A Data Augmentation and Validation Pipeline for Improving Emotion Classification in Mobile App Reviews

Kalidsa B. de Oliveira[1], Gabriel M. Lunardi[1], Williamson Silva[2], Thiago L. T. da Silveira[1], Adriano Q. Oliveira[1]

[1]Universidade Federal de Santa Maria - UFSM, Santa Maria, RS, Brasil

[2]Universidade Federal do Cariri – UFCA, Juazeiro do Norte, CE, Brasil

kalidsa.oliveira@ecomp.ufsm.br, williamson.silva@gmail.com

{gabriel.lunardi, thiago.silveira, adriano.q.oliveira@ufsm.br}@ufsm.br

Abstract. *This paper examines GPT-2 based data augmentation to improve sentiment classification of Portuguese mobile app reviews, employing the BERTimbau model. A pipeline is proposed, integrating synthetic data generation with semantic analysis, UMAP for dimensionality reduction, and HDBSCAN for clustering and validation. Results show validated and balanced synthetic augmentation boosts model performance in sparse or imbalanced data scenarios.*

1. Introduction

Sentiment analysis of informal texts, such as reviews for mobile apps, products, and services, is challenging due to the scarcity of annotated data and, when available, the imbalance among emotion classes [Wankhade et al. 2022]. Although Data Augmentation (DA) is a promising approach to mitigate these problems by creating synthetic instances, its effective application extends beyond simple text generation through semantic substitutions or restructuring [Sujana and Kao 2023]. A key research gap lies in the lack of methods to validate the quality and relevance of the generated data, ensuring they genuinely enrich the training set without introducing noise [Ding et al. 2024].

To address this gap, this paper proposes and evaluates a computational pipeline for generating and validating synthetic data. Using a dataset of 3,012 emotion-annotated app reviews as a case use, our approach integrates transformer-based models to create synthetic data, specifically combining GPT-2 and BERT, both adapted for Brazilian Portuguese. To ensure the quality and consistency of this augmented dataset, the pipeline employs the Language-agnostic BERT Sentence Embedding (LaBSE) technique for semantic filtering and a combination of Uniform Manifold Approximation and Projection (UMAP) with Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) to identify and analyze potential outliers.

Thus, models based on BERT have become prominent for their performance in various NLP tasks, especially in low-resource languages like Brazilian Portuguese [Aguiar 2025]. The primary goal, therefore, is to assess the impact of this synthetic data on the predictive performance of BERTimbau. The choice of a Small Language Model (SLM) like BERTimbau is motivated by its ability to offer a high-performance yet cost-effective alternative to resource-intensive Large Language Models (LLMs) [Gartner, Inc. 2025]. BERTimbau has 12 layers and approximately 110 million parameters, which classifies it as a SLM, in contrast to LLMs that often exceed billions of parameters [Moreira et al. 2023]. This makes SLMs an efficient and accessible tool for specialized tasks, particularly for languages with fewer resources.

2. Related Work

The methodological choices in this work are grounded in the recent success of BERT-based models for Natural Language Processing in Portuguese. Recent studies have reported promising results by employing variants of BERT for various tasks. For instance, [Silva et al. 2023] demonstrated significant gains in governmental text classification using BERTimbau and LaBSE semantic embeddings. The authors [Barbosa et al. 2022, de Almeida Neto and de Melo 2023, Carmo et al. 2023] have successfully used BERT-based models for thematic analysis in restaurant reviews and topic analysis in social media respectively. The consistent high performance of these models in Portuguese classification and thematic analysis substantiates our choice to use BERTimbau and LaBSE.

For the validation stage, we draw inspiration from the use of clustering techniques for textual analysis. The authors of [Murthy et al. 2025], for example, employed UMAP and HDBSCAN to identify topics in a large volume of hurricane-related tweets. Our work, however, utilizes this same combination of techniques for a distinct purpose: instead of topic modeling, we apply it to detect outliers and evaluate the semantic consistency of the augmented dataset, treating clustering as a quality control tool.

Finally, the evaluation of our method uses BERTimbau, a model whose effectiveness for Portuguese was highlighted by [Borges 2025] by showing its superiority in news clustering tasks compared to methods like TF-IDF. While the work of [Borges 2025] focuses on clustering, our contribution is to use BERTimbau as a classifier to quantitatively measure the impact of our validated augmentation process. Thus, the novelty of our research lies in the synergistic integration of these fronts, offering a complete workflow to generate, validate, and evaluate the impact of synthetic data in Portuguese.

