fine-tuning of the models, adapted to the classification task.

Each model was trained using StratifiedKFold Cross-Validation, ensuring a robust analysis of the results. Finally,

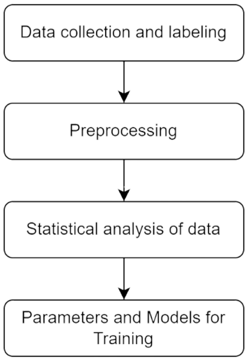
Data collection and labeling
Preprocessing
Statistical analysis of data
Parameters and Models for Training

Fig. 1. Flowchart of the Methodology Carried Out

the models were evaluated with the test data and their metrics extracted for analysis.

A. Data Collection and Labeling

Initially, a set of six state news media outlets was selected. The task of extracting news headlines was carried out using the Web Scraping process, using Python scripts and the Google Colab virtual environment.

For this study, the database was composed mainly of data from the years 2023 and 2024, which account for 7,109 of the 8,638 total samples. Even after the full labeling of the 2023 and 2024 samples, the positive and negative classes showed low representation.

Therefore, to increase the amount of data for these classes, additional headlines from previous years were mapped and labeled based on keywords associated with these contexts, such as 'report', 'expertise', 'delay' and 'POLITEC', resulting in 1,529 new samples, with a predominance of headlines defined as positive or negative. In the Table I is illustrated the temporal distribution of the samples by class.

As a result, the distribution of the database shows a concentration in the most recent years, especially 2023 and 2024, and a lower volume in previous years, with a gradual reduction as time goes on.

The labeling was carried out by a single person, based on the criteria previously defined by the institution's strategic team:

- positive headlines refer to content that highlights achievements, effective actions or institutional recognition;
- neutral headlines are those that only mention the institution in everyday or informative contexts, without value judgment;
- negative headlines present criticism, institutional failures or the involvement of civil servants in crimes and scandals.

This approach reduced costs and ensured consistency in the application of criteria, but it also carries the risk of subjective bias, since the interpretation of headlines depends on individual judgment.

TABLE I
PROCESSED DATABASE

Year	Positive	Neutral	Negative	Total
2005	1	0	0	1
2006	4	1	3	8
2007	10	12	3	25
2008	26	18	6	50
2009	9	15	4	13
2010	16	0	6	22
2011	16	0	5	21
2012	21	3	12	36
2013	23	2	15	40
2014	69	0	26	95
2015	95	3	32	130
2016	107	4	50	161
2017	111	4	63	178
2018	115	3	33	151
2019	90	0	33	123
2020	96	0	52	148
2021	140	0	35	175
2022	120	0	17	137
2023	273	3740	38	4051
2024	181	2853	24	3058
Total	1523	6658	457	8638

B. Preprocessing

In this study, linguistic normalization was used to refine the data before modeling with BERT. Initially, each word in a headline was tokenized. Then, uppercase letters were converted to lowercase and special characters were removed, keeping only accented letters. Next, stopwords were removed, using a set of stopwords corresponding to prepositions and articles in Portuguese. In order to bring uniformity and avoiding divergences between words, the lematization technique was applyed using the SpaCy library with the Portuguese model.

The lemmatization process converts the text into smaller units (tokens) and assigns them a corresponding lemma, which is important for dealing with common variations in the Portuguese language. All these procedures return lemmatized sentences, bringing uniformity and avoiding divergences between words, allowing BERT to be more efficient in its operations and capture relevant semantic patterns. However, despite these benefits, lemmatization increases processing time and may lead to the loss of subtle semantic nuances, especially considering the morphological complexity of Portuguese.

C. Statistical analysis of data

The statistical analysis of the textual data explored the distribution and structural characteristics of the news items in the database. In Fig. 2 is showed the percentage distribution graph by class, with neutral headlines predominating (77.07%), followed by positive headlines (17.63%) and negative headlines (5.29%). This asymmetry has implications for model training, highlighting possible balancing strategies or weight adjustments to avoid classification biases.

