Exploring Sentiment Analysis Approaches in a Public Agency Security News Dataset

Thiago Ruiz Lobo
Institute of Computing
Federal University of Mato Grosso
Brazil, 78060-900
thiago.lobo@sou.ufmt.br

Claudia Aparecida Martins
Institute of Computing
Federal University of Mato Grosso
Brazil, 78060-900
claudia@ic.ufmt.br

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 paper investigates the application of deep learning algorithms in sentiment analysis of headline news about a public security institution. Four deep learning methods were applied in combination with three textual representations, resulting in twelve trained models. For each combination, a class-based analysis of the results was conducted. Models using BERT as the textual representation achieved strong performance, with an F1-score of approximately 90%.

1 Introduction

Perícia Oficial e Identificação Técnica (POLITEC) is a state public security agency responsible for producing expert reports that inform judicial decisions and for issuing identity cards to the state's population. To guide its actions in a structured manner, the Strategic Actions Center (SAC) prepared the institution's strategic plan for 2024-2027, which contains 16 Strategic Objectives (SO) with targets and indicators. Among the objectives, SO 02 aims to strengthen the institutional image of agency, seeking to understand and classify news published in 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.

Computational techniques can be used in the process of identifying and classifying textual data. Natural Language Processing (NLP) is an area related to text processing and interpretation, enabling the extraction of information and patterns, the identification of entities, such as places or organizations, the categorization of documents according to content, or based on the emotion that a given text may express, classifying it as positive, neutral or negative, for example. In this way, combined with the field of Machine Learning (ML), it is possible to perform text processing for news classification and to evaluate comments by detecting positive or negative sentiments present in various media, such as social networks and news pages. This task is known as Sentiment Analysis (SA) and can be applied in multiple domains, enabling companies or institutions to make decisions based on the evaluation of feelings about news published in the media (Prasad et al., 2023).

In this context, this work aims to apply, analyze, and compare the performance of Deep Learning (DL) techniques, in the analysis of sentiments related to news headlines about the agency in the main media outlets of the state. This work is presented as follows: Section 2 presents Related Work. Section 3 describes the Materials and Methods used in the news headline classification processing. In Section 4, the performance of the Results obtained is presented and analyzed, and, finally, in Section 5, the Conclusion is presented.

2 Related Work

The SA field has a large range of application and basically can be divided into two fronts, the Opinion Mining area, which includes tasks related with subjectivity detection and polarity classification and the Emotion Mining area, that goes beyond the common polarity Analysis of a text, and searches for specifics human emotions presents in texts (Islam et al., 2024).

Therefore, 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 (Grant and Booth, 2009). To this end, searches were carried out in the IEEE Xplore, ScienceDirect, Springer Nature and SBC-OpenLib databases to identify recent works that applied the use of SA in Opinion or Emotion Mining, specifically applied

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 1, pages 878–888

to news and news headlines.

Nowadays, with the dissemination of unreliable information, the growth of toxic environments in social media and the, the flagship of SA resides in tasks related with fake news detection, hateful speech detection and customer review analysis.

Ayyasamy et al. (2025) explore a new method for increasingly accurate fake news detection, which combines a Long Short-Term Memory (LSTM) network for contextual feature extraction with a Convolutional Gaussian Perceptron Neural Network (CGPNN) for classification and a metaheuristic Moth-Flame Whale Optimization (MFWO) algorithm for hyperparameter tuning. The combination of these features were tested in established fake news detection datasets, achieving up to 98% accuracy, 95% F1-score, and statistically significant improvements ($p<0.05$) over transformer-based and graph neural network models.

Similarly, Dhal et al. (2023) developed a sentiment-based for Fake News Classification (FNC). The deep network consists of a stacked layer based feature extraction approach. Here, stack layer first extracts the global features by the network of Bidirectional Long Short-Term Memory (Bi-LSTM) and Bidirectional-Gated Recurrent Unit (Bi-GRU).

In the spectrum of hateful speech detection, Soto et al. (2019) compares the results of two word embeddings (word2vec and wang2vec) combined with a CNN. The developed models were tested in three hateful speech datasets showing that embeddings with lower dimensions can achieve satisfactory accuracy in this type of task. Still in this context, Ayetiran and Özgöbek (2024) worked on a unified DL model that can be used for multi modal fake news, hate speech and offensive language detection.

Even so, SA can be used to better understand financial news and their impact. Didolkar and Lokulwar (2025) developed a system based on DL techniques for sentiment analysis of financial news. The system aims to identify sentiment trends in financial news to aid decision-making. To this end, the work compares three types of neural network architectures: RNN, CNN, and Transformers, showing that the BERT model, based on the Transformers architecture, achieved 90% accuracy, higher than the 80% accuracy obtained by models based on RNN and 83% accuracy obtained by the model based on CNN.