3. Pre-processing

The aforementioned dataset was constructed based on Ekman's model of fundamental emotions by [Siqueira et al. 2024]. It features reviews manually classified into seven basic emotions —- happiness, surprise, sadness, neutral, fear, disgust, and anger – through a collaborative annotation process validated by multiple human evaluators. The dataset's aim is to support the development of applications for user experience (*UX*) analysis [Costa et al. 2024, Soares et al. 2025]. These comments contain grammatical errors, emojis, and, notably, subjective expressions of feelings and class imbalance.

For the text data pre-processing stage, a Python pipeline was developed to normalize the user comments, making them suitable for the subsequent stages of DA, sentiment analysis, and polarity classification. Initially, emojis present in the texts were converted to their respective textual descriptions in Portuguese using the 'emoji' library, with the aim of preserving the semantic and emotional charge often associated with these symbols. Next, accented characters were normalized to their basic ASCII form using the unicodedata library, reducing lexical variability caused by accents and facilitating text standardization.

Numbers identified in the messages were converted to their full-text form with the support of the num2words library, configured for Portuguese, in order to maintain lexical coherence and allow for better processing by language models. Characters not belonging to the Latin alphabet were removed with the help of the regex library, preserving only letters, spaces, and the characters : and _, the latter being temporarily kept due to their use in the intermediate representation of emojis. Furthermore, sequences of words repeated more than twice were reduced to a single occurrence to minimize noise typical of informal language. Finally, all text was converted to lowercase, and consecutive white spaces were replaced with a single space, also removing any leading or trailing spaces from the sentences.

After applying the DA technique, the dataset was subjected to the same preprocessing steps described previously, with the addition of stopword removal. Subsequently, tokenization and lemmatization of the texts were performed to achieve linguistic normalization. The resulting corpus was then reused in the training process of the BERTimbau model in order to evaluate the effects of DA on the model's performance.

4. Methodology

To mitigate data imbalance, we implemented a DA process based on prior work like in [de Oliveira et al. 2025]. This approach was chosen due to the large volume of data generated daily by users, which makes manual collection unfeasible, and as an alternative to web scraping. All the pipeline, datasets and other artifacts are available at Github[1]

For the DA process, we selected the pierreguillou/gpt2-small-portuguese model, a version of GPT-2 adapted for Brazilian Portuguese. The decision to use this model over more recent Large Language Models (LLMs) was based on two factors. First, its open-source nature and lower computational cost make it an accessible alternative that aligns with our work's focus on proposing a low-cost and highly reproducible pipeline — a principle also applied in the choice of the BERTimbau classifier. Second, using a model specifically tailored for Portuguese ensures greater linguistic fidelity in the text generation.

To this end, we began by counting the number of examples per sentiment class, as shown in Table 1. This was fundamental to guide the DA, allowing the number of examples in each class to be balanced based on the majority class, which is disgust. The logic for DA involved randomly selecting examples from minority classes from the original (pre-processed) set and from content automatically generated by the GPT-2 model. We established that synthesized comments must contain a minimum of three words to avoid overly short and semantically poor samples. After generating new examples, we applied a similarity filtering step using the sentence-transformers/LaBSE model, as described by [Feng et al. 2022]. The purpose of LaBSE is to transform sentences into dense vector representations while preserving semantic meaning, with support for 109 languages. From these vectors, cosine similarity is calculated to identify and remove duplicates, promoting diversity and quality in the augmented data. In total, the model generated 3,653 comments; of these, 2,260 (62%) were removed for having a similarity score of 92% or higher, leaving 1,393 comments.

[1] https://github.com/Kalidsa/sbbd-2025-apps_reviews

Table 1. Sentiments count

Sentiment	before DA	after DA
Sadness	864	893
Happiness	319	556
Disgust	**952**	**952**
Anger	743	828
Fear	47	259
Surprise	4	188
Neutral	82	521

Subsequently, we trained and evaluated the neuralmind/bert-base-portuguese-cased model on both the original and the augmented datasets; this was performed separately for the sentiment and polarity classification tasks. The process began with text pre-processing via tokenization and label encoding, using AutoTokenizer from the Transformers library and LabelEncoder from Scikit-Learn, respectively. The data was then partitioned into training and testing subsets using a stratified split (80% for training and 20% for evaluation). This partitioning preserves the original class proportions, which is essential for preventing bias during model training, especially with imbalanced data.

After the split, the data was converted into Dataset objects from the datasets library to ensure compatibility with the Hugging Face Trainer API. Training was conducted using the Trainer with hyperparameters defined based on the study by [omitted for review]: a learning rate of 0.00002, a batch size of 10 for training and validation, 3 epochs, an evaluation strategy of "epoch", and a weight decay of 0.01. After training, predictions were generated on the test set, and performance metrics (precision, recall, F1-score, and accuracy) were calculated.

