Text Classification of News Using Deep Learning and Natural Language Processing Models Based on Transformers for Brazilian Portuguese

Isabel Nadine de SANTANA
Manufacturing and Technology Integrated Campus – SENAI CIMATEC
Salvador, Bahia, Brazil

Raphael Souza de OLIVEIRA
Manufacturing and Technology Integrated Campus – SENAI CIMATEC
Salvador, Bahia, Brazil

Erick Giovani Sperandio NASCIMENTO
Manufacturing and Technology Integrated Campus – SENAI CIMATEC
Salvador, Bahia, Brazil

ABSTRACT

This work proposes the use of a fine-tuned Transformer-based Natural Language Processing (NLP) model called BERTimbau to generate the word embeddings from texts published in a Brazilian newspaper, to create a robust NLP model to classify news in Portuguese, a task that is costly for humans to perform for big amounts of data. To assess this approach, besides the generation of the embeddings by the fine-tuned BERTimbau, a comparative analysis was conducted using the Word2Vec technique. The first step of the work was to rearrange the news from nineteen to ten categories to reduce the existence of class imbalance in the corpus, using the K-means and TF-IDF techniques. In the Word2Vec step, the CBOW and Skip-gram architectures were applied. In BERTimbau and Word2Vec steps, the Doc2Vec method was used to represent each news as a unique embedding, generating a document embedding for each news. The metrics accuracy, weighted accuracy, precision, recall, F1-Score, AUC ROC and AUC PRC were applied to evaluate the results. It was noticed that the fine-tuned BERTimbau captured distinctions in the texts of the different categories, showing that the classification model based on the fine-tuned BERTimbau has a superior performance than the other explored techniques.

Keywords: Deep Learning, NLP, BERT, BERTimbau, Transformers, Word2Vec.

1. INTRODUCTION

Understanding human language has been one of the greatest challenges of Artificial Intelligence. The Natural Language Processing (NLP) area faces big challenges such as the context understanding, sentiment analysis or figurative language interpretation. Recently, NLP has been boosted with deep learning, performing significant advances and transforming the interaction between humans and machines.

Language models are usually pre-trained to process large sets of general-scope text and audio data of a given language (also known as the linguistic corpus), and serve as the general basis for a diverse range of applications, becoming foundation models, which are then leveraged for solving problems on specific tasks through transfer learning. However, training these foundation models is generally extremely expensive, requiring a huge computational effort. Recently, the emergence of the Transformer-based neural network architecture and the attention mechanism [1] has decreased the required computational resources to perform such tasks. This has made it feasible to train foundation models on large volumes of general scope text and audio data, becoming them the most powerful.

The most part of pre-trained models based on Transformers publicly available was trained in English language, but there are some models in other languages, such as Google's Multilingual BERT, Bidirectional Encoder Representations from Transformers [2], which encompasses 104 languages including Portuguese. However, these models are not trained as a big and representative linguistic corpus of the specific language, opening a gap in the foundation models construction to other languages. In this context, in 2020, the NeuralMind company released a Bert model trained in Brazilian Portuguese named BERTimbau [3]. This initiative opened new possibilities for research in the field of NLP in Portuguese.

One of the main reasons to use BERT is due to the fact that its self-attention mechanism can learn the semantic relationship between words. Also, the model does not process the words input in sequence, opposed to models based in Recurrent Neural Networks (RNN). Furthermore, this model is based on unsupervised learning, which means that they are trained on a corpus with unlabeled data. Thus, this study compares different types of NLP techniques to a specific task of classification in text categories of news written in Brazilian Portuguese from the regional newspaper "A Tribuna" from Vitória/ES, and proposes a methodology for Brazilian Portuguese text classification using BERTimbau as a base, then specializing it as for this case study. News classification tasks become costly when we handle large amounts of data with different textual structures and contexts.