Next, we analyzed the length of the headlines to define the input length of the sequences in the BERT models. In Fig. 3 is presented a box plot that indicates an average length of approximately 8.58 words per headline. The first quartile is

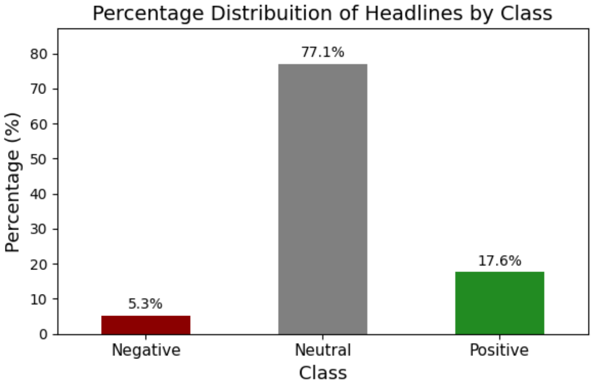

Fig. 2. Headline Distribution Chart by Class

around seven words, while the third quartile is close to ten words. The neutral class concentrates the largest number of outliers, with the largest having around 20 words.

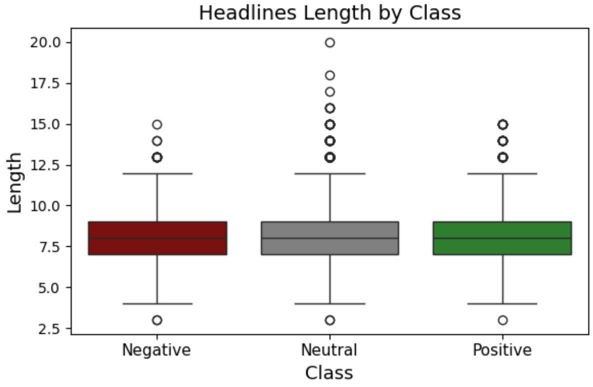

Fig. 3. Boxplot of Headline Length by Class

In order to deepen the analysis of the classes, the ten most frequent words per class were defined, and this stage was an input for the modeling decisions, data balancing, and interpretation of the classification results.

When analyzing the most frequent words in the positive class, the name of the institution predominates, with 566 citations. In addition, terms such as “report” (242), ‘expertise’ (173) and “body” (118) feature prominently, suggesting a focus on possible developments in expertise. Words like “new” (151) and “point out” (112) are also among the most recurrent, indicating possible associations with announcements of new services, positive actions or dissemination of results.

In the negative class, the name of the institution continues to be one of the most frequent words, with 114 occurrences. In addition, the terms “report” (86), “no” (53), ‘wait’ (47) and “body” (46) stand out, which may suggest dissatisfaction related to waiting for results or the absence of conclusions.

In the neutral class, the most frequent words are mostly associated with news events. Terms such as “to be” (2864), “to die” (2023), “man” (1486), “to kill” (1386), and “to shoot” (874) indicate a predominance of factual reports about violence or police incidents. However, words such as “to die” and “to kill”, although recurrent in neutral headlines, tend to carry a negative charge in other contexts. This can make automatic classification difficult, as neutral headlines can be mistakenly classified as negative.

D. Model Training

The training stage began with the delimitation of the BERT models, followed by the respective hyperparameters used. The training was carried out with StratifiedKFold Cross-Validation and, in the end, its metrics were extracted and applied to the data intended for testing.

1) BERT models: For this study, four variants were selected: BERT-Base and BERT-Large, as robust baselines; XLM-RoBERTa, to compare multilingual training against Portuguese-specific adaptation; and DistilBERT, as a lightweight alternative balancing accuracy and efficiency. Compared to autoregressive or encoder-decoder models, these encoder-based architectures are more directly aligned with sequence-level classification, making them well-suited for sentiment and polarity analysis in the Brazilian public security domain.

The BERTimbau model [21], a BERT model pre-trained for the Portuguese language, was used. The BERTimbau Base model, which has 12 layers and 110 million parameters, allows us to observe the textual similarity of the words recognized in the previous stage and to classify the news items that are in the database used. To compare the results between the models, BERTimbau Large Model was also used, considering the performance that can be obtained by twice as many layers (24) and 335 million parameters.

