


Addressed the problem of positive reviews using negative words in sentiment analysis with BERT

Bruno Iglesias^{1,†} and Josias Lima^{1,†*} 

¹ Creathus Institute of Technology, Amazonas, Brazil; bruno.iglesias@creathus.org.br and josias.lima@creathus.org.br

* Correspondence: josias.lima@creathus.org.br

† These authors contributed equally to this work.

Abstract: Context: Customer opinions about products are important for a company. However, there is still a challenge in automatically differentiating a positive review using negative words from a truly negative review. Objective: The implementation of an artificial intelligence model that can effectively make this distinction from customer reviews in Portuguese. Method: This work proposes SACReviews approach based on fine-tuning the BERT model for Sentiment Analysis (SA) of customer opinions. Result: The SACReviews obtained an accuracy of 96% in the analysis of opinions in Portuguese. Conclusions: The SACReviews is effective in analyzing the sentiment of positive reviews using negative words.

Keywords: Sentiment analysis; BERT; Portuguese language; Natural language processing; Product reviews; Machine learning; Opinion mining

Date of Submission: 25-05-2024

Date of Acceptance: 06-06-2024

Citation: Iglesias, B.; Lima, J. Addressed the problem of positive reviews using negative words in sentiment analysis with BERT. *Journal Not Specified* **2024**, *1*, 0. <https://doi.org/>

Received:

Revised:

Accepted:

Published:

Copyright: © 2024 by the authors. Submitted to *Journal Not Specified* for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

I. Introduction

Brazil is one of the countries where users spend the most time online [1]. According to recent industry calculations, Brazil will rank second among 20 countries worldwide in the development of retail e-commerce between 2023 and 2027, with a compound annual growth rate (CAGR) of 14.07%, with global retail e-commerce CAGR estimated at 11.16% during the same period [2]. Several factors are taken into consideration before making an online purchase, including the customer's opinion. Customer reviews have a significant effect on purchasing decisions [3], therefore, it is important for companies to understand how customers feel about their products or services by analyzing the opinions left on digital platforms. There is an area of research called Sentiment Analysis (SA) which analyzes a text to identify whether the emotional tone of the message is positive, negative or neutral [4]. However, not infrequently, a positive comment using negative words may occur in Brazilian Portuguese and is not always correctly identified as positive. An example of this type of comment would be o produto não tem nenhum defeito (in English, the product has no defects), where the words não, nenhum and defeito were used in a positive context, although they may appear negative when considered out of context. This can confuse models, as these words are commonly used in a negative context. Over the years, different researchers have verified different solutions for review sentiment analysis, e.g. Basa and Basarslan [5] analyzed the effectiveness of some classifiers using the IMDb database, where the Support Vector Machine (SVM) classifier obtained the highest accuracy with 90%. Chouikhi et al. [6] used the BERT model for sentiment analysis of reviews in the Arabic language. Gumiel et al. [7] evaluated some classifiers (SVM, Decision Tree, Random Forest, Logistic Regression and BERT-based models) for the analysis of Portuguese texts about diabetes. This article aims to build an artificial intelligence model that is capable of effectively analyzing the sentiments of positive comments (in Portuguese) using negative words. To Version May 31, 2024 submitted to *Journal Not Specified*<https://www.mdpi.com/journal/notspecified> this end, we chose to use the BERT model as a basis, doing the fine-tuning¹ using the *IMDb* database and the database

(*CReviews*), which we built using *ChatGPT*. The model fine-tuning was carried out in two parts, in the first, only the *IMDb* database was used and the accuracy of this model was tested using the *CReviews* database, in the second, the previous model was fine-tuned using the *CReviews* and verified its accuracy with 10% of *CReviews* that was not used during fine-tuning.

The results of the study indicated the viability of the *SACReviews* approach for analyzing sentiments of customer reviews in Portuguese, where there are positive reviews with negative words, since the approach achieved an accuracy of 96%. The main contributions of this work are:

- The *CReviews* database containing 159 positive reviews and 144 negative reviews, available at https://github.com/jlgomes/sentiment_analysis_with_bert;
- Definition of an approach for analyzing sentiments of customer reviews in Portuguese, also considering positive comments using negative words.
- Analysis of the effectiveness of the *SACReviews* approach in analyzing sentiments of positive reviews (in Portuguese) using negative words.

