Embedding generation for Text Classification of User Reviews in Brazilian Portuguese: From Bag-of-Words to Transformers

Frederico Dias Souza (*author*)
Eletrical Engineering Department
Federal University of Rio de Janeiro
fredericods@poli.ufrj.br

João Baptista de Oliveira e Souza Filho (*advisor*)
Eletrical Engineering Department
Federal University of Rio de Janeiro
jbfilho@poli.ufrj.br

***Abstract*—Text Classification is one of the most classical and studied Natural Language Processing (NLP) tasks. To classify documents accurately, a common approach is to provide a robust numerical representation, a process known as embedding. Embedding is a key NLP field that faced a significant advance in the last decade, especially after the popularization of Deep Learning models for solving NLP tasks, including Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Transformer-based Language Models (TLMs). Despite achievements, the literature regarding generating embeddings for Brazilian Portuguese texts still needs further investigation compared to the English language. Therefore, this work provides an experimental study of embedding techniques targeting a binary sentiment classification of user reviews in Brazilian Portuguese. This analysis includes classical (Bag-of-Words) to state-of-the-art (Transformer-based) NLP models. We evaluate the models over five open-source datasets containing pre-defined partitions to encourage reproducibility. The Fine-tuned TLMs attain the best results for all cases, followed by the Feature-based TLM, LSTM, and CNN, with alternate ranks depending on the dataset.**

***Index Terms*—Machine Learning, Deep Learning, Natural Language Processing, Sentiment Analysis, Text Classification**

I. Introduction

Opinions substantially influence human behavior, encouraging organizations to seek individual and collective opinions to leverage their business [1]. For example, streaming services crave user opinions for better movie recommendations, while e-commerce platforms strive to understand customer impressions for enhanced user experiences. With the surge in automatically generated data from digital interactions, opinion mining can now be automated, obviating the need for surveys or interviews. Consequently, there is an increasing demand for algorithms capable of performing automated opinion analysis.

A. Motivation

Text understanding in computers is highly challenging, usually requiring the transformation of text into numerical representations. This process, known as *embedding*, is a key part of text classification and is the main focus of this study.

The best ML methods for dealing with language data include supervised algorithms, a data modelling case which the ground truth is available. Text Classification (TC) is a classical supervised NLP task, which involves assigning labels to texts, and it represents this dissertation's object of study. Particularly, sentiment analysis targets to infer people's opinions expressed in texts, predicting if they are positive or negative [2]–[4].

While NLP solutions utilizing ML models have existed for some time, the last decade witnessed impressive advancements in Deep Learning (DL) algorithms. This began with Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), followed by the impact of Transformers. Transformer-based Language Models (TLMs) have redefined the state-of-the-art in numerous tasks by processing large text portions in parallel, achieving a deeper semantic understanding. Additionally, TLMs offer the advantage of addressing various NLP problems through transfer learning.

Regarding Brazilian Portuguese (PT-BR), several datasets, pre-trained resources, and models have been developed in the recent years for sentiment analysis in PT-BR texts, as can be seen in Pereira's work [5] and *Opinando* project [6]. However, to our best knowledge, studies comparing embeddings generated by the current state-of-the-art models (Transformed-based) with more consolidated approaches still require further exploration compared to the English language.

B. Objectives and Contributions

Due to the growing importance of inferring users' opinions and motivated by the necessity of experimental studies including the more recent NLP models in PT-BR, we conducted a experimental study of embedding alternatives targeting a binary sentiment classification task for PT-BR texts. We contemplated from traditional solutions to the state-of-the-art models in five open-source databases to ensure the generality and reproducibility of the findings. In the following, we summarize the main objectives and contributions of this thesis.

1) It collects five public annotated datasets of PT-BR user reviews for sentiment classification. We propose pre-defined partitions to facilitate comparisons and store them in a public repository[1]. We encourage researchers

[1]https://kaggle.com/datasets/fredericods/ptbr-sentiment-analysis-datasets

to evaluate alternative models using these partitions for direct comparisons with our reported results.
2) It provides a comprehensive study of feature generation techniques in PT-BR, covering strategies based on corpus statistics to transfer learning approaches, including word embedding strategies and TLMs. It also addresses intrinsic embeddings generated by end-to-end model training like CNNs, RNNs, and Fine-tuned TLMs.
3) Despite the recent advances, open-source models have emerged only recently, and more systematic studies still need to be developed, especially for our language. Thus, this work evaluates the predictive performance of 10 TLMs available for PT-BR, including three multilingual and seven language-specific.
4) It aids practitioners in selecting text classification algorithms by providing insights into trade-offs between predictive performance and computational resources of state-of-the-art models.
5) This master's dissertation generated two works published in international conferences and one journal article, as described in Appendix A.