In addition to the models specialized in the Portuguese language, pre-trained multilingual models were used, in different languages, but with better results in English and Portuguese. Also from the BERT family of models, DistilBERT-base-multilingual-cased [22] is based on Wikipedia, in 104 languages, with 6 layers, 768 dimensions, 12 layers, and 134 million parameters - a value close to the others used.

Compared to the DistilBERT model, the XLM-RoBERTa-based model [23] uses a base of 100 text languages, has 279 million parameters, without human labeling, allowing it to be a generalizable model for NLP tasks. It has an automatic label generation process, which performs a masked language model to predict hidden words in a text, without the need for already defined labels.

2) Hyperparameter Selection: The training procedure began by defining the main hyperparameters of the BERT models. All models were trained with six epochs, batch_size of 16, learning rate of 2e-5, and weight_decay of 0.01.

The batch_size and learning rate follow Devlin et al. [4], who showed that batch sizes between 16 and 32 and learning rates between 2e-5 and 5e-5 ensure stable convergence. Although [4] typically recommends up to 4 epochs, we set 6 epochs based on empirical observations with smaller datasets [24]. Weight decay of 0.01 was applied for regularization, preventing overfitting and stabilizing optimization as suggested by [25].

Finally, to ensure robustness and validity of the training, Stratified K-Fold cross-validation with k = 10 was applied, preserving class distributions and providing more reliable estimates of model generalization.

3) Model Training with StratifiedKFold Cross-Validation: After defining the hyperparameters, the dataset was initially partitioned into 90% for training and 10% for final testing. Then, the models were trained using StratifiedKFold Cross-Validation. For each of the ten iterations of Cross-Validation, the models were trained, and their respective F1 Score values were recorded.

In each of the models, the textual inputs were previously processed using the corresponding tokenizer, limiting the length of the inputs to up to 32 words. Each sentence was encoded in vector representations with 768 dimensions, and the output corresponding to the special token [CLS] was used as the sequence representation vector for the classification task.

At the end of the Cross-Validation, each model was trained again using all of the training data and subsequently evaluated on the previously separated test set.

In order to speed up training, the process was carried out using TPU v6e-1 available in the Google Colab virtual environment. This infrastructure significantly accelerated experimentation and eliminated the need for dedicated hardware. However, reliance on free cloud services brings risks, such as limited runtime sessions and reproducibility challenges. In a production or institutional setting, dedicated hardware or cloud credits would be necessary to ensure stability and scalability. The code used in this stage can be accessed at this Google Colab link.

IV. Results

Firstly, this section present the mains results and a statistical analysis of the F1 Scores obtained in each of the folds, followed by an individual analysis of the models, identifying the main errors for each of the classes present. In Table II are showed an overview of the metrics Accuracy (Acc), Precision (P), Recall (R), and F1 Score (F1) obtained by the models. Given the imbalance between the classes, the main metric to be analyzed will be the F1 Score, which represents a harmonic mean between Precision and Recall – Equation (1).

$$F_1 Score = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{1}$$

TABLE II
Main Metrics Models

Model	**Acc (%)**	**P (%)**	**R (%)**	**F1 (%)**
BERTimbau Base	96,18	88,55	93,11	90,51
BERTimbau Large	96,75	91,32	92,01	91,62
DistilBERT	96,29	89,72	89,11	89,41
XLM-RoBERTa	95,13	86,24	87,76	86,98

In Fig. 4 and Fig. 5 are showed the evolution of validation accuracy and loss per epoch. In general, in their respective first epochs, the models achieved accuracies of over 90%, thus highlighting how the BERT models' self-attention mechanism is highly effective in capturing characteristics and contexts present in texts. Furthermore, the BERTimbau Base model showed better stability and lower loss in its final epoch, while DistilBERT showed constant growth in loss.