In the spectrum of SA applied in news polarity, Sreekumar et al. (2023) conducted a comparative study between models developed using DL techniques to classify news from the BBC dataset. The LSTM, BiLSTM, and CNN techniques were used, combined with the use of three types of feature extraction techniques, namely TF-IDF, Word2vec, and GloVe. The models that combined Word2Vec or GloVe with the CNN, LSTM, or BiLSTM techniques presented better results, while the combinations that used TF-IDF achieved the lowest accuracy.

Singh and Jain (2021) applied the concepts of SA to news headlines using Transformers, focusing on Simple Transformers. Four pre-trained models were compared in classifying headlines as negative, neutral, or positive: BERT, RoBERTa, DistilBERT, and XLNe, highlighting the BERT model that obtained the best result, with 90.1%.

Even due to the lack of annotated corpora in the Sindhi language, Soomro et al. (2022) introduced the Sindhi News Headlines Dataset (SNHD), a novel corpus annotated for both SA and Topic Classification. To evaluate the effectiveness, multiple ML, DL and Transformers-based approaches were used.

Wang et al. (2023) present an approach that uses a BERT model pre-trained with a CNN. The BERT model is applied to produce contextualized embeddings, while the CNN classifier uses such embeddings to train the model. This combination was tested on two public datasets, AG News and Amazon Product Reviews, and obtained performances above 90% accuracy.

In the work of Prathaban et al. (2025), it was evaluate a group of ML models for SA and Topic Classification of Indian news articles using VGGNet, Alexnet and IndicBERT model (IBM) to classify news articles based on their sentiment and topics.

In Brazil, Gumiel et al. (2021) addressed sentiment analysis in the health domain, specifically in online forums about Diabetes Mellitus in Portuguese. The authors contributed by creating a new corpus containing 1,290 annotated posts in three sentiment categories (positive, neutral, and negative). The methodology involved comparing classical machine learning algorithms with state-of-the-art models based on BERT.

In parallel, Souza and Souza Filho (2022) presented a comprehensive experimental study focused on generating embeddings for binary sentiment classification in user reviews (e-commerce)

in Brazil, comparing various vector representation techniques, from classical natural language processing methods such as Bag-of-Words and TF-IDF, to deep learning models, including CNN, RNN, and Transformer-based models, such as BERTimbau.

Seno et al. (2024) investigated the application of the ChatGPT language model in aspect-based sentiment analysis (ABSA) in the domain of political discourse in Brazilian Portuguese. For aspect detection, the results indicated that simple heuristics based on named entity recognition outperformed ChatGPT. However, in polarity classification, ChatGPT demonstrated superiority over traditional knowledge-based methods and fine-tuned BERT models, achieving a macro F-measure of 57.88%. The study concludes that, while promising for associating sentiments with aspects, the model still faces challenges in accurately identifying aspects in informal social media texts.

Specifically in Brazilian Public Security Domain, Nascimento et al. (2025) initiated the construction of a Portuguese-language corpus composed of news articles related to public security, through the automatic collection and preprocessing of journalistic content. This linguistic resource is intended to support the development and evaluation of Natural Language Processing applications focused on analyzing public security discourse in Brazil.

In turn, Lobo and Martins (2024) investigated the impact of news headlines on public security agencies by comparing different ML techniques applied to sentiment analysis of institutional news. In that study, traditional ML algorithms were combined with Bag-of-Words and TF-IDF representations, as well as oversampling strategies to mitigate class imbalance. However, due to the imbalance between sentiment classes and the sensitivity of the task, the models showed difficulty in correctly identifying negative and news. This result suggests that more expressive textual representations, such as word embeddings and Transformer-based language models, may offer important gains in performance.

Taken together with previous research applying sentiment analysis to other areas, including hate speech detection, fake news identification, and financial market sentiment, these studies reinforce the relevance and versatility of polarity classification techniques in journalistic contexts. Nevertheless, applications involving Brazilian Portuguese news in the domain of public security remain scarce, particularly when the content is produced by public institutions. In this context, the present work advances the state of the art by applying modern NLP techniques to a manually annotated, domain-specific dataset, with the objective of analyzing the polarity of news related to a Brazilian public security agency.

3 Materials and Methods

In order to identify the best model for sentiment analysis in headlines, this work proposed a methodology, that can be followed below, based on the combination of different semantic representations through word embeddings, neural network architectures, and hyperparameter variations, using Stratified K-Fold cross-validation.

3.1 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 each news item, its publication date, website, headline, and news article were extracted. In the end, a set of 28,572 news items distributed between the years 2005 and 2024 was stored in a CSV file.

However, for this study, the database was composed initially of 7,109 samples from the years 2023 and 2024. Even after the full labeling of the 2023 and 2024 samples, the positive and negative classes showed low representation.

Therefore, to increase the number of samples in these classes, 1,529 additional headlines from previous years were identified based on keywords associated with negative and positive contexts and subsequently labeled. Words such as “POLITEC”, “laudos”, and “Perícia” were examples of terms used in a positive context, while words such as "suspende", “atrasar” and “espera” represented a negative context. In the Figure 1 is shown the distribution of samples across each class, resulting in a final dataset of 8,638 samples.

