Desarrollo de una herramienta de análisis de sentimiento en tiempo real para comprender la percepción pública a través de Google Noticias

Christian Daniel España Chamorro
Juan Sebastián Villada Osorio

Monografía presentada para optar al título de Especialista en Analítica y Ciencia de Datos

Asesor
Javier Fernando Botía Valderrama, Doctor en ingeniería electrónica

Universidad de Antioquia
Facultad de Ingeniería
Especialización en Analítica y Ciencia de Datos
Medellín, Antioquia, Colombia
2024

Cita	(España Chamorro & Villada Osorio, 2024)
Referencia **Estilo APA 7 (2020)**	España Chamorro, C. D., & Villada Osorio, J. S. (2024). *Desarrollar una herramienta de análisis de sentimientos que pueda extraer información relevante de Google noticias en tiempo real para comprender la percepción pública sobre un tema específico* Trabajo de grado especialización]. Universidad de Antioquia, Medellín, Colombia.

Especialización en Analítica y Ciencia de Datos, Cohorte VI.

Universidad de Antioquia - www.udea.edu.co

Rector: John Jairo Arboleda Céspedes.
Decano: Julio Cesar Saldarriaga Molina
Jefe departamento: Danny Alexandro Munera Ramirez

El contenido de esta obra corresponde al derecho de expresión de los autores y no compromete el pensamiento institucional de la Universidad de Antioquia ni desata su responsabilidad frente a terceros. Los autores asumen la responsabilidad por los derechos de autor y conexos.

Table of contents

1. Resumen

Hoy en día, en la era digital, los datos de las noticias digitales son la mina de oro de la opinión pública y de las tendencias emergentes en el tiempo. De este proyecto esperamos que surja una potente máquina analítica que nos permita explorar esta riqueza de datos para encontrar significado dentro de ellos. Además del hecho de que queremos que esta herramienta evolucione a través de niveles más altos de sofisticación de la ciencia y el análisis de datos, la visión es que esta herramienta esté en progreso y se actualice continuamente para aprender y adaptarse a entornos cambiantes según sea necesario. Esta información puede tener un valor incalculable para empresas, investigadores y organismos gubernamentales, interesados en conocer mejor las opiniones del público en general sobre diversos temas de la sociedad actual. Se entrenaron desde cero cuatro modelos siguiendo la arquitectura de BERT, para la clasificación de noticias en positivas o negativas. El mejor modelo tuvo un accuracy de 72%, precisión de 80%, recall de 80%, F1 score de 80%, adicionalmente la grafica ROC muestra un 89% de probabilidades de que el modelo clasifique correctamente un ejemplo positivo de uno negativo.

Palabras clave: NLP, Transformers, BERT, Tokenización, Embeddings, Sentiment analysis, pytorch.

github.com/CDspana/Especializacion

2. Abstract

Today in digital era, digital news data is the gold mine of public opinion and time-emerging trend. Coming out of this project we expect a powerful analytical machine that will allow us to explore this wealth of data to find meaning within that data. Besides the fact that we want this tool to evolve across higher levels of sophistication of data science and analysis, the vision is that this tool is in progress and continuously updated to learn and adapt to changing environments as needed. This information can be invaluable to companies, researchers, and government agencies, as they are interested in learning more about the views of the general public on various issues in current society. Four models were trained from scratch following the BERT architecture for the classification of news into positive or negative. The best model obtained an accuracy of 72%, a precision of 80%, a recall of 80% and an F1 score of 80%. In addition, the ROC plot shows an 89% probability that the model correctly classifies a positive example from a negative one.

Keywords: NLP, Transformers, BERT, Tokenización, Embeddings, Sentiment analysis, pytorch.

github.com/CDspana/Especializacion

3. Introduction

Today we are bombarded with information as soon as we check the headlines, just by connecting to an online news source. The question, therefore, remains: how can we strategically process the overwhelming flood of information? As we can see, the speed at which this data is produced means that it is crucial to track public opinion in real time through the tools available.