The remainder of this paper is organized as follows: Related works are presented in Section 2. Section 3 provides an overview of the materials and methods. Section 4 presents the results and discussion of the study. Finally, Section 5 shows the conclusion and future work.

II. Related Works

In recent years, many natural language processing (NLP) researchers have analyzed the effectiveness of the BERT model for problem solving. For example, Geetha and Renuka [8] compared Naïve Bayes, Long Short Term Memory (LSTM) and Support Vector Machine (SVM) classifiers with the BERT model for the sentiment analysis problem. The BERT model showed better performance in Accuracy (88.48%), Precision (88.09%), Recall (86.22%) and F1-Score (89.41%) compared to the other machine learning methods in experimental evaluation. Vásquez et al. [9] applied two Bert-based approaches to classifying reviews about Mexican tourist sites into five classes. The first approach consisted of tuning Beto, a Bert-like model pre-trained in Spanish. The second approach focused on combining Bert embeddings with TF-IDF weighted feature vectors. The results obtained using the standalone BERT model came first in the task. Lopes et al. [10] present an approach for Aspect Extraction based on pre-trained Multilingual BERT (Google) and Portuguese (BERTimbau) models. Experiments show that Aspect Extraction based on pre-trained BERT for Portuguese achieved Balanced Accuracy of up to 93% on a corpus of reviews about the hosting industry. Chouikhi et al. [6] propose the integration of an Arabic BERT tokenizer instead of a basic BERT tokenizer for sentiment analysis of reviews in the Arabic language, with an experimental study proving the efficiency of the proposed approach in terms of quality and classification accuracy compared to Arabic models BERT and AraBERT. Gumiel et al. [7] provide a database and a study comparing models for the Sentiment Analysis task in Portuguese texts about Diabetes Mellitus. The database contains 1,290 posts retrieved from online health community forums in Portuguese and annotated by two annotators according to 3 sentiment categories (Positive, Neutral and Negative). The study evaluated traditional machine learning classifiers (Support Vector Machine, Decision Tree, Random Forest and Logistic Regression) and state-of-the-art (BERT-based models), where the BERTimbau model obtained the best result for the Precision (83.12%), Recall (82.67%) and F1-Score (82.53%) metrics.

¹ Fine-tuning is the process of taking a pre-trained machine learning model and training it further on a new dataset.

Xu et al. [11] build a dataset for Review Reading Comprehension (RRC) called ReviewRC based on a popular benchmark for aspect-based sentiment analysis. They explored a new post-training approach on the popular language model BERT to improve the performance of fine-tuning BERT for RRC. To show the generality of the approach, the proposed post-training was also applied to some other review-based tasks, such as aspect extraction and aspect sentiment classification in aspect-based sentiment analysis. The experimental results demonstrated that the proposed post-training is highly effective.

Our article differs from related works because it is focused on using the BERT model in the problem of analyzing sentiments of positive (in Portuguese) reviews using negative words, which are often misclassified as negative.

III. Materials and Methods

3.1 Proposal

SAC Reviews approach (originating from the expression *Sentiment Analysis Customer Reviews*) was designed to automatically identify customer emotion in their reviews in Portuguese using machine learning, taking into account positive reviews using negative words and not only positive reviews with positive words. It has four main processes: (1)

Using the BERT model, (2) Data preparation, (3) Fine-tuning execution, and (4) Review classification. These processes are presented in Figure 1.