Finally, considering that the synthetic data was generated from a model trained on diverse Portuguese text sources — many of which do not directly correspond to the specific domain of mobile app reviews — we employed a combination of UMAP and HDBSCAN. UMAP was used for dimensionality reduction, preserving the topological structure of the data and facilitating the visualization of latent patterns. Next, we applied the HDBSCAN algorithm, Figure 1, an unsupervised density-based clustering technique with an intrinsic ability to identify noise, to group semantically similar comments and detect isolated observations. This approach aimed to identify potential data discrepancies, enabling a qualitative assessment of the compatibility and consistency of the synthetic data with the target domain.

5. Results

The purpose of applying DA was to promote balance in the original dataset by increasing the representation of minority classes and, consequently, improving the model's performance and the fairness of its evaluation metrics. A significant increase in the total number of examples was observed after applying DA. However, this increase was not sufficient to satisfactorily correct the imbalance between categories, especially for low-frequency classes such as "surprise" and "fear". This factor may have contributed to the GPT-2-based model's limited ability to adequately generalize comments associated with these emotions, which is reflected in its lower performance for these classes. Furthermore, the

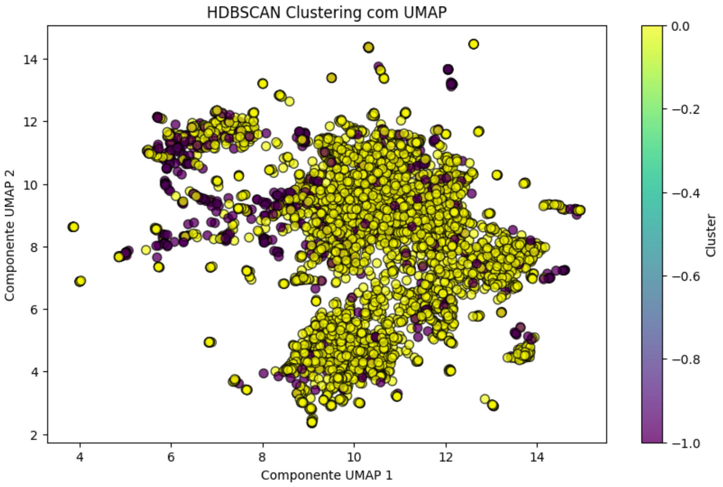

Figure 1. Visualization of the HDBSCAN clustering applied to the UMAP components. Colors represent different cluster labels, including those considered noise (in purple).

exclusion of similar texts during the similarity verification process may also have negatively impacted the model's learning capacity, further limiting its effectiveness on the least represented emotions.

For the clustering of comments, we used the Silhouette Coefficient and Adjusted Rand Index (ARI) metrics to evaluate intra-cluster cohesion and inter-cluster separation, respectively. To complement this analysis, we calculated the Euclidean distance of each point to the centroid of its respective cluster in the two-dimensional space generated by UMAP. The goal was to quantify the deviation of each comment from its cluster's center of mass. This approach allows us to identify which comments align with the group's predominant context and which are significantly distant, potentially being considered out of context (outliers).

The identification of outliers was refined by analyzing the 150 most frequent words in the corpus. Comments with a significantly greater distance to their centroid were considered noise candidates and stored for qualitative inspection. In total, 519 comments were flagged as potential outliers based on this distance. Of these, 43 had one or no occurrences of words from the most frequent vocabulary, suggesting they are irrelevant to the analyzed domain. By analyzing the datasets separately, we identified 302 outliers in the original set and 217 in the augmented set. This indicates that noise collected via web scraping was incorporated into the base dataset and subsequently propagated during the DA process.

The results we obtained with the BERTimbau model on the sentiment and polarity classification tasks demonstrate the positive impacts of DA on the model's predictive quality. In Table 2, for the sentiment classification task, the model trained with augmented data achieved consistent improvements across all evaluated metrics, with an average increase of 10.13% in all metrics. For the polarity task, as shown in Table 3, the average increase in metrics was 6.14%. These results indicate that the DA technique contributed significantly to improving the model's ability to generalize and capture variations in the input data, especially in contexts with greater lexical and structural diversity. However, it is worth noting that outliers were included in these training and evaluation processes to assess the effectiveness of DA under real-world conditions, which often include noise and irrelevant data from automated collection methods.