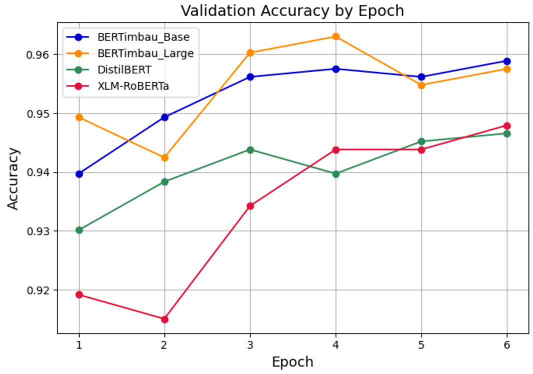

Fig. 4. Evolution of Accuracy per Epoch

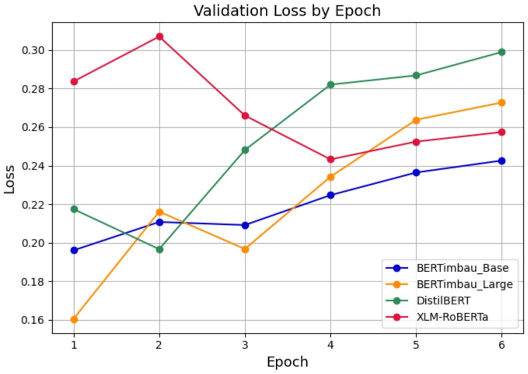

Fig. 5. Evolution of Loss per Epoch

In general, the models obtained satisfactory results, especially the BERTimbau Base and Large models. The following subsections present a statistical analysis of the F1 Scores obtained in each of the folds during the StratifiedKFold Cross-Validation, followed by an individual analysis of the models, identifying the main errors for each of the classes present.

A. Statistical Analysis of Model Stability via Stratified KFold

With the initial aim of evaluating the performance of each model's folds, we casually use the average, standard deviation and Confidence Interval (CI) calculation to measure the uncertainty associated with this average. The CI is a measure that helps ensure the degree of reliability that an observed average actually reflects the true performance of the models used – Equation (2).

$$CI = \bar{x} \pm t_{1-\frac{\alpha}{2},n-1} \cdot \frac{s}{\sqrt{n}} \tag{2}$$

The performance obtained is this work is illustrated in Fig. 6. Models with narrower intervals (BERTimbau and

DistilBERT) have greater stability and statistical reliability. The narrower intervals of the models indicate greater similarity between the F1 Score values of the folds of each model during Stratified KFold.

However, models with wider CIs, such as XLM-RoBERTa, reflect greater uncertainty about their average performance, suggesting that their effectiveness may be more sensitive to the composition of the training data, given the variation in F1 Score values according to the set of folds used for training and validation.

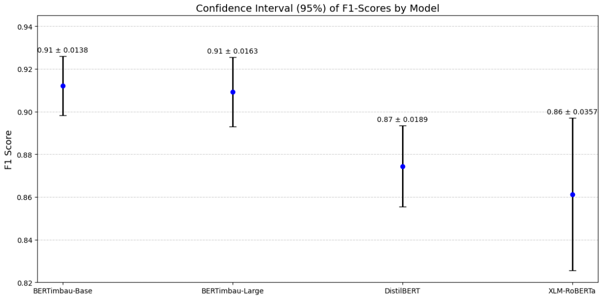

Fig. 6. Confidence Interval

After statistical analysis of the F1 Score obtained by Cross-Validation, the Friedman test was applied to verify significant differences among the models. The test returned a p-value of 0.0009, very close to zero, confirming that the models' performances are not equivalent and that there are relevant statistical differences.

Subsequently, Nemenyi's post hoc test indicated that BERTimbau Base outperformed DistilBERT (p = 0.0286) and XLM-RoBERTa (p = 0.0030), while BERTimbau Large showed superiority only over XLM-RoBERTa (p = 0.0286). These results align with the confidence interval analysis, reinforcing that the BERTimbau models present greater stability, robustness, and reliability for sentiment classification in Portuguese news headlines.

B. Model Performance Analysis by Sentiment Class

The BERTimbau Base and Large models performed better than the others in all classes, especially the negative class, where the F1 Score of the models was 82% and 84% respectively, as shown in Fig. 7. This superiority over the DistilBERT and XLM-RoBERTa models is due to the fact that both are multilingual models, trained on a wide variety of linguistic text.