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;

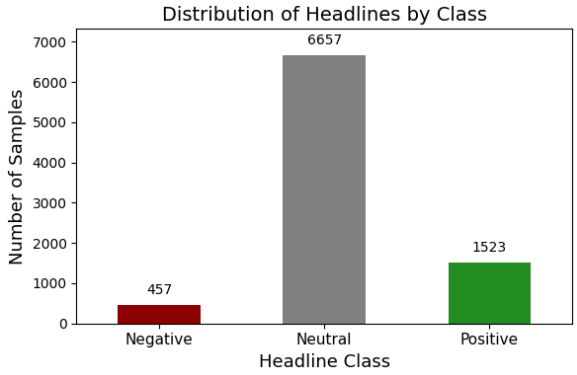

Figure 1: Distribution of samples by class

- Negative headlines present criticism, institutional failures or the involvement of civil servants in crimes and scandals.

3.2 Data Analysis

The statistical analysis of the textual data initially explored the length of the headlines to define the input length of the sequences in the BERT models. In the Figure 2 is presented a box plot that indicates an average length of approximately 8.58 words per headline. The first quartile is 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.

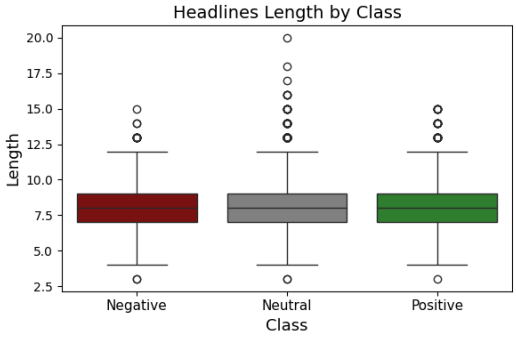

Figure 2: Boxplot of Headline Length by Class

Next, 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 "laudo" (242), "perícia" (173) and "corpo" (118) feature prominently, suggesting a focus on possible developments in expertise. Words like "novo" (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 "laudo" (86), "não" (53), "espera" (47) and "corpo" (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 "ser" (2864), "morrer" (2023), "homem" (1486), "matar" (1386), and "atirar" (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.

3.3 Preprocessing

The preprocessing of textual information is generally carried out in four stages: tokenization, data cleaning, normalization, and removal of stopwords (Aydin and Zeliha, 2022). Thus, initially, the tokenization process was carried out for each news headline, that is, the text was separated into smaller units (tokens). Afterward, in the data cleaning stage, only possible punctuation marks and special characters were eliminated. Next, the uppercase letters were normalized to lowercase, and words considered stopwords were eliminated.

During the stopword removal process, it was observed that standard libraries of the Portuguese language often exclude terms such as "no" and "without", which have strong semantic value, especially in headlines with negative connotations. Considering this aspect, it was decided to develop a customized set of stopwords, constructed from the analysis of the 200 most frequent words in the database, removing only prepositions and articles present ("em", "de", "e", "no", "na", "da", "do", "o" and "a").

3.4 Feature Extraction Methods

The Word2Vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014) embeddings provides efficient representations for words using vectors of real numbers in an n-dimensional space. These

models learn syntactic and semantic relationships through large data corpora.

With the dissemination of these embeddings, the Interinstitutional Center for Computational Linguistics (Núcleo Interinstitucional de Linguística Computacional - NILC) developed the version of these methods adapted to brazilian corpora (Hartmann et al., 2017).

Some years later, with the advancement of text representation techniques and the emergence of pre-trained models based on Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019), BERTimbau was developed, a version of BERT trained specifically for Brazilian Portuguese (Souza et al., 2020).

In this context, this work employed the Word2Vec Skip-Gram and GloVe word embeddings provided by NILC, both with 50-dimensional vectors, as well as the BERTimbau-Base model with 768-dimensional contextual embeddings. All models are publicly available on Hugging Face.

To reduce the size of the neural network inputs, a maximum length of 30 tokens per headline was defined, given that the longest headline in the dataset contains 20 tokens. Headlines longer than 30 tokens were truncated, whereas shorter ones were padded with zeros. The choice of a value larger than the maximum observed length was intended to allow the trained model to properly handle longer headlines that may appear during deployment.

As a result, each headline was represented as a 30×50 matrix when using Word2Vec or GloVe embeddings, and as a 30×768 matrix when using BERTimbau-Base. After completing these preprocessing steps, the dataset was prepared to serve as input to the neural networks.

3.5 Implemented and Model Training Algorithms

This work applied four classics of Artificial Neural Networks (ANNs). From the family of RNN, three architectures were chosen, the LSTM (Hochreiter and Schmidhuber, 1997), Bi-LSTM (Graves and Schmidhuber, 2005) and Gated Recurrent Unit (GRU) (Cho et al., 2014). The last ANN was the Text Convolutional Neural Network (TextCNN), a variation of CNN, commonly used for challenges related to text classification (Yang, 2022).