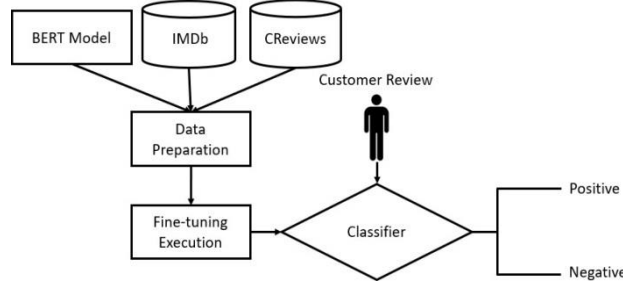


Figure 1. Approach SACReviews

Below, each step that makes up the processes presented in Figure 1 will be detailed.

3.1.1 Using the BERT Model

BERT model is a neural network-based technique for natural language processing (NLP) pre-training developed by Google, it has the ability to understand bidirectional context in a sentence, which means it can understand the meaning of a word in relation to the surrounding words [12]. This improvement in contextual understanding has led to excellent performance on several natural language processing tasks, which is why we chose to use this model as the basis for our SACReviews approach, where we will do the fine-tuning of the BERT model.

3.1.2 Data Preparation

A search was carried out in the technical literature and the database IMDb was found, which is widely used by artificial intelligence (AI) researchers [5], which is why it was chosen. It contains 49,444 film reviews written in Portuguese, including 24,679 positive reviews and 24,765 negative reviews, both with equal distribution by class. As a complement to this database, it was decided to create a new database called CReviews where it has positive reviews with negative words and negative reviews, for this, we chose to use ChatGPT to generate these customer reviews, where the base has 159 positive reviews (of which 122 are positive comments without negative words and 37 are positive comments with negative words) and 144 negative reviews, further details are in the Table 1. To generate reviews, we asked ChatGPT to generate a list of positive reviews, then a list of positive reviews containing negative words, and then a list of negative reviews. The current number of reviews occurred after we sorted the reviews alphabetically and carried out a manual check

Table 1. CReviews database details

Reviews without negative words	Quantity	Positive	Number of tokens	Number of lemmata	Number of 1568 negative tokens	Average review length	Standard deviation of review length
	122		1590		0	13.033	4.652
Positive with negative words	37		250	250	45	6.757	1.403
Negative	144		1664	1636	134	11.556	4.638

to remove comments that were too close, for greater reliability of the database. All reviews from each of the two databases were labeled as positive or not, according to Equation 1.

$$Revision(x) = \begin{cases} 1, & \text{if } x \text{ is a positive opinion} \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

BERT model is different from directional models, which read the text input sequentially from left to right or right to left, the transformer encoder reads the entire sequence of words at once [6]. To do this, the text will be tokenized by the *BertTokenizer* which will leave the tokenized text in the format that the *BERT* model expects.

3.1.3 Fine-tuning Execution

SACReviews sentiment classification model was fine-tuned in two phases, with the first phase involving fine-tuning the BERT language model with the IMDb database and the second phase involving fine-tuning the previous model with the CReviews database. Assessing model performance is a fundamental step in machine learning problem solving. Normally, this evaluation is done empirically, in which the data is divided into one or more fine-tuning sets and one or more test sets. The objective is to simulate the model when subjected to data never seen before. Thus, the model is built with fine-tuning examples and only has access to test examples at the time of performance evaluation. During fine-tuning, the database was split at 70% for fine-tuning, 20% for validation and 10% for testing. The hyperparameters adopted for this research were:

- Max length = 512;
- Batch size = 16;
- Optimizer = Adam;
- Epochs = 11; 147
- Steps per epochs = 200;
- Epochs validation samples = 50;
- Requires grad = False;
- Loss func = CrossEntropy.

To evaluate the tests, some performance measures are commonly used, in order to evaluate the effectiveness of the models, including *accuracy*, *precision*, *recall* and *F1-Score* [5,6,12]. In this work, the data represents customer reviews that can be positive or negative.

So these are the two classes of interest.

3.1.4 Classification of Reviews

In this process, the *SACReviews* approach created in the fine-tuning execution process will be used. It will receive as input a list of customer opinions and return whether the opinion is positive or negative, this will help to understand how customers feel about using the product or service.