The methodology to the classification demonstrated in this paper was performed through a Machine Learning model fed by numerical vectors that represent the words, named word embeddings (WE) [4]. The experiment also did a comparative study between the WE made from BERTimbau and those made from an older but widely used NLP technique called Word2Vec [5]. This paper is organized as follows: Section 1 presents the Introduction. Section 2 presents Theoretical Reference. Section

ISBN: 978-1-950492-66-4
ISSN: 2771-0947

https://doi.org/10.54808/WMSCI2022.03.134

3 presents the Methodology. Section 4 presents the Results and Discussion and finally Section 5 presents the Conclusion.

2. THEORETICAL REFERENCE

One of the biggest challenges in NLP is training language models on large linguistic corpora. Before the advent of WE, a widely used technique was Bag of Words (BoW). It transforms the analyzed corpus into a sparse matrix. The rows of this matrix represent the number of documents to be analyzed, and the columns represent the presence or absence of a given token (word) in the document. A similar technique, used in this study as a baseline, is TF-IDF (Term-Frequency - Inverse Document Frequency). It calculates the frequency of a term in a document divided by the inverse of its frequency in all other documents in the corpus, thus indicating the importance of a word for a document [6]. BoW and TF-IDF are techniques that fail to capture context and semantics of words and add drawbacks such as high sparsity of word representations, since the size of each vector is equal to the number of distinct words in the corpus, but they carry very little relevant information.

In 2011, the first techniques for building word representation in a dense vector space, the WE [4] , appeared. The generation of a WE considers words that are close to the target word, in order to capture the context and semantics. In this way, representations of similar words occupy dense and close vector spaces, and therefore, it becomes possible to calculate the degree of similarity of words through similarity analysis techniques, such as cosine similarity [7]. In 2013, Google presented the unsupervised learning technique Word2Vec [5]. For the WE generation, Word2Vec has the CBOW (Continuous Bag of Word) and Skip-gram approaches. CBOW aims to predict a target word from a number of context words, located before and after that word. The number of words is defined in the "window size". Skip-gram, on the other hand, tries to predict a number of context words from an input target word. The CBOW technique has the advantage of being simpler and its training time is shorter than Skip-gram. In addition, the first technique captures syntactic relations between words better than the latter. However, the Skip-gram technique can better capture semantic relationships between words [3]. An example of a learned relationship between WEs is realized by subtracting two vectors and adding another WE to the result. For example, Paris - France + Italy = Rome. [5].

Word2Vec builds context-free WE, that is, every word in your vocabulary is represented by the same dense vector regardless of what a word means in a given sentence. So its biggest drawback is that it cannot differentiate words that are homographs. Another weakness is that the Word2Vec-based model cannot create WE from words that are not present in its vocabulary. WE were commonly used as embedding layers in RNNs such as LSTM and GRU, which require higher computational power to train. Because of their intrinsically recurrent architecture, they cannot be trained in parallel, which reduces the gains in training these networks on supercomputers.

At that same time, the concept of attention mechanism [1] and the "Sequence-to-Sequence" approach for solving some problems, such as translation between languages, surged. In 2018, BERT emerged as a model based on the neural network architecture called Transformers, which promoted a major revolution in the field of NLP. The Transformer has an Encoder-Decoder architecture. The model receives as input a sequence of words, which will be encoded in WE, and later they will be decoded in words as outputs, for a given NLP task (e.g. text translation). The original proposed architecture is composed of a stack of six encoders and six decoders that are identical, but do not share the same weights. Each encoder has a MultiHead Self-Attection layer, and a feed-forward neural network. The Positional Encoding is present at the input of the model, being responsible for assigning the position of each word in a sentence, thus preserving the order of the words.

The Self-Attention layer is responsible for generating a WE for each word. The WE of a word is based on the weighted sum of all the other words in the sequence, where those that are important to the target word will receive greater importance. The idea of this mechanism is to signal that the meaning of the target word can be better explained by looking at the other words in the sequence, storing contextual information in each representation. The feed-forward takes as input the output produced by the attention mechanism and sends it to the next encoder. The structure of the decoder is similar to the encoder structure, but with the addition of an intermediate attention layer that signals the decoder to "pay attention" to the most relevant words in the sentence. Despite the existence of other predecessor models based on Transformers, BERT was the first to capture the context to the left and right of the target word in a bidirectional way. This makes it easier for the model to find relationships among words.