The use of BERT models trained exclusively on a large body of Portuguese-language data produced good results for all classes of headlines, especially for negative headlines, where the high imbalance of the classes and chance impacts on erroneous training for minority classes.

C. Model Error Analysis and Limitations

Error analysis is essential to identify the limitations and inconsistencies of machine learning models. Confusion matrices (see Appendix A) reveal that models trained on Brazilian corpora performed consistently across all classes, with a clear improvement in the negative class compared to multilingual models.

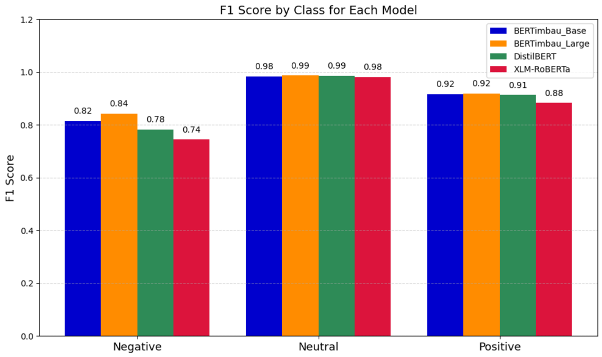

Fig. 7. F1 Score by model class

TABLE III
MISCLASSIFIED HEADLINES

Headline	**Class**	M1	M2	M3	M4
Firefighters try to avoid responsibility for soldier's death; Court summons Politec (H1)	Neg	Pos	Neu	Neg	Neu
Elderly man's family requests judicial measures to speed up investigation (H2)	Neg	Pos	Neu	Neg	Neu
Politec of Água Boa faces difficulties in carrying out its work (H3)	Neg	Pos	Neu	Neg	Neu
Forensic experts of Mato Grosso suspend strike call after government confirms salary payments (H4)	Pos	Neg	Neg	Neg	Neg
After negative tests, grandfather is released following rape accusation in MT (H5)	Pos	Neu	Neu	X	Neu
Politec reveals error in building project (H6)	Pos	Neg	Neg	X	Neg
Civil Police incinerates around 20 kilos of narcotics in the countryside (H7)	Neu	Pos	Pos	Pos	Pos
PJC incinerates 150 kilos of seized cocaine (H8)	Neu	Pos	Pos	Pos	Pos
Applications for the recruitment test of prison agents start tomorrow (H9)	Neu	X	X	X	Pos

M1 = BERTimbau-Base, M2 = BERTimbau-Large, M3 = DistilBERT, M4 = XLM-RoBERTa

Nonetheless, quantitative evaluation alone cannot fully explain the linguistic and contextual factors behind misclassifications. For this reason, combining metrics with qualitative analysis of misclassified headlines provides a broader view of model performance, helping to identify patterns of errors, linguistic biases, and task-specific challenges. Table III presents nine headline samples, three per class, along with the corresponding outputs from each model.

1) Samples of negative headlines mistakes: In headlines H1 and H2, the summons between institutions indicates potential delays, a negative scenario, yet the models misclassified them without clear reason. In H3, misclassification occurred only in multilingual models, likely influenced by the positive adjective "Boa," as preprocessing failed to recognize it as the municipality name Água Boa, ignoring context.

2) Samples of positive headlines mistakes: The presence of words such as "suspend", "negative", and "error" in headlines H4, H5, and H6 seems to have influenced the classification by the models. Despite the overall positive context of these headlines, the models interpreted these words in isolation as indicators of negativity or neutrality, resulting in incorrect classifications.

3) Samples of neutral headlines mistakes: Although headlines H7, H8, and H9 describe positive governmental actions, they do not directly involve Politec nor present explicit positive or negative judgments about the institution. For this reason, these headlines should be considered neutral. However, it is observed that the models incorrectly classified M7 and M8 as positive, possibly due to the isolated interpretation of words such as "incinerates" or "cocaine", which may have been associated with a positive impact in terms of police action. Headline M9, in turn, presents greater classification difficulty, reflecting the complexity of recognizing neutrality in informative news.