Sentiment analysis is an incredible tool at our disposal. It is the field that combines artificial intelligence and data science to analyze different emotions and opinions in text and report individual sentiments. Sentiment analysis has proven its usefulness in different situations, from measuring customer satisfaction to policy predictions.

However, unique challenges arise when attempting to use this technology for real-time news analysis. It is difficult to understand and analyze the language due to its intrinsic complexity and the speed at which it evolves. Especially in the context of Colombia, which is noted for its constant and dynamic political and social change, there is a need for tools capable of adapting and providing up-to-date and relevant information.

We intend to create a tool focused on sentiment analysis that operates in real time and is specifically adapted for Colombian news. We want to offer companies, researchers and government entities an effective tool that allows them to instantly know public opinion on various topics of interest. To do so, we use some of the most cutting-edge approaches in natural language processing and machine learning.

To this end, we leverage prior research in similar domains of news and social media sentiment analysis. Puertas [1]

4. Related Works

To gain insight into what people think about certain topics, researchers have conducted numerous studies in the past, using various methods and approaches to analyze public sentiment in news articles. While these works have made great progress, we must keep in mind that they may not work perfectly in all situations.

Puertas and her team of researchers [1] studied the emotions expressed in Twitter posts to predict the outcome of the Colombian presidential elections. While their method worked well for classifying tweets, it is important to remember that cultural and contextual factors unique to Colombia were not taken into account when analyzing emotional expressions. We should also keep in mind that the information we get from Twitter may not be completely accurate or representative of the entire population, which could affect the results and make them less reliable for everyone.

Kumar and Bhushan [2] shared their brilliant idea of using computers to understand people's feelings by observing what they say on social media, in the year 2023. The model we use to understand what people think on social media does not it is perfect. It can be easily changed and there is a lot of false information out there. Additionally, it is difficult to draw broad conclusions because the way people talk and the rules they follow may be different in different places and cultures.

Chauhan and colleagues [3] delved into the field of detailed sentiment analysis, employing advanced techniques such as deep and machine learning to unravel hidden emotions. Although their sophisticated methods contributed to better sentiment classification, they required a large amount of labeled data, which can be expensive and require considerable computational power. Understanding deep learning models can be complicated due to their encrypted nature, which is difficult to decrypt. Furthermore, it is not always clear why a model made a certain decision.

The previous studies we reviewed were really beneficial for our project, as they indicated the relative effectiveness of various methods when analyzing the feelings expressed by people in news and social networks. However, it is essential to address these limitations with a view to developing a more effective tool for analyzing real-time sentiments in Colombia. Our purpose is to build on this solid foundation by creating a specialized tool capable of understanding and analyzing human emotions and thoughts, using genuinely intelligent algorithms and computer programs.

5. Solution architecture

The following is the architecture designed to successfully carry out the degree project. Initially, news are downloaded from google news, which are not labeled. A pre-trained BERT model from hugging face was used to label the news. The BERT model was built and four models were trained from which the best one was chosen. The next step is to perform again a news extraction and to compare the model created from scratch with the pre-trained model.