II. Experimental Procedure

We defined a set of sequential steps to obtain sentiment predictions from texts, involving collecting annotated user reviews, cleaning the text, splitting it into tokens, potentially removing irrelevant tokens, and generating a embedding to feed the classification model, as depicted in Figure 1. We used five datasets (Appendix B) to evaluate all approaches.

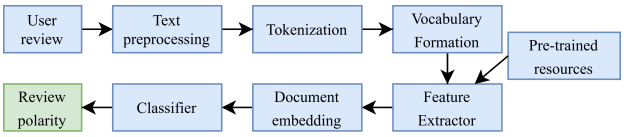

Fig. 1. General scheme of text classification on user reviews.

We categorized the analyzed models into three families: Bag-of-Words, Classical Deep Learning, and Transformer-based Large Language Models. Our focus in this work was on TLMs, exploring two design strategies: feature-based and fine-tuned. In the feature-based approach (FB TLM), TLMs were used to generate document embeddings followed by a Logistic Regression classifier. In contrast, the fine-tuned model (FT TLM) involved fine-tuning the weights of the TLM specifically for sentiment analysis. Appendix C provides further details about the embedding generation process, focusing on the TLMs.

III. Results and Discussion

In the dissertation's original text, we described the results for each model family in separate sections, trained on a training-validation-test scheme, and raised the test ROC-AUC. Here, we detailed only the results for the TLMs. After identifying the best performance configuration for each family, in Section III-C, we conducted a 10-fold cross-validation experiment, computing the average and standard deviation values of the performance metrics and submitting these results to statistical testing for a more rigorous analysis.

A. Feature-Based Large Language Models

The analysis shown in Table I considered ten TLMs, five token aggregation modalities (additional ones were explored in the original text), and five datasets. After ranking the results for each database, an average rank for each TLM was computed to provide an overall performance index, regardless of the database and the aggregation modality. The table also includes the average time, the reserved and allocated vRAMs[2].

TABLE I
Average ranks for the Transformer-based Language Models, and the average time, the reserved vRAM (GB), the allocated vRAM (GB) required to process a batch with size equal to 128.

Model	Average Rank	Time (s)	Reserved vRAM	Allocate vRAM
BERTimbau Large	9.4	1.0	1.7	1.3
PTT5 Large	9.7	1.0	3.3	2.8
BERTimbau Base	16.8	0.5	0.8	0.4
PTT5 Base	17.2	0.5	1.3	0.8
XLM-R Large	22.7	1.5	10.3	2.1
XLM-R Base	26.6	0.9	8.8	1.0
PTT5 Small	36.3	0.5	0.5	0.2
GPT2 Small	37.0	0.6	3.6	0.5
m-BERT	38.0	0.5	1.1	0.7
GPTNeo Small	45.4	0.6	3.4	0.5

Large model versions consistently outperform the base models. PTT5 is competitive with BERTimbau, followed by XLM-Roberta (XLM-R), which are significantly better alternatives than m-BERT, GPT2 Small, and GPTNeo Small. The GPT models are more suitable for content generation than text classification due to their architecture based only on decoder blocks and unidirectional auto-regressive token processing.

Regarding the remaining parameters reported in Table I, the reserved vRAM is correlated with the model size, while the allocated vRAM is with the embedding size. Large models generally require up to twice the processing time and memory space compared to their Base versions. Despite being multilingual, XLM-R achieved interesting results at the expense of significant computational resources. PTT5 Large performs competitively with BERTimbau in performance and processing time but requires significantly more memory. Therefore, BERTimbau Base offers the most attractive trade-off between performance and computational resources.

Figure 2 depicts the predictive performance for each dataset, model, and aggregation modality, restricting only to the top-seven TLMs. Modalities using multiple tokens generally outperformed single token approaches. Unlike BERT, models like XLM-Roberta Large achieved competitive results when using the "last" token, likely due to the absense of a dedicated classification token (CLS).