The training procedure began with the delimitation of the hyperparameter combinations, with two options for the number of neurons in the first layer, two dropout values, and two learning rate values, thus totaling eight possible hyperparameter combinations.

Each one of the four architectures were built in a similar way, all with two intermediate layers, with the first layer varying between 64 and 128 neurons and the second layer fixed at 64 neurons. The dropout value varied between 0.3 and 0.5, applied in all layers to mitigate overfitting. The output layer of all architectures was configured with the softmax activation function, allowing the obtaining of probabilities for each of the three classes. Finally, the learning rate value of the Adam optimizer varied between 0.001 (default value) and 0.003, being selected due to its recognized efficiency in adapting the learning rates of the parameters. The categorical_crossentropy loss function was chosen because it is widely used in multi-class classification problems.

It is worth noting that for the recurrent architectures (LSTM, BiLSTM, and GRU), the activation functions used in the gates were the hyperbolic tangent (tanh) and the sigmoid (sigmoid), according to the standard implementation of these networks in the Keras library. These functions control, respectively, the output values and the mechanisms for retaining and forgetting information throughout the sequences.

In the case of TextCNN, a single 1D convolutional layer (Conv1D) with a ReLU activation function was also adopted, followed by a global pooling layer (GlobalMaxPooling1D). This combination aims to capture relevant local patterns in the text sequences and aggregate them into a fixed and robust representation for classification.

After configuring the neural network architectures, the model training process began by dividing the data into training and testing sets, with 80% of the data allocated to training and 20% to testing.

Each architecture was trained with all eight hyperparameter combinations. In each combination, the stratified K-fold with k=10 is used, training the model for 30 epochs and batch_size=128. Afterwards, the average F1-score macro of the folds is stored, which represents the score of the combination. The process is repeated for all combinations, aiming to identify the combination with the best average F1-score Macro. And, finally, the final model is trained with the best combination identified, for 50 epochs and batch_size=128.

In order to speed up the training, the process was executed in parallel through three Colab Notebooks, one for each embedding. The notebooks

focused on the Word2Vec and GloVe word embeddings used the A100 GPU, while the notebook with textual representations via BERT used the TPU v6e-1. The code and the dataset used in the methodology stage can be accessed through the following Github link.

It is important to note that the early stopping strategy was not adopted during the training procedure. This decision was made to ensure consistency and comparability across all hyperparameter combinations and architectures evaluated in the experiments. Since each configuration was trained under the same fixed number of epochs during the stratified k-fold cross-validation procedure, the use of early stopping could result in different training durations for each fold and configuration, potentially introducing variability unrelated to the model architecture or hyperparameters themselves.

4 Results

In the Table 1 are shown the results of the best hyperparameter combinations for each architecture and embedding. The models that used BERT outperformed the others, with emphasis on BERT + GRU, which achieved the highest macro F1-score (91.19%), although with a high execution time (88 seconds). In contrast, TextCNN, despite presenting inferior performance with static embeddings such as Word2Vec (79.32%) and GloVe (82.84%), showed computational efficiency, with execution times between 11 and 15 seconds.

The following sections detail the results obtained by each type of word embedding with the accuracy, loss, and F1-score metrics per class for analysis of the results.

4.1 Results using Word2Vec representation

In the Figure 3 are shown the behavior of accuracy and loss at each epoch for the best models generated with Word2Vec as Word Embedding. These models stand out for achieving an accuracy higher than 90% before ten epochs of training, with a loss close to 10% in the last epochs.

In the Table 2 is presented that the F1-scores obtained are similar in all classes, except for the TextCNN model. It is worth mentioning that for the positive and negative classes, the LSTM and GRU models obtained the best results, with 87.93% and 87.68%, respectively, for the positive class and 74.72% and 76.19%, respectively, for the negative class.

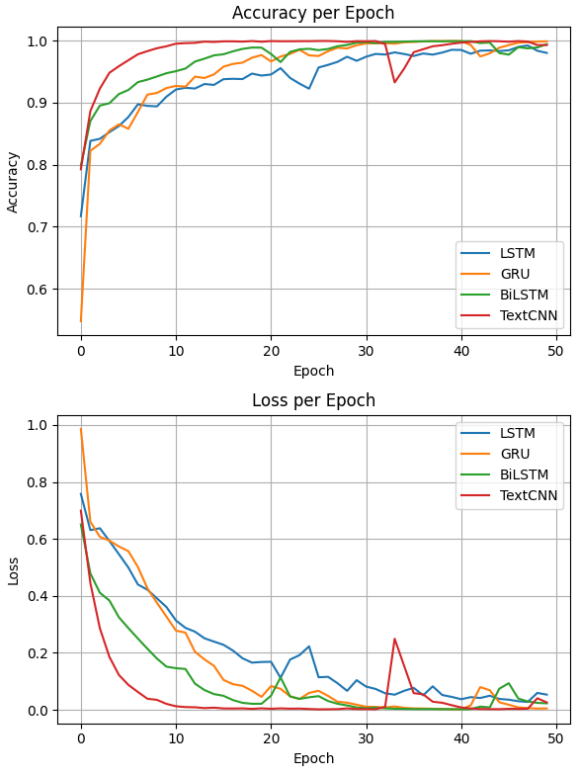

Figure 3: Accuracy and loss per epoch of models with pre-trained Word2Vec Embedding

4.2 Results using GloVe representation

In the Figure 4 are shown the behavior of accuracy and loss at each epoch for the best models generated with GloVe as Word Embedding. Except for the LSTM model, which obtained an accuracy close to 95% and a loss close to 20%, all the others stood out for obtaining an accuracy close to 99% at the end of all epochs and a loss below 5%.