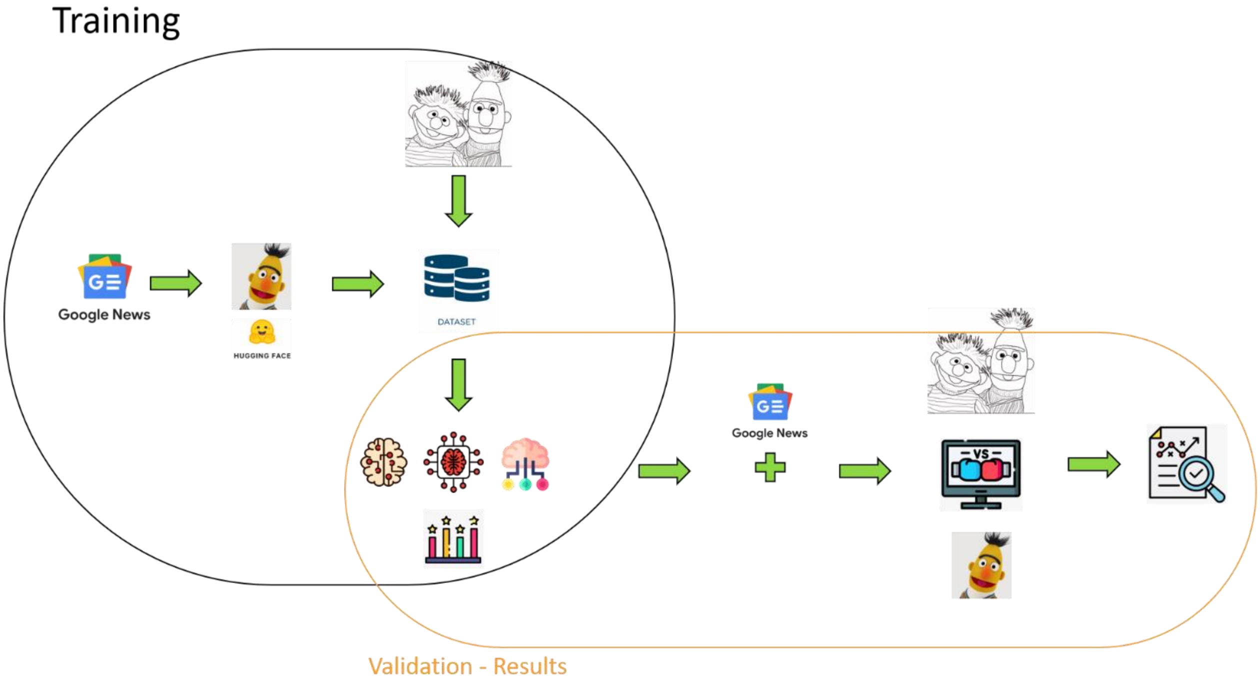

Figure 1. Architecture

6. Methodology

6.1 Dataset

To prepare the BERT model, Spanish news was extracted from Google News through the Gnews library. The dataset is composed of 65.6% news from Colombia and Latin America and 34% news from Spain, extracted from Kaggle with news from larepublica.co.

Temporary Coverage:

The period of Colombia and Latin America news spans from 2000 to 2024; this temporal range ensures that historically significant and more recent events are covered, offering a broad and varied context.

Search Topics:

In order for the system to cover most relevant and topical issues, the following topics are used as search topics:

TABLE I
NEWS TOPICS

Topics
Peace process and post-conflict
Economy and economic development
Foreign policy and international relations
Health care system and access to health care
Security and drug trafficking
Rural and agrarian development
Social movements and human rights
Education and access to education
Climate change and environment
Infrastructure and transportation

The final dataset has a total of 30,000 news items, combining news from Google News [4] and Kaggle [5]. The combination prevents imbalance in the representation of the different areas and subjects in the dataset.