This new WE building process, unlike RNN architectures, does not depend on the sequence of the words. So multiple sentences can be processed in parallel. Such independence made it possible to train these models on large corpora, making these models very powerful. As a consequence, we are now able to work with these large pre-trained models taking advantage of transfer learning in any NLP task. Regarding BERT, this model has a stack of twelve encoders and twelve attention mechanisms for the Base version and 24 encoders and 16 attention mechanisms for the Large version. The feed-forward layers are 768 for BERT-Base and 1024 for BERT-Large. In addition, they have 110 and 355 million parameters, respectively. BERT was trained on Wikipedia and the Book Corpus [2]. BERTimbau is the first BERT model that was trained on the Brazilian Portuguese BrWaC (Brazilian Web as Corpus) [3]. Its pre-trained model has Large and Base versions.

3. METHODOLOGY

The study was implemented following the steps illustrated in Figure 1. It was developed with the Python programming language (version 3.7.11). The K-means algorithm from Scikit Learn was used to plot the clusters. WE for Word2Vec were built with the open source library Gensim (version 3.6.0). In the fine-tuning step of BERT, the pre-trained BERTimbau model from the BERT-base version was used. The open source PyTorch library was used for GPU usage. BERTimbau was manipulated with the Transformers Hugging Face library (version 4.9.0), from the open-source community of pre-trained Deep Learning models called “Hugging Face” [8]. The model used for news classification was the XGBClassifier, from the XGBoost library (version 0.90), which uses the Gradient Boosting technique.

The data analysis was performed on a collection of news from the newspaper "A Tribuna", collected between the years 2004 and 2007. The database was labeled according to the newspaper sections, with a total of 42,123 instances and is distributed in nineteen unbalanced categories. About 77% of the data is concentrated in six categories, being “Economy” the largest category with 6,557 instances and the smallest category, “Everything to do”, with only 30 instances. In the data exploration step, it was noticed that most of the texts have between 100 and 450 words, but there is a significant amount of texts with 500 to 1000 words. The average is 485 words, with the smallest text containing 14 words and the longest 5,523 words.

In order to group the news in more general categories, the first step was to generate word clouds for each category to know the most cited words and try to extract some information about the subject and context of the existing categories. Before the clouds were generated, the text was cleaned by removing very frequent words, HTML tags, numbers, accentuation, punctuation, prepositions and articles. By observing the word clouds, it was possible to clearly identify the subject matter in some classes, while the distinction was not so clear in others.

To reduce the unbalance of the dataset and facilitate the regrouping process of the minority categories, the k-means and TF-IDF techniques were used in order to identify whether such categories could be incorporated into the larger ones or simply be discarded. The definition of the initial number of clusters was based on the elbow method using the inertia metric, resulting in six clusters. An aspect also taken into consideration was the fact that almost 80% of the instances were distributed in only six of the nineteen existing categories. After the clusters were defined, heat maps were plotted which indicated the similarity of the various categories with the six major clusters. Some minority categories that were similar to each other were merged into a new category. Others were incorporated into those that were similar and already existed. In this way, the initial nineteen classes were rearranged into only ten.

The next step was to build WE using Word2Vec with the CBOW and Skip-gram architectures. The goal was to feed a classification model, which served as a baseline for comparing classification results with WE built by Transformers in the following steps. The Word2Vec model from the Gensim API[1] was used to generate the WE of the words. The size of the output WE vector was 768 to be equivalent to the one built by BERT, which also produces an embedding of the same size. The window-size chosen for Skip-gram and CBOW was five words.

After the WE were generated, each news text had the amount of WE equivalent to the amount of words. Then, the Doc2Vec method was used to create only one WE per text. For each news story, a single WE was calculated by averaging the WE derived from each text. To train the classification model, the doc2vecs were divided into train and test in the proportion 80% and 20% respectively. The model chosen to be the baseline was Word2Vec with CBOW. This architecture was chosen for having the fastest processing, given the size of the dataset used in this study.