[2]The allocated vRAM corresponds to the portion of GPU memory currently used, while the reserved vRAM is related to the cache memory allocated.

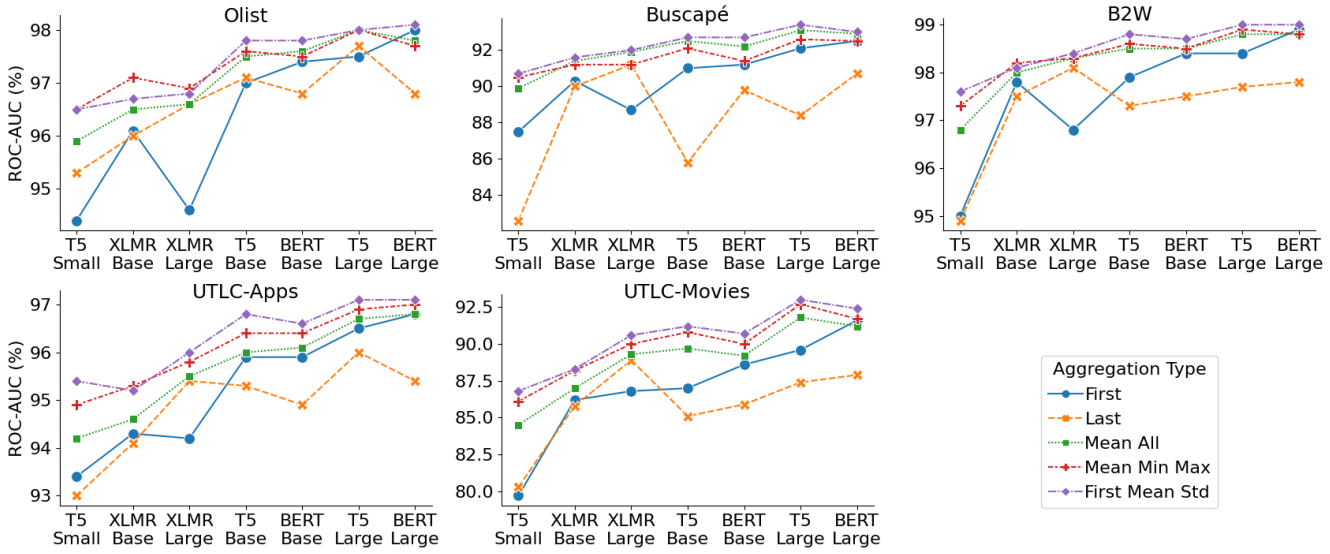

Fig. 2. Values of the ROC-AUC (%) per Large Language Model, Embedding Type, and Dataset (see text).

B. Fine-tuned Large Language Models

Table II displays results for fine-tuned BERT models and the best-performing feature-based BERT. Fine-tuning surpasses all pre-trained and baseline alternatives across datasets, particularly benefiting *UTLC-Movies*. The m-BERT model exhibits the largest performance increase but falls short of finetuned BERTimbau, which benefits from a large pre-training dataset.

TABLE II
VALUES OF THE ROC-AUC (%) FOR THE FINE-TUNED (FT) BERT MODELS, COMPARED TO THE FEATURE-BASED (FB) BERT WITH THE AGGREGATION MODALITY "FIRST+MEAN+STD".

TLM	Type	Olist	Buscapé	B2W	UTLC Apps	UTLC Movies
m-BERT	FB	97.0	90.0	97.3	94.7	84.9
	FT	97.6	91.8	98.6	97.4	94.1
	delta	0.6	1.8	1.3	2.7	9.2
BERtimbau Base	FB	97.8	92.7	98.7	96.6	90.7
	FT	98.5	93.4	99.2	97.9	95.6
	delta	0.7	0.7	0.5	1.3	4.9
BERtimbau Large	FB	98.1	93.0	99.0	97.1	92.4
	FT	98.6	94.1	99.3	97.9	95.8
	delta	0.5	1.1	0.3	0.8	3.4

C. Statistical Models' Comparison

Table III shows mean and standard deviation for Accuracy, F1-Score, and ROC-AUC across the ten test folds. UTLC-Movies exhibits the lowest performance, likely due to its focus on nuanced movie reviews rather than product reviews.