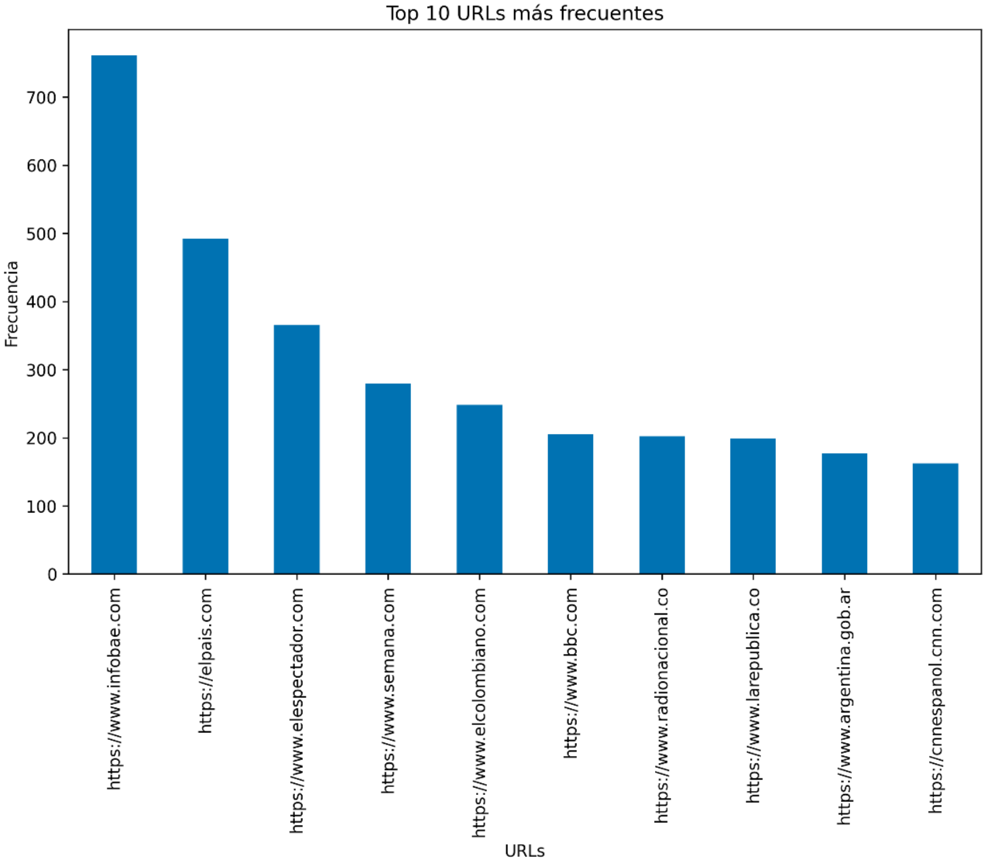

Figure 2. Top 10 news media presented in the dataset.

6.2 Preprocessing

Extensive news cleaning was performed to prepare the data for BERT model training. Preprocessing steps included:

Special Character Cleaning:

News items were identified as having the line break character \n. Using regular expressions (regex), this and other unwanted special characters were removed to make the text uniform.

Emoji preservation:

Unlike other approaches that remove emojis, in this work it was decided to keep them. Emojis are important components of modern language, especially on digital platforms, as they convey emotions and contexts concisely.

Automatic Labeling:

The extracted news items were not initially labeled. A pre-trained Hugging Face model was used for text classification and labeling of the news items 6}. This model allowed assigning sentiment labels (positive or negative) to each news item.

Dataset balancing:

After labeling, 15,000 positive news items and 15,000 negative news items were selected, thus achieving a balanced dataset. This balancing is crucial to avoid biases in the training of the model and to ensure an equal representation of both sentiments.

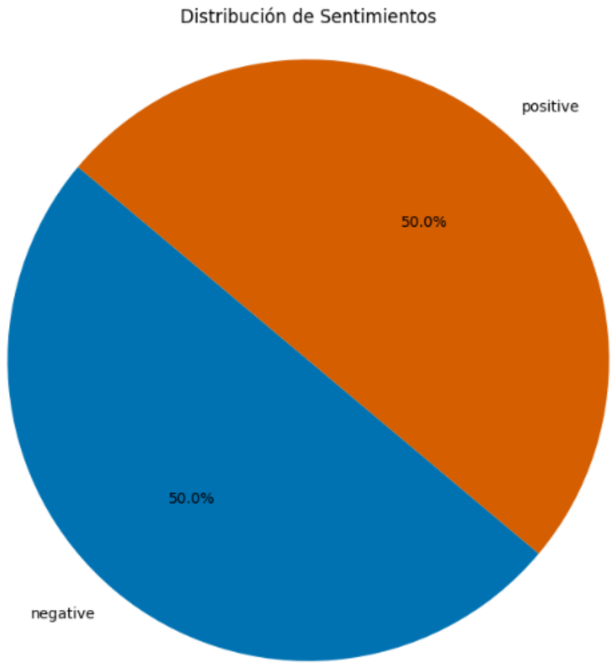

Figure 3. Distribution of the classes in the dataset

6.3 Tokenization

Tokenization is a fundamental step in natural language processing. For this work, a tokenizer was created following the WordPiece approach as established by Hugging Face [7]. This approach ensures efficient segmentation of words into sub-words or tokens, improving the model's ability to handle large and diverse vocabularies.