In the Table 3, it can be seen that all models obtained similar performances in the three classes evaluated, especially in the positive class, with F1-scores above 85%, with the exception of the TextCNN model, which presented the lowest performance with 65% for the negative class.

4.3 Results using BERT representation

In the Figure 5 are shown the behavior of accuracy and loss per epoch of the models trained with the pre-trained BERT embedding. The use of BERT enhanced the good results previously achieved by the other Word Embeddings, with accuracies above 95% in less than 10 epochs. In addition, a rapid reduction in loss is noted in the first epochs, with values stabilizing close to 1% in the last iterations for most models.

In the Table 4 is presented that the models using

Models	Units	Dropout	Learning Rate	F1-Score (%)	Time (s)
Word2Vec + LSTM	128	0.3	0.003	86.77	22
Word2Vec + GRU	128	0.3	0.003	87.23	21
Word2Vec + BiLSTM	128	0.3	0.003	85.73	34
Word2Vec + TextCNN	128	0.3	0.003	79.32	**11**
GloVe + LSTM	64	0.3	0.003	84.90	22
GloVe + GRU	128	0.3	0.003	85.25	22
GloVe + BiLSTM	64	0.3	0.003	85.24	34
GloVe + TextCNN	128	0.5	0.001	82.84	**12**
BERT + LSTM	128	0.3	0.003	88.80	130
BERT + GRU	64	0.5	0.003	**91.19**	88
BERT + BiLSTM	64	0.5	0.001	90.61	102
BERT + TextCNN	64	0.5	0.001	90.64	15

Table 1: Best Matches by Model

Model	Neutral	Positive	Negative
LSTM	97.65%	87.93%	74.72%
GRU	97.83%	87.68%	76.19%
BiLSTM	97.96%	85.95%	73.29%
TextCNN	97.04%	79.56%	61.36%

Table 2: F1-score per Class: Word2Vec pre-trained embedding

Model	Neutral	Positive	Negative
LSTM	97.38%	85.44%	71.87%
GRU	97.18%	86.13%	72.43%
BiLSTM	97.52%	86.69%	71.50%
TextCNN	97.50%	86.04%	65.00%

Table 3: F1-score per Class: GloVe pre-trained embedding

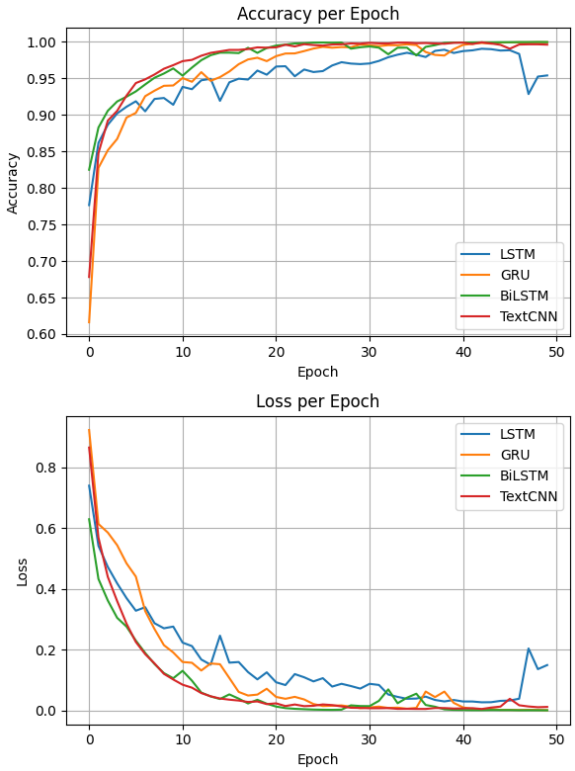

Figure 4: Accuracy and loss per epoch of models with pre-trained GloVe Embedding

BERT representation outperformed other results developed with other types of pre-trained embeddings. The use of the BERT model as a textual representation allowed for gains in the positive classes and, mainly, in the negative class, which presents a greater challenge due to its imbalance, showing that the combination of BERT models for textual representation together with recurrent architectures has great potential in sentiment analysis in imbalanced datasets.