V. Conclusion

As news spreads through the media, public perceptions of an organization can be shaped and propagated in different ways. Developing efficient sentiment analysis models that capture textual nuances allows for the rapid identification of headline sentiment regarding POLITEC, allowing the institution to prioritize responses, conduct communication campaigns, adjust strategies, and mitigate impacts on public perception.

While traditional machine learning models and architecture may suffice for simple classification tasks, the use of BERT proves preferable in scenarios with complex or ambiguous headlines. Its ability to capture semantic subtleties makes it superior to conventional models in accurately identifying sentiment in challenging textual contexts.

In this context, to support POLITEC's strategic objectives, this work analyzed several pre-trained BERT models in the language of the news, classifying headlines into Negative, Neutral, and Positive categories.

The dataset analysis showed a predominance of neutral headlines, followed by positive and negative ones, with the latter requiring greater attention due to their potential institutional impact. For classification, the monolingual BERT Base and Large models achieved the best results, with F1 Scores of 90.51% and 91.62%, surpassing the multilingual models DistilBERT (89.41%) and XLM-RoBERTa (86.98%). Among them, BERT-Large consistently stood out across all classes.

As limitations, this study may present temporal bias, as it does not cover news published prior to the analyzed period. For future work, we propose exploring Few-Shot and Zero-Shot approaches, especially with the T5 architecture, as well as developing an application to provide an intuitive interface for news classification.

References

[1] O. Prasad, S. Nandi, V. Dogra and D. Diwakar, "A systematic review of NLP methods for Sentiment classification of Online News Articles," in 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT), Delhi, India, 2023, pp. 1-9, doi: 10.1109/ICCCNT56998.2023.10308056.

[2] T. Lobo and C. Martins, "Comparativo de Algoritmos de Aprendizado de Máquina para a Classificação de Notícias sobre a Politec em Mato Grosso", in Anais da XIII Escola Regional de Informática de Mato Grosso, Alto Araguaia/MT, 2024, pp. 72-77, doi: https://doi.org/10.5753/eri-mt.2024.245831.

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser and I. Polosukhin, "Attention Is All You Need", in Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17), Red Hook, NY, USA, 2017, pp 6000–6010.

[4] J. Devlin, M. W. Chang, K. Lee and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding", in Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, 2019, vol 1, pp. 4171-4186, .

[5] M. Grant and A. Booth, "A typology of reviews: an analysis of 14 review types and associated methodologies", in. Health Information & Libraries Journal, v. 26, n. 2, p. 91–108, 2009.

[6] F. Akter, S. A. Tushar, S. A. Shawan, M. Keya, S. A. Khushbu and S. Isalm, "Sentiment Forecasting Method on Approach of Supervised Learning by News Comments," 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT), Kharagpur, India, 2021, pp. 1-7, doi: 10.1109/ICCCNT51525.2021.9579717.

[7] K. L. Kohsasih, B. H. Hayadi, Robet, C. Juliandy, O. Pribadi and Andi, "Sentiment Analysis for Financial News Using RNN-LSTM Network," 2022 4th International Conference on Cybernetics and Intelligent System (ICORIS), Prapat, Indonesia, 2022, pp. 1-6, doi: 10.1109/ICORIS56080.2022.10031595.

[8] R. R. Sornalakshmi, M. Ramu, K. Raghuveer, V. A. Vuyyuru, D. Venkateswarlu and A. Balakumar, "Harmful News Detection using BERT model through sentimental analysis" in 2024 International Conference on Intelligent Systems and Advanced Applications (ICISAA), Pune, India, 2024, pp. 1-5, doi: 10.1109/ICISAA62385.2024.10828690.

[9] K. Mujahid, S. Bhatti and M. Memon, "Classification of URDU headline news using Bidirectional Encoder Representation from Transformer and Traditional Machine learning Algorithm" in 2021 6th International Multi-Topic ICT Conference (IMTIC), Jamshoro & Karachi, Pakistan, 2021, pp. 1-5, doi: 10.1109/IMTIC53841.2021.9719828.

[10] P. S, S. M, V. K. S and S. N. S, "News Category Classification using Natural Language Processing Transformer" in 2023 Second International Conference on Augmented Intelligence and Sustainable Systems (ICAISS), Trichy, India, 2023, pp. 1185-1189, doi: 10.1109/ICAISS58487.2023.10250566.