1. Reserved Tokens

Reserved tokens required for the BERT model such as "[PAD]", "[UNK]", "[CLS]", "[SEP]", "[MASK]" were included. These tokens play specific roles during training and inference, such as indicating the start of a sequence, the end of a sequence, and unknown tokens.

2. Examples of Generated Tokens

By applying the tokenizer on the corpus, tokens such as ['J', 'b', 'Ü', 'Â', 'g'] were generated. These examples show the diversity and capability of the tokenizer to handle special characters and emojis.

6.4 Feature Extraction

In the BERT model, feature extraction is performed through the transformer encoding layers. These layers are essential for transforming the input tokens into vector representations that the model can process. The main components include:

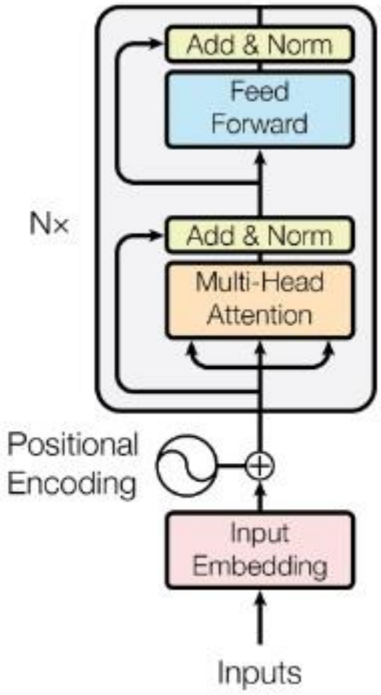

Figure 4. Decoder of transformers. [8]

1. Input Embeddings

- Input tokens are converted into embeddings of a fixed dimension of 256.
- This process includes the sum of word embeddings and positional embeddings, which represent the position of each token in the sequence.

2. Positional Encoding

- Positional information is added to the word embeddings so that the model takes into account the order of the tokens in the sequence.
- The length is the same as the input embeddings.

3. Encoder Layers

- These layers perform most of the feature extraction work.
- Each encoder block layer contains a multi-head attention layer and a feed-forward neural network layer.

- The multi-head attention allows the model to focus on different parts of the sequence simultaneously, while the feed-forward network performs additional transformations on the extracted features.
- In this work, 8 and 12 encoder layers were used, allowing significant depth in processing and feature extraction.

7. Training

For the training process the dataset was divided into 90% training data and 10% test data. Four models were trained following the BERT architecture varying the hyperparameters as follows:

TABLE II

TABLE OF MODELS

Model Name	Encoder layer	Cantidad de tokens	Epocas
Model 1	12	256	10
Model 2	8	512	10
Model 3	12	512	7
Model 4	12	512	15

The metrics used to measure performance and compare models were precision, Recall and F1. However, metrics such as ROC curves were used to make the decision to select the best model.

Metrics such as ROC curves show the relationship between true positive rate (TPR) and false positive rate (FPR) at different decision thresholds. TPR, also known as sensitivity, measures the proportion of true positives that are correctly identified, while FPR measures the proportion of false positives. In addition, we used precision-recall curves, a graphical tool used to evaluate the performance of a binary classification model, which provides detailed information on the relationship between precision and recall at different decision thresholds.

8. Results and Analysis

8.1 Experiment

For all training, the loss function BCEWithLogitsLoss was used, with Adam as optimizer and a learning rate of 2e-5. The following table shows the metrics obtained for the four trained models.