In summary, the model that performed best for the News Headline Sentiment Analysis challenge was the one that combined the use of BERT with GRU, achieving an F1-Score of 91.19%. On the other hand, the combination of textual representation via Word2Vec and the convolutional layer-based technique TextCNN performed worst, with an F1-Score of 79.32%. These achievements demonstrate that robust textual representations, such as those provided by pre-trained BERT-type models, are capable of achieving results that are not obtained by other approaches.

4.4 Error Analysis of Models

To evaluate the models performance in the SA task, a set of nine headline news related to public safety was selected. Each sentence was previously manually labeled according to its polarity (negative,

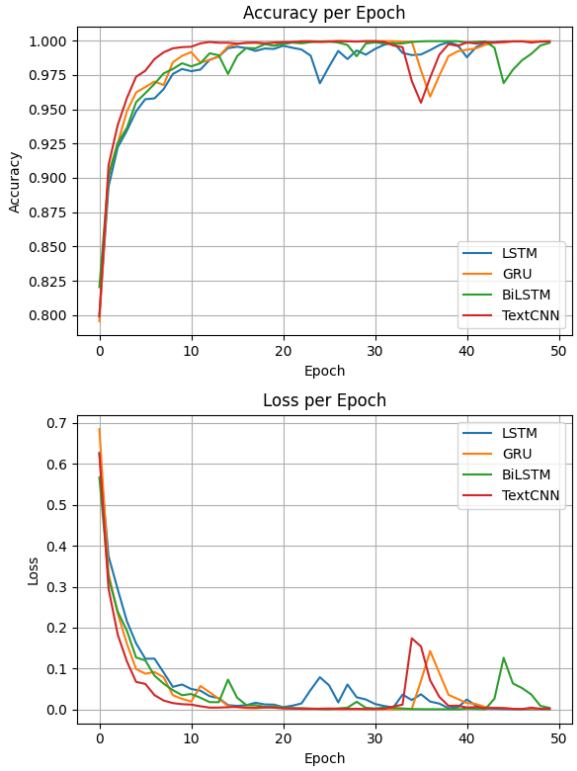

Figure 5: Accuracy and loss per epoch of models with pre-trained BERT Embedding

Model	Neutral	Positive	Negative
LSTM	97.85%	87.43%	81.11%
GRU	97.76%	90.07%	84.94%
BiLSTM	97.86%	89.18%	84.78%
TextCNN	98.11%	89.93%	84.78%

Table 4: F1-score per Class: BERT pre-trained embedding

neutral, or positive) to allow for a direct comparison between the reference label and the prediction made by each model. The headlines (H) used in the experiment are:

- H1 - The court warns that delays in forensic analysis can lead to the release of criminals in the state **(Negative)**
- H2 - AL Citizen Space informs of the suspension of services this Friday (26) **(Negative)**
- H3 - Forensic Police does not set a deadline for the investigation into the shopping mall fire; shop owners are moving to a specific neighborhood. **(Negative)**
- H4 - Criminals invade house and shoot man dead near his wife. **(Neutral)**
- H5 - Woman under investigation for mistreatment violates ankle monitor again, and association requests revocation of the precautionary measure. **(Neutral)**
- H6 - Police find drugs and stolen motorcycle while arresting pharmacy thieves in the state. **(Neutral)**
- H7 - Project identifies 42 people buried as indigents in the city. **(Positive)**
- H8 - Forensic Police confirms presence of genetic material from suspect in victim of sexual crime in the state. **(Positive)**
- H9 - Forensic Police task force completes 80 ballistics reports in the city **(Positive)**

It was observed that none of the architectures presented classification errors for the headlines H1, H3, H4, H6, H8, and H9 indicating high consistency of the models in these instances. The BERT+BiLSTM model was the only one capable of correctly predicting all classes, showing no errors among the nine headlines evaluated. The only sentences with an error rate were H2, H5, and H7.

In the Table 5 is presented the headlines that were misclassified by the models. These errors may indicate the presence of words with positive or negative polarity that, in some way, mistakenly influence the classification performed by the models. Through tools such as NILC-Metrix (Leal et al., 2023), it is possible to extract a set of linguistic metrics from a sentence. Among these metrics, two stand out that measure the proportion of positive and negative words in a text, positive_words and negative_words. The identification of word polarity is performed using the LIWC Brazilian Dictionary for Sentiment Analysis (Pennebaker et al., 2001) (Filho et al., 2013).

When analyzing the three headlines that presented classification errors, it is observed that the H2 headline has a ratio of 0.375 positive words and 0.25 negative words. This result suggests that the higher incidence of positive terms may have contributed to the misclassification by the models.

In the case of the H5 and H7 headlines, both show a higher proportion of negative words. The H5 headline has 0.66 negative words and 0.33 positive words, while the H7 headline has a proportion of 1.4 negative words, with no occurrence of positive terms. This context may explain both the incorrect classification of H7, originally labeled as

positive, and the inadequate interpretation observed in the models for H5.