[11] F. Al-Quayed, D. Javed, N. Z. Jhanjhi, M. Humayun and T. S. Alnusairi, "A Hybrid Transformer-Based Model for Optimizing Fake News Detection," in IEEE Access, vol. 12, pp. 160822-160834, 2024, doi: 10.1109/ACCESS.2024.3476432

[12] A. Didolkar and P. Lokulwar, "Sentiment Detection in Financial News: A Deep learning Approach to Extreme and Moderate Classification", in 2025 International Conference on Intelligent and Innovative Technologies in Computing, Electrical and Electronics (IITCEE), Bangalore, India, 2025, pp. 1-8, doi: 10.1109/IITCEE64140.2025.10915503.

[13] A. Singh and G. Jain, "Sentiment Analysis of News Headlines Using Simple Transformers," in 2021 Asian Conference on Innovation in Technology (ASIANCON), PUNE, India, 2021, pp. 1-6, doi: 10.1109/ASIANCON51346.2021.9544806.

[14] S. A. Soomro, S. S. Yuhaniz, M. A. Dootio, G. Mujtaba and J. A. Siddiqui, "Category-Based Sentiment Analysis of Sindhi News Headlines Using Machine Learning, Deep Learning, and Transformer Models," in IEEE Access, vol. 13, pp. 99985-100001, 2025, doi: 10.1109/ACCESS.2025.3576853.

[15] M. S. Khatun and I. Khan, "Bangla Counterfeit News Identification: Using the Power of BERT," in 2024 IEEE International Conference on Power, Electrical, Electronics and Industrial Applications (PEEIACON), Rajshahi, Bangladesh, 2024, pp. 518-522, doi: 10.1109/PEEIACON63629.2024.10800650.

[16] A. E. O. Carosia, A. E. A. Silva and G. P. Coelho, ”Predicting the Brazilian Stock Market with Sentiment Analysis, Technical Indicators and Stock Prices: A Deep Learning Approach”, in Comput Econ 65, 2351–2378 (2025). doi: https://doi.org/10.1007/s10614-024-10636-y

[17] G. Piorino, V. Moreira, L. H. Q. Lima, A. C. S. Pagano, A. S. Pagano, and A. P. C. da Silva, “Sentiment Analysis of Shared Content in Brazilian Reddit Communities”, in JIS, vol. 16, no. 1, pp. 666–686, Aug. 2025.

[18] L. Barberia, P. Schmalz, N. T. Roman, B. Lombard and T. M. Souza, ”It’s about What and How you say it: A Corpus with Stance and Sentiment Annotation for COVID-19 Vaccines Posts on X/Twitter by Brazilian Political Elites”, in Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities, Albuquerque, USA, 2025. pp. 365-376. doi: https://10.18653/v1/2025.nlp4dh-1.32

[19] C. N. Souza, J. Martínez-Arribas, R. A. Correia, J. A. G. R. Almeida, R. Ladle, A. S. Vaz and A. C. Malhado, ”Using social media and machine learning to understand sentiments towards Brazilian National Parks”, in Biological Psychiatry, Elsevier, 2024. doi:https://doi.org/10.1016/j.biocon.2024.110557.

[20] G. Araújo, T. Melo and C. M. S. Figueiredo, ”Is ChatGPT an effective solver of sentiment analysis tasks in Portuguese? A Preliminary Study” in. Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1, Santiago de Compostela, Galicia/Spain, 2024, pp. 13-21.

[21] F. Souza, R. Nogueira, and R. Lotufo, “BERTimbau: Pretrained BERT Models for Brazilian Portuguese” in Intelligent Systems: 9th Brazilian Conference on Intelligent Systems (BRACIS 2020), Springer, 2020, pp. 403–417. doi: https://doi.org/10.1007/978-3-030-61377-8_28

[22] V. Sanh, L. Debut, J. Chaumond and T. Wolf, “DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter”, in 5th Workshop on Energy Efficient Machine Learning and Cognitive Computing - NeurIPS 2019, 2019. doi: https://doi.org/10.48550/arXiv.1910.01108

[23] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer and V. Stoyanov, “Roberta: A robustly optimized Bert pretraining approach”. in arXiv preprint arXiv:1907.11692, 2019. doi: https://doi.org/10.48550/arXiv.1907.11692

[24] A. Karimi, L. Rossi and A. Prati, ”Adversarial Training for Aspect-Based Sentiment Analysis with BERT,” 2020 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 2021, pp. 8797-8803, doi: 10.1109/ICPR48806.2021.9412167.