TABLE III
TABLE OF METRICS BY MODEL

Model	Epochs	Encoder layers	Tokens	Precision	Recall	F1 Score
Model 1	10	12	256	0.81	0.8	0.8
Model 2	10	8	512	0.81	0.81	0.81
Model 3	7	12	512	0.82	0.81	0.82
Model 4	15	12	512	0.8	0.8	0.8

Figure 5 shows the ROC and the precision-recall curve for model 1. In both graphs it can be seen how the curves have great disparity. Furthermore, in the precision-recall graph it is observed that the curve for the positive class is quite poor.

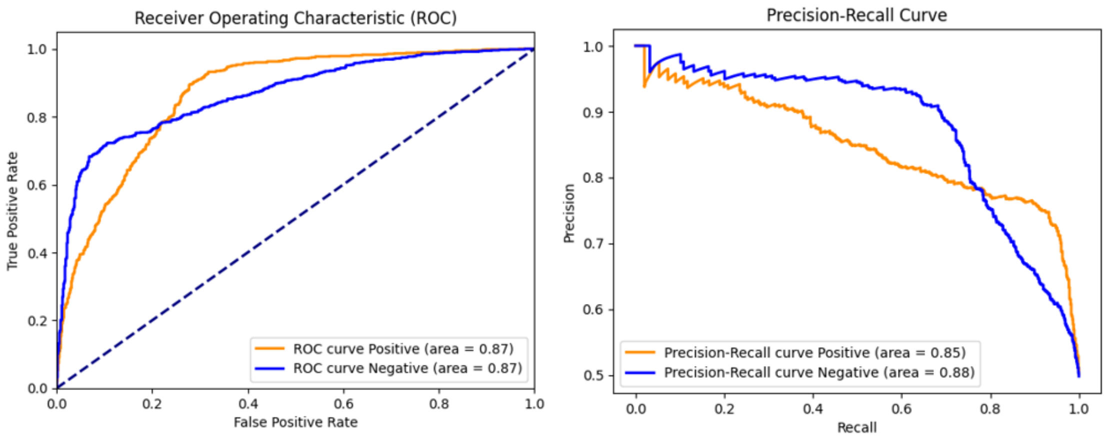

Figure 5. ROC and precisión-recall curves model 1.

Figure 6 shows the ROC and the precision-recall curve for model 2. Here it can be seen that by increasing the token window, the precision-recall curve improves and, in particular, the positive class curve improves drastically. The ROC curves also improve slightly with the change to 512 tokens going from 87% in the previous experiment to 89%.

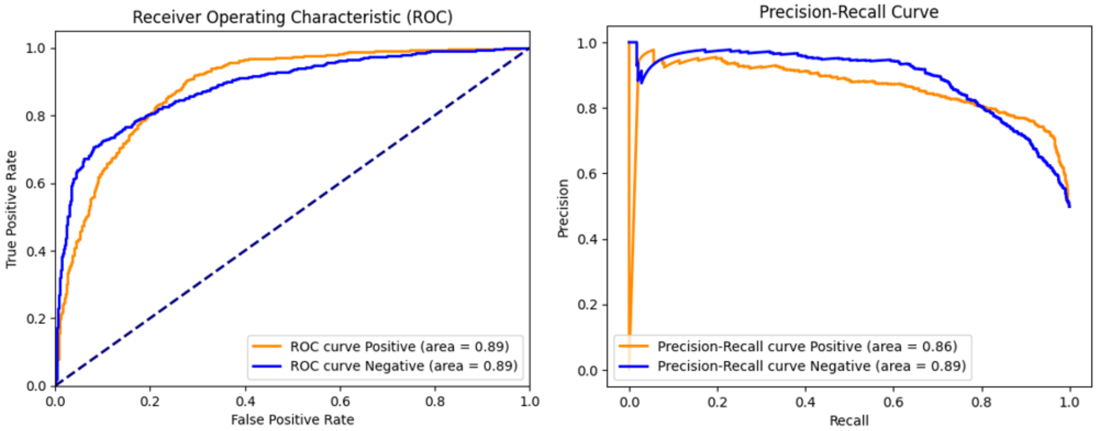

Figure 6. ROC and precisión-recall curves model 2.