IV. Results and Discussion

The results of this study will be shown for each of the models, starting with the BERT model, then the IMDb model, and finally the result for the SACReviews model. More information (e.g. precision micro, precision macro, precision weighted and precision binary) is available at

https://github.com/jlgomes/sentiment_analysis_with_bert. For the BERT model, the accuracy value for fine-tuning and validation will not be shown, as we are only using the already trained model. The accuracy of the BERT model tested on the ChatGPT database was 44% and the confusion matrix values are presented in the first row of Table 2.

Table 2. Confusion Matrices of Different Models

	TN	TP	FN	FP	Accuracy	F1-Score
BERT model tested on ChatGPT base	132	3	156	12	44%	30%
IMDb model in test stage	136	85	0	66	77%	77%
IMDb model tested on ChatGPT base	144	90	69	0	77%	76%
SACReviews model in test stage	69	75	2	4	96%	95%

Figure 2 shows the accuracy result for the *IMDb* model, where we can see that the accuracy value of the validation stage reached 80%. Testing the *IMDb model* on the *ChatGPT* database it achieved 77%, however, it placed 26 (70.3%) of 37 of the positive comments using negative words as negative comments. The values of the confusion matrix of the test stage are presented in the third row of Table 2.

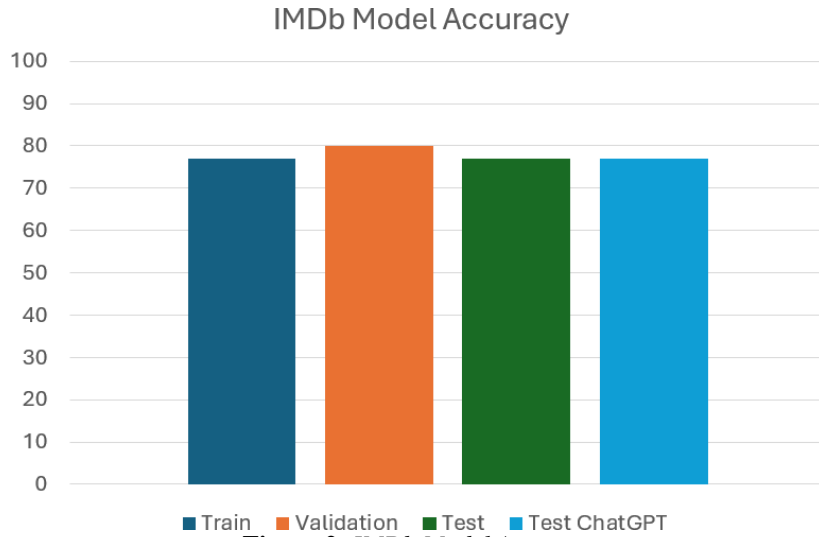


Figure 2. IMDb Model Accuracy

Figure 3 presents the accuracy for the SACReviews approach, where in fine-tuning it obtained an accuracy of 90%, in validation 93% and in the testing stage the accuracy was 96%. The confusion matrix from the SACReviews model test is shown in the fourth row of the table 2. One of the positive comments with negative words that SACReviews miscategorized was: Não vi nenhum defeito (in English, I didn't see any defects). However, we can mention three reviews that were correctly categorized: Não tenho nada para reclamar, serviço impecável (in English, I have nothing to complain about, impeccable service), Não vi nada de errado, recomendo totalmente (in English, I didn't see anything wrong, I totally recommend it) and Não tive problemas com este produto, funciona perfeitamente (in English, I had no problems with this product, it works perfectly). Given these results, we can infer that our approach is effective in analyzing sentiments of reviews in Portuguese, where negative words can be used for positive reviews. Accuracy is the fraction of correctly predicted components. In general, the greater the accuracy, the better the model. And this metric is recommended for balanced databases as in our case, however, we chose to also show other metrics. Figure 4 shows the test precision of the BERT, IMDb and SACReviews models, where we can see that SACReviews had better results, reaching 96%. Precision reveals how well the model can differentiate classes. Figure 5 shows the comparison recall in the test of the three models, where SACReviews obtained the best result with 96% and IMDb only 77%. This metric shows how much a model can recognize of a given class.