[25] I. Loshchilov and F. Hutter, ”Decoupled Weight Decay Regularization”, in Proceedings of the 7th International Conference on Learning Representations (ICLR), 2019.

APPENDIX A
CONFUSION MATRICES OF EACH MODEL

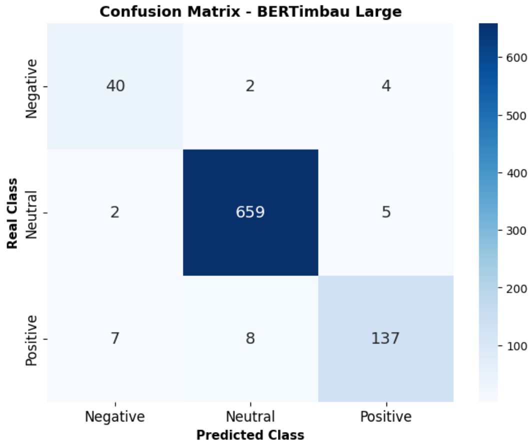

Fig. 8. Confusion Matrix - BERTimbau Large

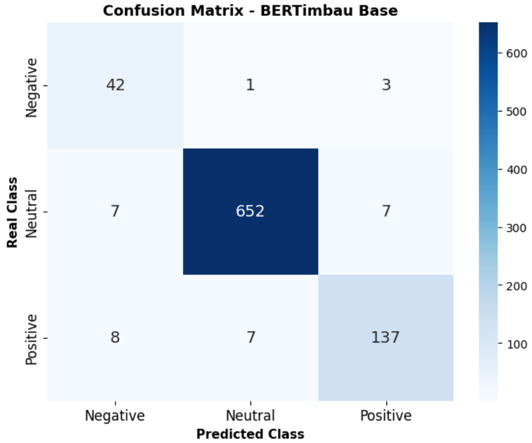

Fig. 9. Confusion Matrix - BERTimbau Base

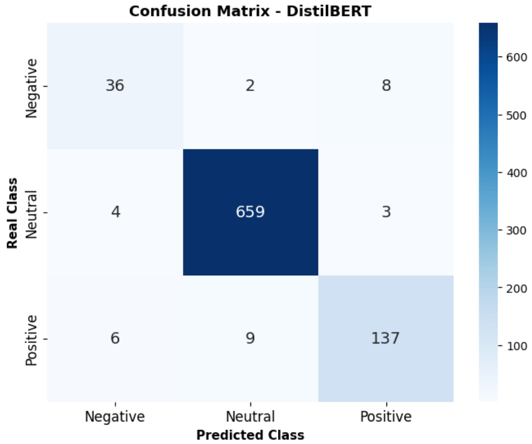

Fig. 10. Confusion Matrix - DistilBERT

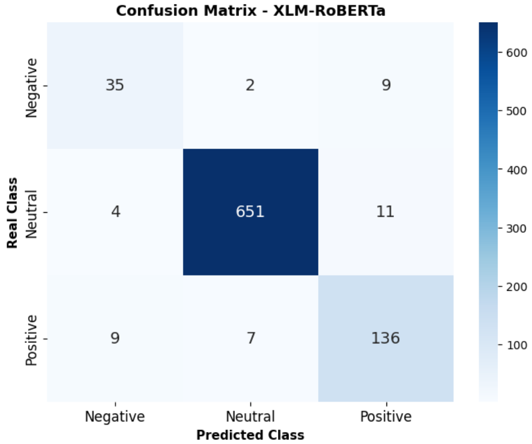

Fig. 11. Confusion Matrix - XLM-RoBERTa