Models	H2	H5	H7
Word2Vec+LSTM	**Pos**	Neu	**Neu**
Word2Vec+GRU	**Neu**	Neu	**Neu**
Word2Vec+BiLSTM	**Neu**	Neu	**Neu**
Word2Vec+TextCNN	**Neu**	Neu	**Neu**
GloVe+LSTM	**Neu**	Neu	**Neu**
GloVe+GRU	Neg	**Neg**	**Neu**
GloVe+BiLSTM	Neg	Neu	**Neu**
GloVe+TextCNN	**Pos**	Neu	**Neu**
BERT+LSTM	**Neu**	Neu	**Neu**
BERT+GRU	**Neu**	Neu	**Neu**
BERT+BiLSTM	Neg	Neu	Pos
BERT+TextCNN	Neu	Neu	**Neu**

Table 5: Selected Examples of Qualitative Error Analysis of Models

5 Limitations

Among the study's limitations, the fact that the annotation process was conducted by only one person stands out. Although criteria previously defined by the communication advisory team and management were adopted, the use of multiple annotators could contribute to greater methodological robustness and enrich the discussion of the results. Even so, it is important to emphasize that, in cases where there was doubt regarding the polarity of the headlines, decisions were discussed jointly with members of the communication advisory team and management, seeking to reduce possible individual biases.

Another aspect to be considered refers to the need for a deeper understanding of the differences between the linguistic characteristics of each of the classes analyzed. The Nilc-Metrix tool, used only at the end of the process, could have been employed more systematically throughout the analysis. Using specific linguistic metrics, it would have been possible to conduct a more detailed and consistent investigation of the distinctions between the classes, contributing to a more robust interpretation of the data.

6 Conclusion

This work involved investigating the application of several word embeddings in conjunction with deep learning techniques to create news headline classification models, specifically in sentiment analysis, within the domain of news headlines related to news of Forensic Police.

From the results obtained, it can be seen that, among the combinations of deep learning models, the models that used the BERT model presented F1-scores higher than 88%, with emphasis on the model that combined the GRU and BERT techniques, which presented the best F1-score values for positive and negative headlines, respectively 90.07% and 84.94%, something satisfactory considering the existing class imbalance.

As future work, it is important to note that the proposed model is already being employed for the monthly evaluation of news related to the institution. For further model improvement, additional approaches based on transformer architectures will be explored and systematically compared. Moreover, in order to mitigate classification errors, an extensive analysis of the linguistic metrics of the evaluated samples will be conducted, focusing on metrics that highlight the presence of positive and negative linguistic features as defined by the NILC-Metrix framework.

7 Generative Artificial Inteligence Usage

The use of Generative Artificial Intelligence in this work occurred primarily in the linguistic refinement process of texts translated into English. In addition, preliminary versions of queries for searching for articles related to the topic were also developed with the aid of these tools, taking advantage of their ability to support the initial formulation of search strategies. Regarding the tools employed, the ChatGPT software was predominantly used.

References

E. Aydin and Zeliha. 2022. Sentiment analysis about turkish tv series with web scraping. In *2022 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)*, pages 1–4.

E. F. Ayetiran and O. Özgöbek. 2024. An inter-modal attention-based deep learning framework using unified modality for multimodal fake news, hate speech and offensive language detection. In *Information Systems*, volume 123, page 102378.

R. K. Ayyasamy, C. Ponnusamy, K. N. Bhargavi, S. Cherukuvada, G. Charles Babu, and D. T. Amutha, S. Gamu. 2025. A hybrid deep learning framework for fake news detection using lstm-cgpnn and metaheuristic

optimization. In *Sci Rep 15, 41522. https://doi-org.ez52.periodicos.capes.gov.br/10.1038/s41598-025-25311-x.*

K. Cho, B. V. Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. 2014. Learning phrase representations using rnn encoder–decoder for statistical machine translation.

J. Devlin, M. Chang, K. Lee, and K. Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT.*

P. Dhal, D. Biswas, Pragya, J. Patra, M. A. Mishra, and P Jha Kumar. 2023. Stacked layer based deep learning approach for fake news classification. In *7th International Conference On Computing, Communication, Control And Automation (ICCUBEA), Pune, India, 2023, pp. 1-5, doi: 10.1109/ICCUBEA58933.2023.10392014.*

A. Didolkar and P. Lokulwar. 2025. Sentiment detection in financial news: A deep learning approach to extreme and moderate classification. In *International Conference on Intelligent and Innovative Technologies in Computing, Electrical and Electronics (IITCEE)*, pages 1–6.

P. P. B. Filho, T. A. S. Pardo, and S. M." Aluísio. 2013. An evaluation of the Brazilian Portuguese LIWC dictionary for sentiment analysis. In *Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology*.

M. J. Grant and A. Booth. 2009. A typology of reviews: an analysis of 14 review types and associated methodologies. *Health Information & Libraries Journal*, 26(2):91–108.

A. Graves and J. Schmidhuber. 2005. Framewise phoneme classification with bidirectional lstm networks.

Y. B. Gumiel, I. Lee, T. A. Soares, T. C. Ferreira, and A. Pagano. 2021. Sentiment analysis in Portuguese texts from online health community forums: Data, model and evaluation. In *Proceedings of the 13th Brazilian Symposium in Information and Human Language Technology*, pages 64–72.