Friedman's Chi-Square Test [7] confirms statistically significant differences in performance ($p < 0.1\%$) among the methods for all datasets. Based on the Posthoc Tukey test, we established the following relations.

- **Olist** BoW<CNN=LSTM<FB TLM<FT TLM
- **Buscapé** BoW<CNN=LSTM=FB TLM<FT TLM

TABLE III
MEAN AND STANDARD DEVIATION OF THE ACCURACY (%), F1-SCORE (%) AND ROC-AUC (%) OBTAINED WITH EACH MODEL AND DATABASE.

Metric	Model Family	Olist	Buscapé	B2W	UTLC Apps	UTLC Movies
Acc	BoW	91.8±0.2	94.8±0.2	94.0±0.3	92.3±0.1	93.1±0.0
	CNN	93.3±0.6	95.7±0.2	94.7±0.6	93.1±0.6	93.7±0.1
	LSTM	93.4±0.6	95.5±0.2	94.4±1.0	93.6±0.1	94.0±0.1
	FB TLM	94.7±0.4	95.6±0.2	96.1±0.2	93.6±0.1	93.2±0.1
	FT TLM	95.3±0.3	96.0±0.1	97.0±0.1	94.9±0.1	95.2±0.1
F1 Score	BoW	94.2±0.2	97.2±0.1	95.7±0.2	95.1±0.0	96.2±0.0
	CNN	95.2±0.4	97.7±0.1	96.2±0.4	95.6±0.4	96.5±0.0
	LSTM	95.3±0.4	97.6±0.1	96.0±0.6	95.9±0.1	96.6±0.1
	FB TLM	96.2±0.3	97.6±0.1	97.2±0.1	95.9±0.0	96.2±0.0
	FT TLM	96.6±0.2	97.8±0.1	97.8±0.1	96.7±0.1	97.3±0.0
ROC AUC	BoW	96.6±0.3	91.9±0.6	98.1±0.1	96.1±0.1	92.8±0.1
	CNN	97.7±0.1	93.3±0.6	98.8±0.1	97.0±0.1	93.7±0.1
	LSTM	97.6±0.2	93.2±0.6	98.8±0.1	97.3±0.1	94.5±0.1
	FB TLM	98.0±0.1	93.0±0.6	99.0±0.1	97.2±0.1	92.9±0.1
	FT TLM	98.4±0.2	94.3±0.4	99.4±0.1	98.1±0.1	96.1±0.1

- **B2W** BoW<CNN=LSTM<FB TLM<FT TLM
- **UTLCApps** BoW<CNN<LSTM=FB TLM<FT TLM
- **UTLCMovies** BoW=FB TLM<CNN<LSTM<FT TLM

As expected, the BoW is the worst, while the FT BERT is always the best. CNN and LSTM perform similarly in Olist, Buscapé, and B2W databases, with LSTM surpassing CNN in UTLC. FB-TLM is comparable or superior to classic Deep Learning models, except for UTLC-Movies, where end-to-end model training performs better due to the dataset's complexity. Considering practical applications, BoW models remain attractive, with simplicity and low computational burden. FB TLM is an excellent intermediate alternative, outperforming classic Deep Learning models with open-source pre-trained models.

REFERENCES

[1] B. Liu, "Sentiment analysis and opinion mining," *Synthesis Lectures on Human Language Technologies*, vol. 5, no. 1, 2012. [Online]. Available: http://dx.doi.org/10.2200/S00416ED1V01Y201204HLT016

[2] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, and J. Gao, "Deep learning based text classification: A comprehensive review," 2020. [Online]. Available: https://arxiv.org/abs/2004.03705

[3] Q. Li, H. Peng, J. Li, C. Xia, R. Yang, L. Sun, P. S. Yu, and L. He, "A survey on text classification: From shallow to deep learning," 2020. [Online]. Available: https://arxiv.org/abs/2008.00364

[4] Kowsari, J. Meimandi, Heidarysafa, Mendu, Barnes, and Brown, "Text classification algorithms: A survey," *Information*, vol. 10, no. 4, 2019.

[5] D. A. Pereira, "A survey of sentiment analysis in the portuguese language," *Artificial Intelligence Review*, vol. 54, no. 2, 2020.