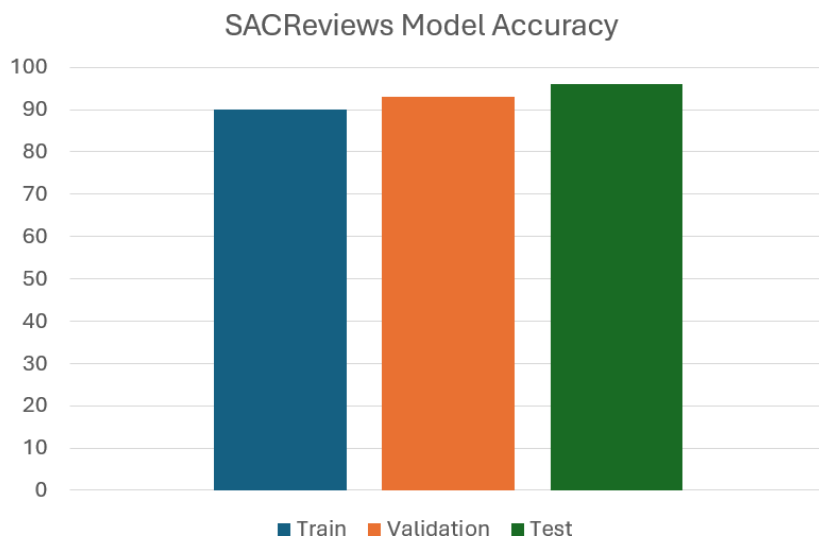


Figure 3. Approach SACReviews Accuracy

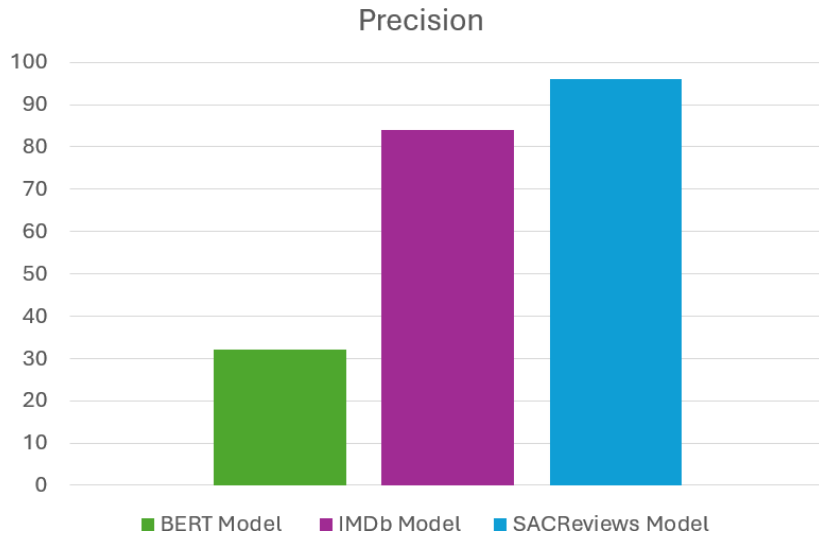


Figure 4. Precision

Figure 6 presents the comparison of the result in the F1-Score metric, in which it is observed that SACReviews managed to produce the best result with 95% compared to 76% of IMDb and 30% of BERT. This metric is the harmonic mean of precision and recall and seeks to approximate a balance between these two measures. 197 With the information presented in this section, we believe we have evidence of the importance of considering positive reviews with negative words during model construction, as the BERT and IMDb models obtained, respectively, 44% and 77% accuracy, while the SACReviews got 96%. The results also indicate that the SACReviews model is viable in analyzing sentiments of customer reviews in Portuguese, where negative words can be used in positive reviews, as it obtained the best values for all accuracy metrics, precision, recall and f1-score. The database used contains synthetic data, therefore the results cannot be generalized, but they are relevant for a feasibility study such as the one carried out. The size of the database is also a threat to the generalization of results. To reduce this threat, the reviews generated were checked by the authors of the article in order to remove reviews that were too similar and a consensus was reached between the authors to define which revisions would remain in the CReviews database.