[1] API: Application Programming Interface

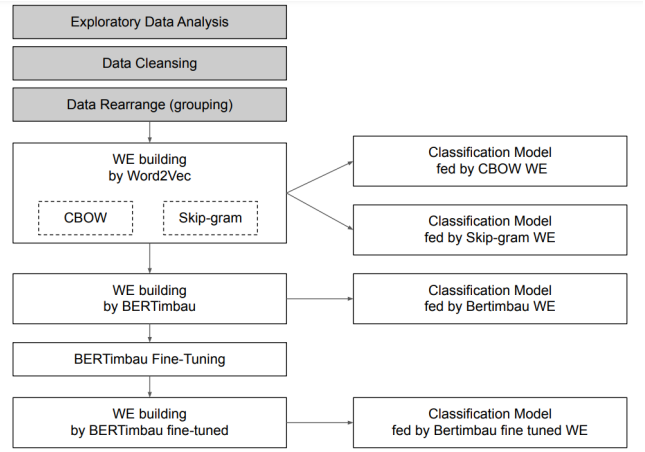

Figure 1. Steps of the methodology developed in this study

In order to discover the power of Transformers in capturing the news context, for this step, the pre-trained BERTimbau Base model was used. The BERT requires a specific input format to be fed. Mandatorily, the word sequences must have a fixed size of 512 words. Using the BERT Tokenizer, words were transformed into tokens, which represent BERT's vocabulary words. BERT's Tokenizer transforms subwords into tokens, enabling BERT to generate a representation of an unknown word. Then, the average of the WE of each sentence in a document was calculated to reduce to just one per sentence. To train the BERT classification model, the doc2vecs were divided into train and test in the proportion 80% and 20% respectively.

The next step was to perform fine-tuning of BERTimbau, which is the process of training part of the pre-trained model on a specific dataset. The BERTimbau (Bert-Base) model fine-tuning technique chosen was to train all layers of the model with the masked-language modeling (MLM) strategy. This approach consists of providing a sentence to the model with some masked input tokens and the output should be the same complete sentence. The model's attempts to guess the tokens makes BERT better understand the usage of the words in a specific context. Thus, at the end of training, the weights of the layers of its entire architecture are updated and adjusted according to the "A Tribuna" dataset. The indicators and metrics used to analyze the results were the Accuracy, Weighted Accuracy, Precision, Recall, F1-Score, Area Under the Curve - Receiver Operating Characteristic Curve (AUC ROC), and Area Under the Curve Precision Recall Curve (AUC PRC).

4. RESULTS AND DISCUSSION

The overall results obtained in classification are illustrated in Table 1. Regarding Word2Vec, the best results were obtained with the Skip-gram approach. The general accuracy obtained was only 57.8%. The weighted accuracy was 57%. The precision was 55%.

Figures 2 and 3 illustrate the classification reports for CBOW and Skip-gram with the metrics Accuracy, Revocation, F1-Score separated by categories. In Figure 3, where the best results are displayed, it is possible to highlight “Sports” and “Tech” as the categories that the model was most accurate with 67%. The “Regional” category obtained the lowest recall result, with a value of only 19%. The best F1-Score result was also for “Sports” with 68%. The model performed below 50% in all

metrics for the “Local News”, “International”, “Opinion” and “Regional” categories. The AUC PRC result for CBOW and Skip-Gram was only 53% and 55% respectively. The “Regional” obtained the curve with the smallest area in both approaches, highlighting the CBOW that obtained a value of only 22%. Finally, two plots are shown with the clusters resulting from the classification using CBOW in Figure 4 and Skip-gram in Figure 5. In both plots, although it is possible to see most of the groups they are not clearly separated.

The results achieved with the WE from BERTimbau were significantly better than those of Word2Vec, according to the metrics shown in column Original from Table 1.