Figure 7 shows the ROC and the precision-recall curve for model 3. For this experiment, the encoder layers were increased and again showed slight improvements in the precision-recall graph for the positive class, going from 86% in the previous one. Experiment up to 87%.

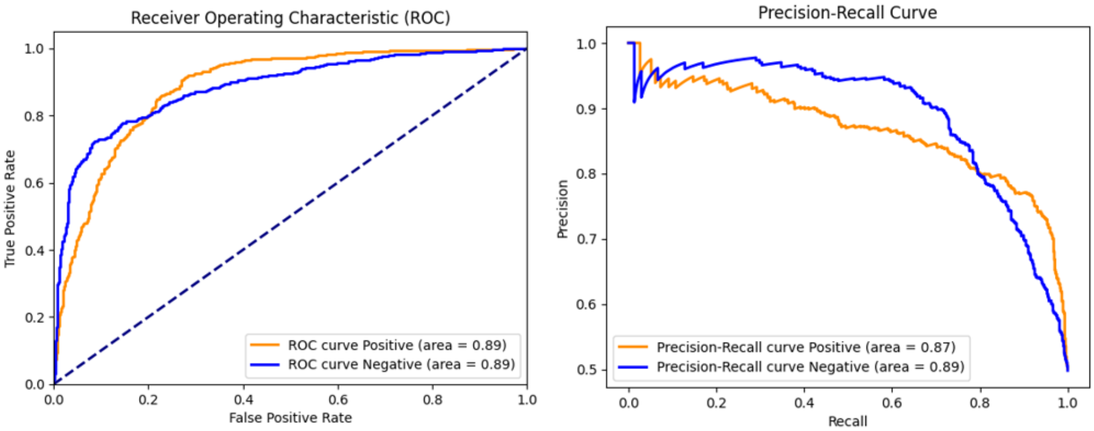

Figure 7. ROC and precisión-recall curves model 3.

Figure 8 shows the ROC and the precision-recall curve for model 4. With the experience of the previous model, it was decided to maintain the token size configurations at 512 and 12 encoder layers and thus double the epochs. When comparing Figure 8 and Figure 7, it can be seen that as the epochs increase, the precision-recall curves improve drastically.

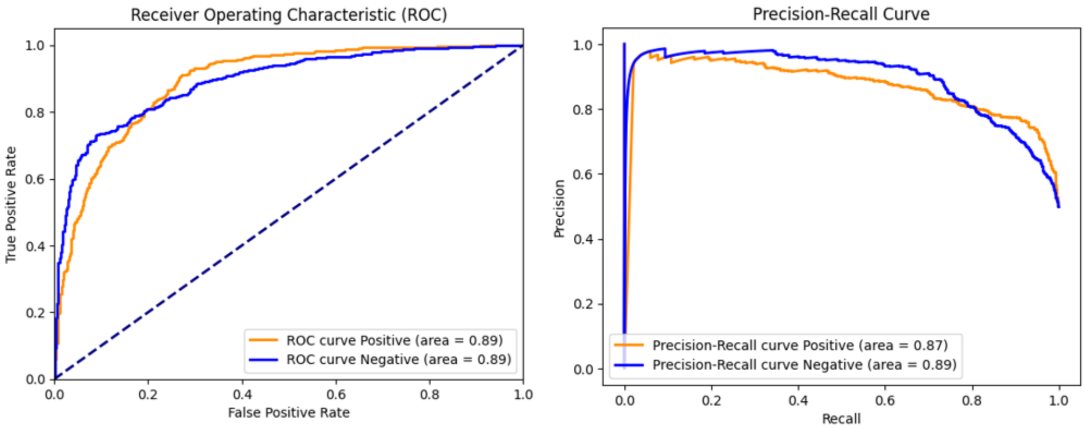

Figure 8. ROC and precisión-recall curves model 4.