[6] "Opinion mining for portuguese," 2019. [Online]. Available: https://sites.google.com/icmc.usp.br/opinando/pgina-inicial

[7] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *JMLR*, vol. 7, p. 1–30, dec 2006.

[8] Olist, "Brazilian e-commerce public dataset by olist," Nov 2018. [Online]. Available: https://www.kaggle.com/olistbr/brazilian-ecommerce

[9] L. Real, M. Oshiro, and A. Mafra, "B2w-reviews01 - an open product reviews corpus," *STIL - Symposium in Information and Human Language Technology*, 2019. [Online]. Available: https://github.com/b2wdigital/b2w-reviews01

[10] N. Hartmann, L. Avanço, P. Balage, M. Duran, M. das Graças Volpe Nunes, T. Pardo, and S. Aluísio, "A large corpus of product reviews in Portuguese: Tackling out-of-vocabulary words," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik, Iceland: European Language Resources Association (ELRA), May 2014.

[11] R. F. d. Sousa, H. B. Brum, and M. d. G. V. Nunes, "A bunch of helpfulness and sentiment corpora in brazilian portuguese," *Symposium in Information and Human Language Technology - STIL*, 2019.

[12] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, and A. Rush, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Oct. 2020, pp. 38–45. [Online]. Available: https://aclanthology.org/2020.emnlp-demos.6

[13] A. Radford, J. Wu, R. Child *et al.*, "Language models are unsupervised multitask learners," 2019. [Online]. Available: https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf

[14] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Jun. 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423

[15] Jay Alammar, "A visual guide to using BERT for the first time," 2019. [Online]. Available: https://jalammar.github.io/a-visual-guide-to-using-bert-for-the-first-time/

[16] C. Sun, X. Qiu, Y. Xu, and X. Huang, "How to fine-tune BERT for text classification?" in *Chinese Computational Linguistics*, M. Sun, X. Huang, H. Ji, Z. Liu, and Y. Liu, Eds. Springer International Publishing, 2019, pp. 194–206.

APPENDIX A
ACADEMIC PUBLICATIONS

This appendix presents the publications produced during the development of this master's dissertation.

- Souza, F.D., Filho, J.B.O. (2021). "Sentiment Analysis on Brazilian Portuguese User Reviews". In: 2021 IEEE Latin American Conference on Computational Intelligence (LA-CCI). https://doi.org/10.1109/LA-CCI48322.2021.9769838
  Abstract: Sentiment Analysis is one of the most classical and primarily studied natural language processing tasks. This problem had a notable advance with the proposition of more complex and scalable machine learning models. Despite this progress, the Brazilian Portuguese language still disposes only of limited linguistic resources, such as datasets dedicated to sentiment classification, especially when considering the existence of predefined partitions in training, testing, and validation sets that would allow a more fair comparison of different algorithm alternatives. Motivated by these issues, this work analyzes the predictive performance of a range of document embedding strategies, assuming the polarity as the system outcome. This analysis includes five sentiment analysis datasets in Brazilian Portuguese, unified in a single dataset, and a reference partitioning in training, testing, and validation sets, both made publicly available through a digital repository. A cross-evaluation of dataset-specific models over different contexts is conducted to evaluate their generalization capabilities and the feasibility of adopting a unique model for addressing all scenarios.
- Souza, F.D., Filho, J.B.O. (2022). "BERT for Sentiment Analysis: Pre-trained and Fine-Tuned Alternatives". In: Computational Processing of the Portuguese Language. PROPOR 2022. Lecture Notes in Computer Science, vol 13208. Springer, Cham. https://doi.org/10.1007/978-3-030-98305-5_20
  Abstract: BERT has revolutionized the NLP field by enabling transfer learning with large language models that can capture complex textual patterns, reaching the state-of-the-art for an expressive number of NLP applications. For text classification tasks, BERT has already been extensively explored. However, aspects like how to better cope with the different embeddings provided by the BERT output layer and the usage of language-specific instead of multilingual models are not well studied in the literature, especially for the Brazilian Portuguese language. The purpose of this article is to conduct an extensive experimental study regarding different strategies for aggregating the features produced in the BERT output layer, with a focus on the sentiment analysis task. The experiments include BERT models trained with Brazilian Portuguese corpora and the multilingual version, contemplating multiple aggregation strategies and open-source datasets with predefined training, validation, and test partitions to facilitate the reproducibility of the results. BERT achieved the highest ROC-AUC values for the majority of cases as compared to TF-IDF. Nonetheless, TF-IDF represents a good trade-off between the predictive performance and computational cost.
- Souza, F.D., Filho, J.B.O. (2022). Embedding generation for text classification of Brazilian Portuguese user reviews: from bag-of-words to transformers. Neural Computing & Applications. https://doi.org/10.1007/s00521-022-08068-6.
  Abstract: Text classification is a Natural Language Processing (NLP) task relevant to many commercial applications, like e-commerce and customer service. Naturally, classifying such excerpts accurately often represents a challenge, due to intrinsic language aspects, like irony and nuance. To accomplish this task, one must provide a robust numerical representation for documents, a process known as embedding. Embedding represents a key NLP field nowadays, having faced a significant advance in the last decade, especially after the introduction of the word-to-vector concept and the popularization of Deep Learning models for solving NLP tasks, including Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Transformer-based Language Models (TLMs). Despite the impressive achievements in this field, the literature coverage regarding generating embeddings for Brazilian Portuguese texts is scarce, especially when considering commercial user reviews. Therefore, this work aims to provide a comprehensive experimental study of embedding approaches targeting a binary sentiment classification of user reviews in Brazilian Portuguese. This study includes from classical (Bag-of-Words) to state-of-the-art (Transformer-based) NLP models. The methods are evaluated with five open-source databases with predefined data partitions made available in an open digital repository to encourage reproducibility. The Fine-tuned TLMs achieved the best results for all cases, being followed by the Feature-based TLM, LSTM, and CNN, with alternate ranks, depending on the database under analysis.