Table 1. Comparative table of the obtained results

Word2Vec			BERTimbau	
Architecture	CBOW	Skip-gram	Original	Fine-tuned
Accuracy	0.557	0.578	0.811	0.877
W. Accuracy	0.550	0.570	0.810	0.880
Precision	0.531	0.554	0.799	0.873
Recall	0.485	0.511	0.774	0.853
F1-Score	0.496	0.524	0.784	0.862
AUC ROC	0.878	0.889	0.967	0.982
AUC PRC	0.528	0.552	0.830	0.897

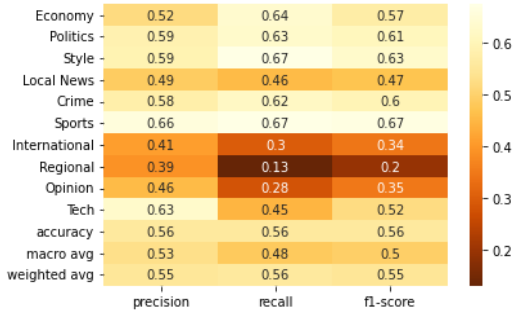

Figure 2. Metrics of the classification performed by the model fed with CBOW WE .

Analyzing the metrics of the original BERTimbau, there was an increase of about 25 percentage points in the metrics of accuracy, precision, revocation, and F1-score. In Figure 6, the metrics by category can be analyzed and it can be seen that the results are more balanced.

The best F1-score result is for the “Sports” category with 93% and soon after, for "Style" with 89%.

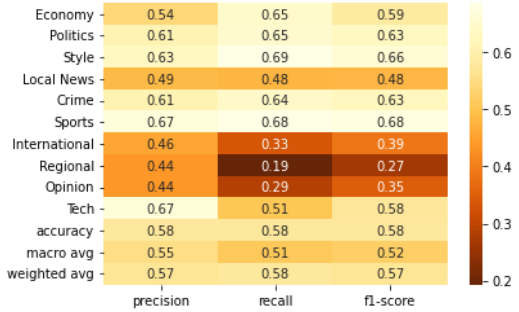

Figure 3. Metrics of the classification performed by the model fed with Skip-gram WE.

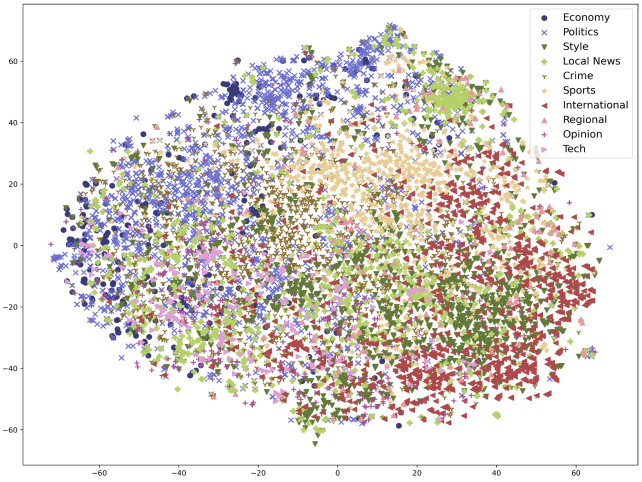

Figure 4. Resulting clusters from CBOW

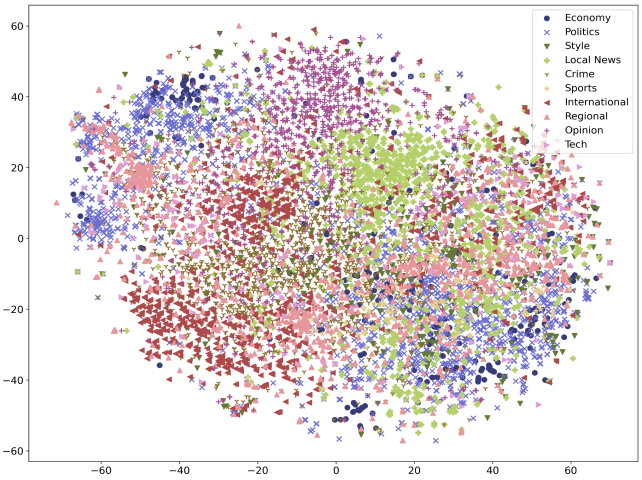

Figure 5. Resulting clusters from Skip-gram

Another relevant point was the increase in the PRC AUC that rose from 55% to 83% demonstrating that BERTimbau was able to capture the nuances of the different subjects, performing well in predicting even for minority categories. The K-means output illustrated in Figure 7 shows a better distinction of the clusters, however there is still not a good distinction between the "Regional", "Local News" and "International" clusters.