8.2 Best model

The ROC and precision-recall curves were used to select the best model. More importantly, the four models trained in this thesis were compared with the BERT model used in the hugging phase to classify the dataset. For this purpose, 213 messages were downloaded and classification was performed using the different models. In this exercise, it is found that the model that comes closest to the pre-trained model is model number 4 with an accuracy of 72%.

TABLE IV
BEST MODEL
Metric Value
Epochs 15
Encoder layers 12
Tokens 512
Precision 0.8
Recall 0.8
F1 Score 0.8

9. Conclusions

We managed to train a model using the BERT architecture, achieving quite reasonable results using 30,000 news articles, reaching an accuracy of 72% compared to a pretrained model. This success is attributed to the observation that training for more epochs, using longer token sequences, and increasing the number of encoding blocks in the transformer significantly improves the model's quality.

Balancing the dataset by removing the neutral class and equalizing positive and negative samples significantly improved the model's ability to discern between polarized sentiments, highlighting the importance of preprocessing and proper data selection in the final model performance.

Although promising results were achieved in the training and validation phase, practical implementation of the model would face additional challenges such as scalability, computational efficiency, and integration with real-time systems. Exploring solutions such as deployment with Flask could facilitate the adoption and application of the model in various environments.

10. Future works

As it has been proved during the development of this project, building a BERT model with the configuration of 512 allowed tokens and 12 encoder layers is generating good models, it would be interesting to train for more epochs and compare the results obtained with the results offered in this project.

As a future work we would like to deploy the models with the help of FastAPI and create a free tool for the community to use.

11. Acknowledgment

We thank the engineering faculty for providing this specialization in analytics and data science because it has given us quite up-to-date knowledge in technology with which we can design solutions in analytics, data science and artificial intelligence to solve problems quickly and accurately or strengthen tools already created.

12. References

[1] E. Puertas, J. C. Martinez-Santos, and P. Andrés Pertuz-Duran, "Presidential preferences in Colombia through Sentiment Analysis," in 2022 IEEE ANDESCON, Barranquilla, Colombia, 2022, pp. 1–5, doi: 10.1109/ANDESCON56260.2022.9989700.

[2] A. Kumar and B. Bhushan, "AI Driven Sentiment Analysis for Social Media Data," in 2023 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS), Greater Noida, India, 2023, pp. 1201–1206, doi: 10.1109/ICCCIS60361.2023.10425434.

[3] R. Chauhan, A. Gusain, P. Kumar, C. Bhatt, and I. Uniyal, "Fine Grained Sentiment Analysis using Machine Learning and Deep Learning," in 2023 International Conference on Sustainable Emerging Innovations in Engineering and Technology (ICSEIET), Ghaziabad, India, 2023, pp. 423–427, doi: 10.1109/ICSEIET58677.2023.10303481.

[4] "gnews," PyPI, [Online]. Available: https://pypi.org/project/gnews/. [Accessed: 18-Jun-2024].

[5] "Spanish News Classification," Kaggle, [Online]. Available: https://www.kaggle.com/datasets/kevinmorgado/spanish-news-classification. [Accessed: 18-Jun-2024].

[6] "distilbert-base-multilingual-cased-sentiments-student," Hugging Face, [Online]. Available: https://huggingface.co/lxyuan/distilbert-base-multilingual-cased-sentiments-student. [Accessed: 18-Jun-2024].

[7] "Chapter 6: Transformers and Pretrained Models," Hugging Face, [Online]. Available: https://huggingface.co/learn/nlp-course/en/chapter6/6?fw=pt. [Accessed: 18-Jun-2024].

[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is All You Need," in Advances in Neural Information Processing Systems (NeurIPS), Long Beach, CA, USA, 2017, pp. 5998–6008, doi: 10.48550/arXiv.1706.03762.

Article IEEE conference

https://www.overleaf.com/read/whqfsgjjcqgy#40673d