N. S. Hartmann, E. Fonseca, C. D. Shulby, M. V. Treviso, J. S. Rodrigues, and S. M. Aluísio. 2017. Portuguese word embeddings: Evaluating on word analogies and natural language tasks. In *Symposium in Information and Human Language Technology (STIL).*

S. Hochreiter and J. Schmidhuber. 1997. Long short-term memory. In *Neural Computation*, volume 9, pages 1735–1780.

S. Islam, M. N. Kabir, N. A. Ghani, K. Z. Zamli, N. Zulkifli, M Rahman, and M. Moni. 2024. Challenges and future in deep learning for sentiment analysis: a comprehensive review and a proposed novel hybrid approach. In *Artif Intell Rev 57, 62. https://doi.org/10.1007/s10462-023-10651-9.*

S. E. Leal, M. S. Duran, C. E. Scarton, S. Hartmann, N, and S. M. Aluísio. 2023. Nilc-metrix: assessing the complexity of written and spoken language in brazilian portuguese. *Lang Resources Evaluation.*

T. Lobo and C. Martins. 2024. 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*, pages 72–77, Porto Alegre, RS, Brasil. SBC.

T. Mikolov, K. Chen, G. Corrado, and J. Dean. 2013. Efficient estimation of word representations in vector space. In *International Conference on Learning Representations (ICLR).*

M. Nascimento, V. Silva, G. Souza, K. Lima, E. Araújo, E. Souza, J. Turet, and V. Carvalho. 2025. Extração de notícias sobre segurança pública para desenvolvimento de corpora em português: uma análise preliminar em cidades do nordeste brasileiro. In *Anais do VI Workshop sobre as Implicações da Computação na Sociedade*, pages 267–276, Porto Alegre, RS, Brasil. SBC.

J. W. Pennebaker, M. E. Francis, and R. J. Booth. 2001. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates.*

J. Pennington, R. Socher, and C. D. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

O. Prasad, S. Nandi, V. Dogra, and D. Diwakar. 2023. A systematic review of nlp methods for sentiment classification of online news articles. In *International Conference on Computing, Communication and Networking Technology*.

B. P. Prathaban, Subash R., B. Sujin, and G. Purushothaman, K. E. Lakshmi. 2025. Precision driven sentiment and topic classification of news articles using indicbert model (ibm). In *2025 8th International Conference on Circuit, Power Computing Technologies (ICCPCT), Kollam, India, 2025, pp. 962-967, doi: 10.1109/ICCPCT65132.2025.11176462.*

E. R. M. Seno, L. G. T. Silva, F. S. I. Anno, F. M. Rocha Junior, and H. M. Caseli. 2024. Aspect-based sentiment analysis in comments on political debates in Portuguese: evaluating the potential of ChatGPT. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese*, pages 312–320.

A. Singh and G. Jain. 2021. Sentiment analysis of news headlines using simple transformers. In *Asian Conference on Innovation in Technology (ASIANCON)*, pages 1–5.

S. A. Soomro, S. S. Yuhaniz, M. A. Dootio, G. Mujtaba Mujtaba, and J. A. Siffiqui. 2022. Category-based

sentiment analysis of sindhi news headlines using machine learning, deep learning, and transformer models. In *in IEEE Access, vol. 13, pp. 99985-100001, 2025, doi: 10.1109/ACCESS.2025.3576853*.

C. Soto, G. Nunes, and J. Gomes. 2019. Avaliação de técnicas de word embedding na tarefa de detecção de discurso de ódio. In *In: ENCONTRO NACIONAL DE INTELIGÊNCIA ARTIFICIAL E COMPUTACIONAL (ENIAC), 16. , 2019, Salvador. Anais [...]. Porto Alegre: Sociedade Brasileira de Computação, 2019 . p. 1020-1031. ISSN 2763-9061. DOI: https://doi.org/10.5753/eniac.2019.9354.*

F. D. Souza and J. B. O. Souza Filho. 2022. Embedding generation for text classification of Brazilian Portuguese user reviews: from bag-of-words to transformers. *Neural Computing and Applications*, 35:24715–24733.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. BERTimbau: pretrained BERT models for Brazilian Portuguese. In *9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23 (to appear)*.

A. Sreekumar, R. Reshma, and B. Athira. 2023. Comparative study of deep learning models for document classification. In *9th International Conference on Smart Computing and Communications (ICSCC)*, pages 1–6.

C. Wang, Y. Li, and Z. Wang. 2023. A novel approach for text classification by combining pre-trained bert model with cnn classifier. In *6th International Conference on Information Systems and Computer Aided Education (ICISCAE)*, pages 1–6.

L. Yang. 2022. A brief introduction of the text classification methods. In *IEEE International Conference on Electrical Engineering, Big Data and Algorithms*.