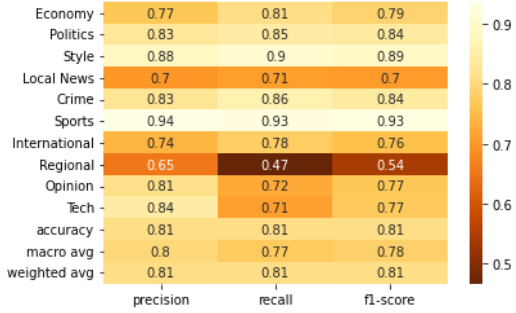

Figure 6. Metrics of the classification performed by the model fed with BERTimbau WE

The next step was to perform the BERTimbau fine-tuning. Perplexity [9] was the metric used to evaluate the fine-tuning of the model. It tries to show how likely the model is to be

confused when choosing a word. So the lower the Perplexity, the better the model. The value obtained at the end of the BERTimbau training was approximately 4.21. After that, the trained model was used to generate new WE for another dataset classification. With the new WE created, a better performance than the original BERTimbau was obtained. All the results of the applied metrics had a relevant increase, demonstrating that the fine-tuned BERTimbau absorbed characteristics of the texts of database context. Table 1 (fine-tuned column) also shows the results of the overall metrics obtained.

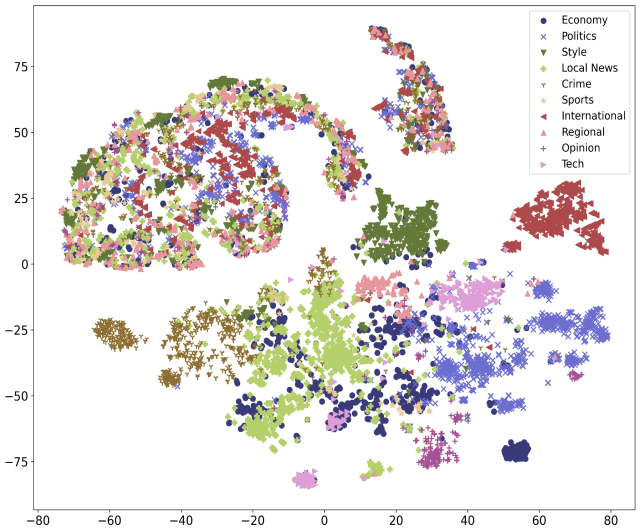

Figure 7. Resulting clusters from Bertimbau

The accuracy of the model reached approximately 87.7%. The weighted accuracy was 88%. The precision was approximately 87%, highlighting the 95% precision achieved for “Sports” (Figure 8). Another interesting result was the 82% precision for the “Regional” category, a considerable improvement over the 65% result obtained by the original BERTimbau. The highest recall was also for the “Sports” and the lowest was 68% for the "Regional" category. As in the original BERTimbau, the best F1-Score result was from the “Sports”. There was an improvement over the predecessor from 93% to 96%. The ROC curve increased from 97% to 98%. The AUC of all categories obtained a percentage greater than 97%. The “Sports” category obtained an AUC of 100%. The AUC PRC obtained a value of 89.7%. The categories with the smallest area were “Opinion” with 79% and “Regional” with 80%. The “Regional” category AUC PRC was only 56% using the original BERTimbau.

The plot shown in Figure 9, clearly demonstrates the ten clusters defined, and it is evident how much better the fine-tuned BERTimbau model performed compared to the other models developed in this study. When observing the obtained results, it can be seen that the percentages achieved with the techniques that used BERTimbau were considerably higher. The main reason for this is because the WE built by BERT are able to capture the context of words, unlike the WE built with Word2Vec, which result in context-free representations of words. This difference is evidenced when analyzing the results of the PRC curves for each category of news. The results for Word2Vec show curve areas below 50%, indicating the underfit of the model predictions for the minority categories.