APPENDIX B
DATASETS

We considered the following five datasets: *Olist* [8], *B2W* [9], *Buscapé* [10], *UTLC-Apps*, and *UTLC-Movies* [11].

1) **Olist**: in 2018, Olist, Brazilian marketplace, launched the "Brazilian E-Commerce Public Dataset by Olist" [8], a database with approximately 100,000 orders from 2016 to 2018 provisioned by several marketplaces in Brazil. This work adopted the *olist order reviews dataset*, which disposes of the user comments plus a label with a satisfaction rate ranging from 1 to 5.
2) **B2W**: in 2019, B2W Digital, one of the most prominent Latin American e-commerce, released the *B2W Reviews01* [9], an open corpus of product reviews with more than 130,000 user reviews. We considered only as target feature the user ratings from 1 to 5 stars.
3) **Buscapé**: as described by Hartmann et al. [10], the *Corpus Buscapé* is a large corpus of product reviews in Portuguese, crawled in 2013, with more than 80,000 samples from the *Buscapé*, a product and price search website. Unlike the datasets previously described, the Opinando labels' range is from 0 to 5, leading us to remove the comments rated as zero.
4) **UTLC-Apps and UTLC-Movies**: the *UTLCorpus* [11] is the most extensive set considered here, having more than 2 million reviews. It includes movie reviews collected from the *Filmow*, a famous movie Social network, and mobile apps comments collected from the Google Play Store. Here, the *UTLCorpus* was split into two different datasets: the *UTLC-Movies* and the *UTLC-Apps*. Similar to the Buscapé database, reviews with a rating equal to 0 were excluded.

In all evaluations, we considered the binary polarity target feature. Table IV summarizes the number of samples, document length, vocabulary size, and the polarity distribution.

It is worth mentioning that the dataset domains significantly influence how the models represent the texts [1]. Four datasets (Olist, Buscapé, B2W, and UTLC-Apps) have product reviews, and the dataset UTLC-Movies contains movie reviews, which tends to be more nuanced, so the models may struggle in the sentiment classification task.

TABLE IV
SUMMARY OF THE DATASETS USED IN THIS WORK: NUMBER OF SAMPLES FOR THE TRAINING, VALIDATION AND TEST SETS, MEAN/MEDIAN DOCUMENT LENGTH, VOCABULARY SIZE, AND THE POLARITY DISTRIBUTION.

