Classification in Journalistic Texts with Transformers

1st Igor D. R. Silva
Universidade Federal de Pernambuco
Recife, Brasil
idrs@cin.ufpe.br

2nd Lucas N. Brandão
Universidade Federal de Pernambuco
Recife, Brasil
lnb@cin.ufpe.br

3rd Matheus I. G. Batista
Universidade Federal de Pernambuco
Recife, Brasil
migb@cin.ufpe.br

***Abstract*—This project aimed to perform textual classification of a series of news items taken from a public dataset, using pre-trained neural network NLP models, including those based on the Transformer architecture. The objective was to demonstrate the efficiency of this approach for this task and verify its performance.**

***Index Terms*—NLP, text-classification, Transformer.**

I. Introduction

Language has long been the subject of many scholars, always proving to be a challenging and complex topic, just like its speakers. According to Rita Mae Brown, language is the road of a culture. It tells us where people come from and where they are going. The study of modern linguistics and artificial intelligence emerged together in 1958 when linguist Noam Chomsky proposed his theory based on syntactic models. The field of Natural Language Processing (NLP) has grown significantly since then, with the advent of Transformers [**b5**], neural network architectures based on attention mechanisms, in 2017 with the release of the article "Attention is all you need." This milestone, resulting from several years of research and contributions from various researchers in the field, facilitated machine processing, understanding of context, subject, and main message in a text.

II. Objective

With this in mind, this project aimed to explore and use the state-of-the-art Transformer model in a database of journalistic texts. The objective was to achieve text classification using this model to align with the pre-established labels for each text in the dataset and verify the effectiveness of the model.

III. Justification

Text classification is a fundamental area in natural language processing. Through it, it becomes possible to understand the category or classification of a message, enabling improvements in human-machine interaction, monitoring brand reputation on social media, understanding the market, analyzing customer feedback, identifying fake news, and much more.

Transformers networks are essential in text classification due to their ability to capture complex contexts and relevant information in text sequences, allowing models to efficiently learn to assign categories based on available data. They have made it possible to achieve high-performance results in this task and many other natural language processing applications. Their creation was undoubtedly one of the major milestones in the field, revolutionizing the way text classification was done.

IV. Methodology

In this project, a text classification experiment was carried out in a database containing journalistic texts. For this, the deep learning technique was used: Transformers Models.

A. Dataset

The database used in the project was "ag news" [**b1**], which is a dataset widely used for text classification and text mining tasks. It contains news collected from various web sources and is commonly used to train and evaluate machine learning and natural language processing (NLP) models for classifying text into four main news categories.

The four categories of "ag news" are: 'World' (news related to global events and happenings) 'Sports' (news related to sports and sporting events), 'Business' (news related to business, economics and finance) and ' Sic/Tech' (news related to science and technology). Each news example in the dataset is labeled with one of these categories.

The dataset is divided into 120000 cases for training, 30000 for each category, and 7600 for training. For our project, we took a smaller portion of the dataset, using 1000 examples for training and 1000 examples for testing.

B. The Transformers Model

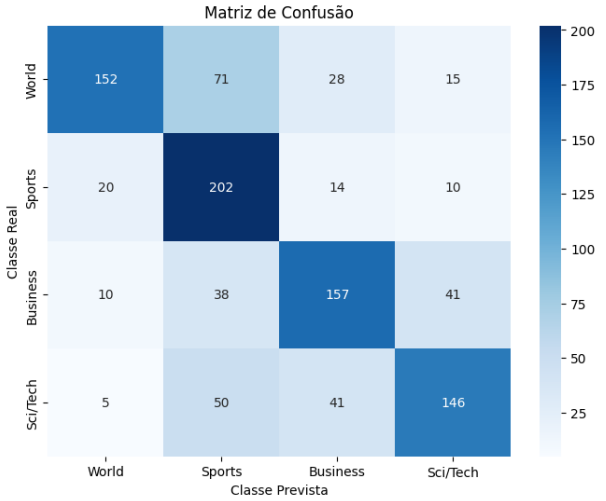
For the project, the "twitter-roberta-base-sentiment-lates" model [**b3**] [**b4**] was used, which is a pre-trained variation of BERT (Bidirectional Encoder Representations from Transformers). This model was tuned for sentiment analysis with the TweetEval benchmark. With it, it is possible to "tokenize" the text examples in the dataset.

We used Google Colab to carry out our project and, to train our model, taking into account the size of the portion of the dataset we decided to use, we chose a number of epochs equal to 50. As the model was being trained, we were able to see the reduction in training loss rate from 1.285 to just 0.002.

C. Results

To check the results obtained in our experiment, we used precision and accuracy as metrics. In the end, we obtain a precision of 0.68 and an accuracy of 0.66.

In addition to precision and accuracy, we also made a confusion matrix of our model:



REFERENCES

[1] X. Zhang, J. J. Zhao, and Y. LeCun, "Character-level Convolutional Networks for Text Classification," in Proceedings of NIPS, 2015.

[2] D. Loureiro, F. Barbieri, L. Neves, L. Espinosa Anke, and J. Camacho-collados, "TimeLMs: Diachronic Language Models from Twitter," in Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Dublin, Ireland, May 2022, pp. 251-260, Association for Computational Linguistics.

[3] J. Camacho-collados, K. Rezaee, T. Riahi, A. Ushio, D. Loureiro, D. Antypas, J. Boisson, L. Espinosa Anke, F. Liu, E. Martínez Cámara, and others, "TweetNLP: Cutting-Edge Natural Language Processing for Social Media," in Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Abu Dhabi, UAE, Dec. 2022, pp. 38-49, Association for Computational Linguistics.

[4] P. Liu, X. Qiu, and X. Huang, "Recurrent Neural Network for Text Classification with Multi-Task Learning," in Fudan University, School of Computer Science, Fudan University, 825 Zhangheng Road, Shanghai, China, May. 2016, Shanghai Key Laboratory of Intelligent Information Processing.

[5] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Advances in neural information processing systems, pages 5998–6008.

[6] AmirAli Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018b. Multimodal language analysis in the wild: CMUMOSEI dataset and interpretable dynamic fusion graph. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2236–2246, Melbourne, Australia. Association for Computational Linguistics.