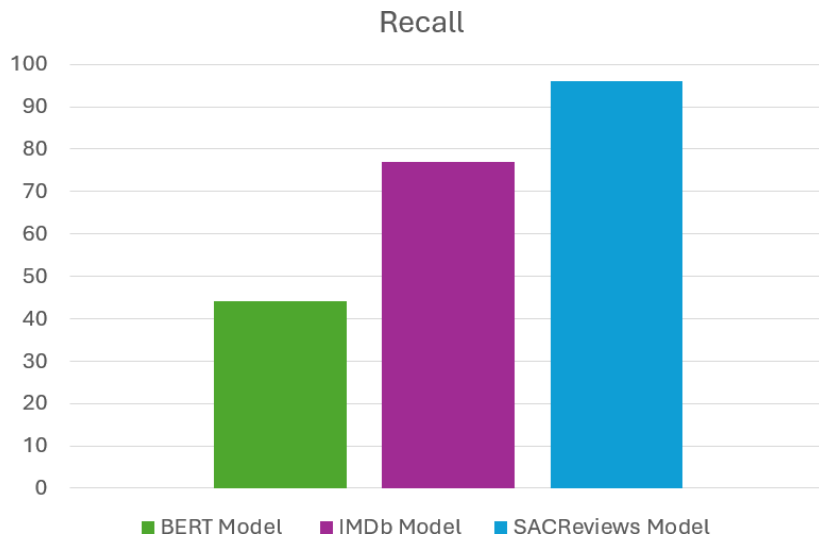


Figure 5. Recall

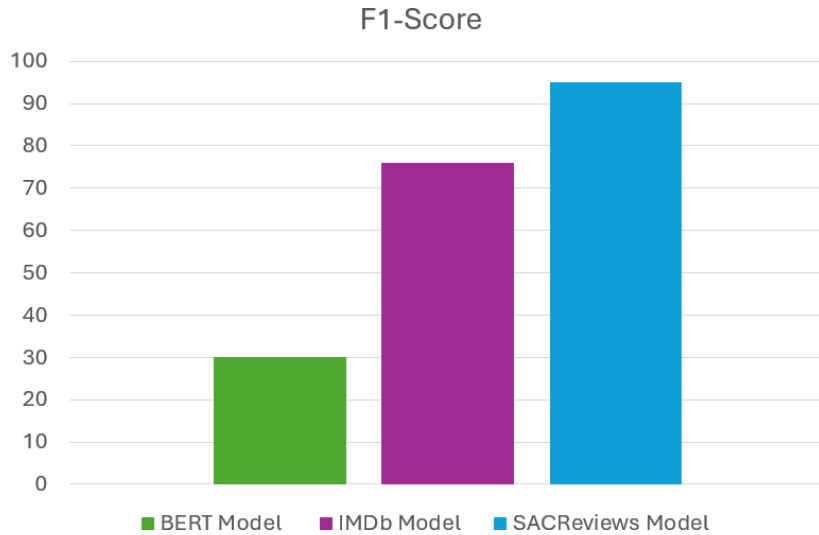


Figure 6. F1-Score

V. Conclusions

In this work, the SACReviews approach was presented, which aims to analyze the sentiments of customer reviews in Portuguese, but precisely the analysis of positive reviews using negative words. The study results demonstrated the need to consider positive comments with negative words during the construction of the model, as the BERT and IMDb models achieved an accuracy of 44% and 77%, compared to 96% for SACReviews. As well as it shows the possibility of using the SACReviews approach, as it achieved the best results for all metrics accuracy (96%), precision (96%), recall (96%) and f1-score (95%). For the fine-tuning of the SACReviews approach, the CReviews database was created, which contains 159 positive reviews and 144 negative reviews and is available for download. As future work, it would be interesting to evaluate the approach with a large database that contains positive reviews using negative words and with a database with real data.

Author Contributions: *Bruno Iglesias and Josias Lima:* Contributing to the conceptualization of the research, developing the methodology, analyzing the results and to the writing of the manuscript.