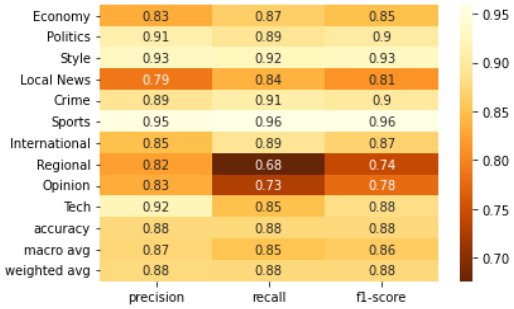

Figure 8. Metrics of the classification performed by the model fed with fine tuned BERTimbau WE

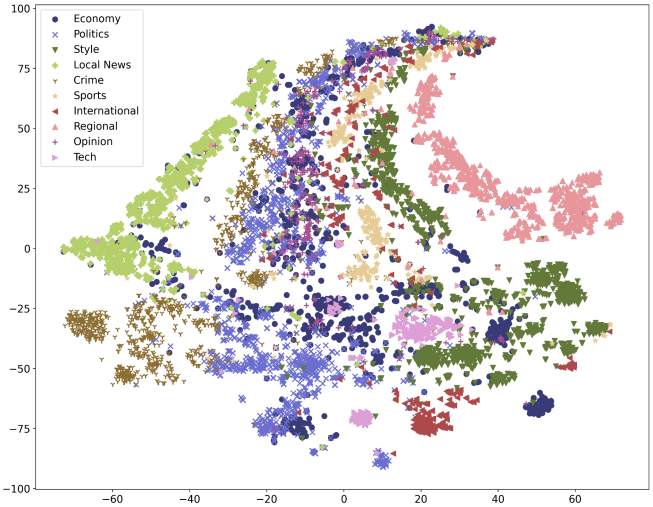

Figure 9. Resulting clusters from fine-tuned Bertimbau

Proceeding with the analysis of the results, it is perceived that when the fine-tuning of BERTimbau was performed, the model specialized itself in the language of the explored dataset. Such circumstance is evidenced in the PRC curve of the "Regional" category. In the original BERTimbau model, its area was 56%. With the fine-tuned BERTimbau, its result increases to 80%. Overall, after the fine-tuning, there was a significant improvement in the classification for all categories.

5. CONCLUSIONS

By performing the comparative study of WE building techniques, it is evident that the dense vector-based representations of words generated by the models based on the Transformers architecture are far superior in relation to their predecessors. The representations generated through the classic Bag of Words (BOW) technique, besides not capturing context, have the problem of the high dimensionality of the vectors defined by the size of the vocabulary, making the training computationally costly for large corpora. Word2Vec generates a WE for each word in your vocabulary. This feature makes this representation context-free . However, trying to reduce all the contexts of a word into a single vector representation did not prove to be a very efficient method. This limitation was perceived when analyzing the results. It was noticed that the model confused categories such as "Politics" and "Economy" whose contexts are distinct but share a similar vocabulary.

BERT, on the other hand, generates numerous representations of each word, depending on the context in which it is presented. This makes its WE context-dependent, so it is possible to capture the semantic nuances of different texts. When applying the original BERTimbau, a considerable improvement in the performance of the news classification task was noticed. After fine-tuning, there is an increase in the classification performance, especially when analyzing the result of the PRC curves for each category. During the fine-tuning process, a difficulty encountered was related to the fine-tuned BERTimbau Tokenizer. The token building time of the new words coming from "A Tribuna " dataset was much higher than the building time of a token from the main vocabulary. So, it became necessary to use the original BERTimbau Tokenizer to generate the new tokens. For future comparative studies, it is suggested to improve the Doc2Vec generation technique in order to choose the one that best preserves the news features. Another interesting suggestion is to use the BERTimbau-large version and other models based on Transformers architecture. Explainable AI can also be applied to better understand the performance of attention mechanisms.