Dataset	Train/Valid/Test Samples (10^3)	Mean/Median Length	Vocab size (1 gram)	Labels distribution (Positive)
Olist	30 / 4 / 4	7/6	3.272	70.0%
Buscapé	59 / 7 / 7	25/17	13.470	90.8%
B2W	93 / 12 / 12	14/10	12.758	69.2%
UTLCApps	775 / 97 / 97	7/5	28.283	77.5%
UTLCMovies	952 / 119 / 119	21/10	69.711	88.4%

APPENDIX C
MODELS

Table V provides an overview of all experiments performed in this work. For didactic reasons, we divided the analyzed models into three families - Bag-of-Words, Classical Deep Learning, and Transformer-based Large Language Models - and presented the techniques in order of increasing complexity in the text. Table V also shows the variations performed in each model. For the sake of brevity, we detailed here only the embedding generation process for the TLMs.

A. Transformer-based Language Models

The Transformer-based Language Models considered three multilingual and seven language-specific models. We evaluated only open-source TLMs for Portuguese released by the Hugging Face [12] initiative, an open-source NLP community. Brazilian Portuguese variants of the GPT (Generative Pre-trained Transformer) on this platform do not have papers attached but were included in the experiments of this work, such as the GPT-Neo Small Portuguese[3] and GPorTuguese-2[4], representing fine-tuned versions of the GPT-Neo 125M[5] and GPT-2 Small [13], respectively. Table VI summarizes the dimensionality of the related embeddings and the number of model parameters, aiming to provide some guiding information over the practical trade-offs between complexity and performance observed with these models in our experiments.

TABLE VI
SUMMARY OF SOME CHARACTERISTICS OF THE TRANSFORMER-BASED LANGUAGE MODELS EVALUATED IN THIS WORK.

Model	Multilingual	Embedding length	Parameters ($\times 10^6$)
PTT5 Small	No	512	60
m-BERT	Yes	768	110
BERTimbau Base	No	768	110
GPT2 Small	No	768	117
XLM-Roberta Base	Yes	768	125
GPTNeo Small	No	768	125
PTT5 Base	No	768	220
BERTimbau Large	No	1024	345
XLM-Roberta Large	Yes	1024	355
PTT5 Large	No	1024	770

The experiments included two design strategies: pre-trained (or feature-based) and fine-tuned. The pre-trained solution considered just using the TLM for producing document embeddings, which were subsequently fed to a Logistic Regression classifier. Conversely, for the fine-tuned model, the TLM weights were fine-tuned to the sentiment analysis task. The experiments solely exploring pre-trained models for generating document embeddings will be referred to here as Feature-based TLM (FB TLM). In turn, those involving fine-tuning are named here as Fine-tuned TLM (FT TLM).

[3]www.huggingface.co/HeyLucasLeao/gpt-neo-small-portuguese
[4]www.huggingface.co/pierreguillou/gpt2-small-portuguese
[5]www.huggingface.co/EleutherAI/gpt-neo-125M

B. Feature-based TLMs

Previous works [14] reported that creative combinations of the token representations provided by the BERT outputs might lead to a significant performance improvement in the NER task, even without any fine-tuning of model parameters. Such findings strongly motivated the following experiments. As shown in Figure 3, TLMs output a 3D tensor whose number of rows is equal to the batch size, the number of columns corresponds to the number of tokens, and the embedding size defines the depth. In this way, each column of this tensor represents a sequence of embeddings corresponding to some position-specific token of the documents integrating this batch. The evaluated TLMs models assumed documents constituted by one to sixty tokens. To each token, these models produced a representation having from 512 to 1024 dimensions, as shown in Table VI, referred to as token embedding. In the following, we describe the approaches evaluated in this work to combine these embeddings. In parenthesis, we exhibited the number of vector concatenations of each case. Thus, the size of the vectors used for document representations has from 512 (512×1) up to 3072 (1024×3) dimensions. Due to a high computational burden, for the TLMs T5 Large, XLM-RoBERTa Base, and XLM-RoBERTa Large, we limited the aggregation types to ”first”, ”last”, ”mean all”, ”first + mean + std”, and ”mean + min + max”.