Funding: This research was funded by the Creathus Institute of Technology.

Data Availability Statement: The data is available at https://github.com/jlgomes/sentiment_analysis_with_bert.

Conflicts of Interest: The authors declare no conflict of interest.

References

- [1]. DataReportal. Digital 2022 global digital overview, 2022. <https://datareportal.com/reports/digital-2022-global-overview-report> (Accessed 11 January 2024).
- [2]. Statista. Retail e-commerce sales compound annual growth rate (CAGR) from 2023 to 2027, by country, 2024. <https://www.statista.com/forecasts/220177/b2c-e-commerce-sales-cagr-forecast-for-selected-countries> (Accessed 11 January 2024).
- [3]. Fatchiyah, A.; Sukmono, R.A. The Effect of Experiential Marketing and Brand Image on Purchase Decisions Through Word of Mouth as Intervening Variables. *Indonesian Journal of Innovation Studies* 2021, 16, 10–21070.
- [4]. Nandwani, P.; Verma, R. A review on sentiment analysis and emotion detection from text. *Social Network Analysis and Mining* 2021, 11, 81.
- [5]. Bas,a, S.N.; Basarslan, M.S. Sentiment Analysis Using Machine Learning Techniques on IMDB Dataset. In Proceedings of the 2023 7th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT). IEEE, 2023, pp. 1–5.
- [6]. Chouikhi, H.; Chniter, H.; Jarray, F. Arabic sentiment analysis using BERT model. In Proceedings of the Advances in Computational Collective Intelligence: 13th International Conference, ICCCI 2021, Kallithea, Rhodes, Greece, September 29–October 1, 2021, Proceedings 13. Springer, 2021, pp. 621–632.
- [7]. Gumiel, Y.B.; Lee, I.; Soares, T.A.; Ferreira, T.C.; Pagano, A. Sentiment analysis in Portuguese texts from online health community forums: data, model and evaluation. In Proceedings of the Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana. SBC, 2021, pp. 64–72. 249
- [8]. Geetha, M.; Renuka, D.K. Improving the performance of aspect based sentiment analysis using fine-tuned Bert Base Uncased

- model. *International Journal of Intelligent Networks* 2021, 2, 64–69.
- [9]. Vásquez, J.; Gómez-Adorno, H.; Bel-Enguix, G. Bert-based Approach for Sentiment Analysis of Spanish Reviews from TripAdvisor. In *Proceedings of the IberLEF@ SEPLN, 2021*, pp. 165–170.
- [10]. Lopes, E.; Correa, U.; Freitas, L. Exploring bert for aspect extraction in portuguese language. In *Proceedings of the The International FLAIRS Conference Proceedings, 2021*, Vol. 34.
- [11]. Xu, H.; Liu, B.; Shu, L.; Yu, P.S. BERT post-training for review reading comprehension and aspect-based sentiment analysis. *arXiv preprint arXiv:1904.02232* 2019.
- [12]. Biesner, D.; Ramamurthy, R.; Stenzel, R.; Lübbering, M.; Hillebrand, L.; Ladi, A.; Pielka, M.; Loitz, R.; Bauckhage, C.; Sifa, R. Anonymization of German financial documents using neural network-based language models with contextual word representations. *International Journal of Data Science and Analytics* 2022, pp. 1–11.

Bruno holds a degree in Software Engineering at the Federal University of Amazonas (UFAM). He is studying Mathematics at Estácio College. He works as a researcher and a developer of artificial intelligence technologies, machine vision, robotics and autonomous navigation using python as the main language at the Creathus institute.

Josias is a PhD student and Master at the Graduate Program in Informatics (PPGI) at the Institute of Computing at UFAM. Bachelor in Computer Science from North University Center (UNINORTE - Manaus - Brazil). His research interests involve Software Engineering, with an emphasis on Software Testing. He worked with the development of web and mobile systems using Java, PHP and JavaScript at Domma-Information Technology (Manaus - Brazil), at CITS Amazonas and currently works as a researcher at the Creathus institute.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.