Transformer-based language models have taken NLP to a level of excellence when it comes to human language interpretation. Several advances have been noticed in information seeking, speech recognition, Text-to-Speech and dialog systems. Models like BERT and its successors, such as RoBERTa [10] and GPT-3 [11] and its predecessors, GPT-1 and GPT-2, from Open AI, are being used by industry, commerce, health care, justice, but little has been produced for the Portuguese language. One of the main reasons is the little investment to build these models using large datasets in Portuguese.

The creation and specialization of language models for Portuguese, such as the fine-tuned BERTimbau, has captured language regionalisms, increasing its vocabulary and improving the classification of news. Journalists may benefit when they need to categorize their large collections into more general subjects. They may also have greater support in searching for topics of interest to them, better filtering the huge amount of unstructured information they are exposed to. The model can also be used to improve recommendation systems according to the type of information consumed by readers. It could also improve aggregation systems, by collecting several different textual sources, linking news by similar contexts. Finally, beyond the famous virtual assistants, Transformers-based models can contribute a lot to the improvement of people's daily lives.

6. ACKNOWLEDGEMENTS

The authors thank the Reference Center on Artificial Intelligence and Supercomputing Center for Industrial Innovation, both from SENAI CIMATEC, for the scientific, technical and computational resources support, as well as the NVIDIA/CIMATEC AI Joint Lab for the technical support. We would like to express our deeply felt gratitude to Professor Dr. Júnia Matos and Professor Dr. Bruno Menezes for the comprehensive and detailed peer-review of this document.

7. REFERENCES

[1] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I. 2017. **Attention Is All You Need.** doi: 10.48550/ARXIV.1706.03762.

[2] Devlin J, Chang M-W, Lee K, Toutanova K. 2018**. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.** doi: 10.48550/ARXIV.1810.04805.

[3] Souza F, Nogueira R, Lotufo R. 2020. **BERTimbau: Pretrained BERT Models for Brazilian Portuguese.** Intelligent Systems :403–417. doi: 10.1007/978-3-030-61377-8_28.

[4] Collobert R, Weston J, Bottou L, Karlen M, Kavukcuoglu K, Kuksa P. 2011. **Natural Language Processing (almost) from Scratch.** doi: 10.48550/ARXIV.1103.0398.

[5] Mikolov T, Chen K, Corrado G, Dean J. 2013. **Efficient Estimation of Word Representations in Vector Space.** doi: 10.48550/ARXIV.1301.3781.

[6] Stein RA, Silva ADB. 2016. **Análise assintótica de algoritmo para geração de matriz termo-documento contendo TF-IDF.** . doi: 10.6084/M9.FIGSHARE.4220691.V1.

[7] Sidorov G, Gelbukh A, Gómez-Adorno H, Pinto D. 2014. **Soft Similarity and Soft Cosine Measure: Similarity of Features in Vector Space Model**. **CyS.** doi: 10.13053/cys-18-3-2043.

[8] Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, Cistac P, Rault T, Louf R, Funtowicz M, Davison J, Shleifer S, von Platen P, Ma C, Jernite Y, Plu J, Xu C, Scao TL, Gugger S, Drame M, Lhoest Q, Rush AM. 2019. **HuggingFace's Transformers: State-of-the-art Natural Language Processing.** doi: 10.48550/ARXIV.1910.03771.

[9] Thiago Faleiros, Alneu De Andrade Lopes. 2016. **Modelos probabilístico de tópicos: desvendando o Latent Dirichlet Allocation.** doi: 10.13140/RG.2.1.3763.2880.

[10] Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov V. 2019. **RoBERTa: A Robustly Optimized BERT Pretraining Approach.** doi: 10.48550/ARXIV.1907.11692.

[11] Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, Agarwal S, Herbert-Voss A, Krueger G, Henighan T, Child R, Ramesh A, Ziegler DM, Wu J, Winter C, Hesse C, Chen M, Sigler E, Litwin M, Gray S, Chess B, Clark J, Berner C, McCandlish S, Radford A, Sutskever I, Amodei D. 2020. **Language Models are Few-Shot Learners.** doi: 10.48550/ARXIV.2005.14165.