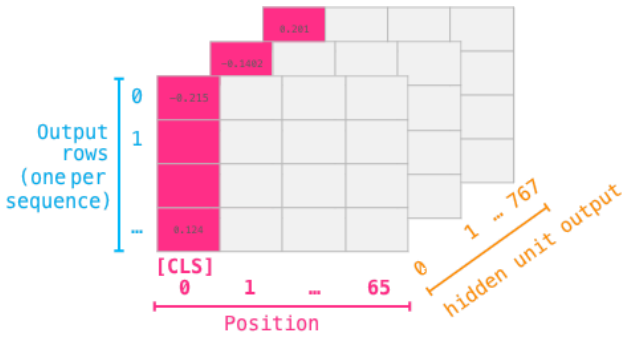

Fig. 3. Example of BERT Base output. Adapted from [15].

1) **first** (1): this embedding corresponds to the first token. For BERT models, it is equivalent to the [CLS] special token, created with the purpose of sentence classification, and considering as the default BERT document embedding;
2) **second** (1): embedding corresponding to the 2nd token;
3) **last** (1): embedding corresponding to the last token;
4) **sum all** (1): sum of all token embeddings;
5) **mean all** (1): average of all token embeddings;
6) **sum all except first** (1): sum of all token embeddings, ignoring the first one;
7) **mean all except first** (1): average of all token embeddings, ignoring the first one;
8) **sum + first** (2): concatenation of the sum of all token embeddings and the first token embedding;
9) **mean + first** (2): concatenation of the average of all token embeddings and the first token embedding;

TABLE V
SUMMARY OF THE EVALUATED MODEL VARIATIONS.

Model Family	Model	Variations
Bag-of-Words (BoW)	TF-IDF	**Feature selection methods:** frequency, chi2, fvalue
	TF-IDF+SVD (LSA)	-
	Bag of Word Vectors Averaged (avgBoWV)	**Word vector model:** Word2vec, GloVe, FastText **Word vector dimension:** 50, 100, 300
	IDF Weighted Bag of Word Vectors (idfBoWV)	**Word vector dimension:** 50, 100, 300
	Bag of Word Vectors with first principal component removal (BoWV-PC)	**Weighting scheme:** unweighted, idf-weighted **Word vector dimension:** 50, 100, 300
Classical Deep Learning (CDL)	Convolutional neural networks (CNN)	**Filter sizes:** [2], [2,3], [2,3,4], [2,3,4,5] **Feature map size**: 50, 100, 200, 400
	Long short term memory neural networks (LSTM)	**Layers:** 1, 2 **Hidden size:** 64, 128, 256 **Pooling layer:** average pooling, max pooling, avg and max concatenated
Trasformer-based Large Language Model (TLM)	Feature-based TLM (FB TLM)	**Models:** see Table VI **Aggregation types:** see Section C-B
	Finetuned TLM (FT TLM)	-

10) **first + mean + std** (3): concatenation of the embeddings of the first token, the average, and the standard deviation of the remaining tokens;
11) **first + mean + max** (3): concatenation of the embeddings of the first token, the average, and maximum of all token embeddings;
12) **mean + min + max** (3): concatenation of the average, minimum, and maximum of all token embeddings;
13) **quantiles 25, 50, 75** (3): concatenation of the quantiles 25%, 50%, and 75% of all token embeddings.

One of the major differences between BERT and alternative TLMs is its capability to accomplish the Next Sentence Prediction (NSP) task. Besides, BERT has a token dedicated to sentence classification (CLS). It is possible to include this token when fine-tuning the other models, which would make them more suitable for text classification. However, this work restricts analyzing them in a feature-based approach.

Unlike BoW, which is based on static word embeddings, the document embeddings generated by TLMs are contextual, resulting in significant gains in the predictive performance of the models, as will be shown later.

C. Fine-tuned TLM

The fine-tuning experiments were restricted to the BERT models (m-BERT, BERTimbau Base, and BERTimbau Large) since it would be too computationally demanding to fine-tune all TLMs. In addition, with these three BERT variants, we could verify the influence of fine-tuning on multi-lingual vs. language-specific models and on Based vs. Large size models.

Design choices, such as the training algorithm and the model hyperparameters, were based on Sun et al. [16]. A Logistic Regression network was added on the top of the layer associated with the [CLS] token, targeting to predict one of the two sentiment classification classes, i.e., assuming as target-values 0 or 1. The network training adopted the Adam optimizer with weight decay, slanted triangular learning rates with a warm-up proportion of 0.1, and a maximum number of epochs equal to 4.