Table 2. Classification Performance Results of Sentiments

Sentimento	BERTimbau without DA	BERTimbau with DA
Precision	67.5%	75.2%
Recall	67.5%	74.6%
F1-Score	66.3%	74.7%
Accuracy	67.5%	74.6%

Table 3. Classification Performance Results of Polarity

Sentimento	BERTimbau without DA	BERTimbau with DA
Precision	82.1%	88.1%
Recall	82.9%	87.9%
F1-Score	82.3%	87.9%
Accuracy	82.9%	87.9%

6. Final Remarks

This paper evaluated the effectiveness of a DA and validation pipeline for sentiment classification in app reviews. The experiments demonstrated that while generation with GPT-2 has limitations in creating diverse samples, our validation method was important for ensuring the quality of the final dataset. The propagation of outliers from the original set to the augmented one also reinforces the need for the filtering and consistency analysis steps that we propose. The positive impact we observed on the BERTimbau classifier's metrics indicates that even a moderate data increase, when properly filtered and validated, contributes to the model's performance on informal and imbalanced texts.

For future work, we intend to investigate more advanced generation techniques, such as Generative Adversarial Networks (GANs), and the evaluation of larger-capacity language models accessed via specialized APIs, aiming to expand the diversity and quality of the generated synthetic data.

Acknowledgments

The authors thank CNPq for the support of the Universal Projects 405973/2021-7 and 402086/2023-6, as well as FAPERGS for the project: ARD/ARC – Process 24/2551-0000645-1.

References

Aguiar, M. S. (2025). Comparative analysis of the performance of large language models in the classification of legal texts.

Barbosa, M., Valle, P., Nakamura, W., Guerino, G., Finger, A., Lunardi, G., and Silva, W. (2022). Um estudo exploratório sobre métodos de avaliaçao de user experience em chatbots. In *Escola Regional de Engenharia de Software (ERES)*, pages 21–30. SBC.

Borges, W. A. (2025). Uso do bertimbau para o pré-processamento e agrupamento de comentários de notícias. *Informática na Educação: teoria e prática*, 28(1):1–20.

Carmo, I., Rêgo, A. L. C., Barreto, M., Schuler, M., Heine, A., Villas, M. V., and Lifschitz, S. (2023). Gerenciamento de dados de redes sociais com análise de redes e modelagem de tópicos. In *Anais do Simpósio Brasileiro de Banco de Dados (SBBD)*.

Costa, R. L. H., Soares, T. S., Lunardi, G. M., Valle, P. H. D., and Silva, W. (2024). Professionals' perceptions of the interaction between user experience and machine learning. In *20th Brazilian Symposium on Information Systems*, pages 1–9.

de Almeida Neto, J. A. and de Melo, T. (2023). Identificação de temas em comentários de restaurantes usando bert e modelos de linguagem generativa. In *Anais do Simpósio Brasileiro de Banco de Dados (SBBD)*.

de Oliveira, K. B., Lunardi, G. M., and Silva, W. (2025). Avaliação de sentimentos de aplicativos: Uma comparação entre modelos de linguagem de grande escala. In *Escola Regional de Banco de Dados (ERBD)*, pages 145–148. SBC.

Ding, B., Qin, C., Zhao, R., Luo, T., Li, X., Chen, G., Xia, W., Hu, J., Luu, A. T., and Joty, S. (2024). Data augmentation using llms: Data perspectives, learning paradigms and challenges.

Feng, F., Yang, Y., Cer, D., Arivazhagan, N., and Wang, W. (2022). Language-agnostic bert sentence embedding.

Gartner, Inc. (2025). Gartner data & analytics summit 2025 orlando: Destaques do terceiro dia. Orlando, Flórida, 5 de março de 2025.

Moreira, L. S., Lunardi, G. M., de Oliveira Ribeiro, M., Silva, W., and Basso, F. P. (2023). A study of algorithm-based detection of fake news in brazilian election: Is bert the best. *IEEE Latin America Transactions*, 21(8):897–903.

Murthy, D., Kurz, S. E., Anand, T., Hornick, S., Lakuduva, N., and Sun, J. (2025). Examining hurricane–related social media topics longitudinally and at scale: A transformer-based approach. *PLOS ONE*, 20(1).

Silva, M. O., Oliveira, G. P., Costa, L. G. L., and Pappa, G. L. (2023). Evaluating domain-adapted language models for governmental text classification tasks in portuguese. In *Anais do Simpósio Brasileiro de Banco de Dados (SBBD)*.

Siqueira, V. X., Costa, R. L. H., Soares, T. S., Lunardi, G. M., and Silva, W. (2024). Dataset anotado de sentimentos a partir de comentários de aplicativos móveis. In *Dataset Showcase Workshop (DSW)*, pages 65–76. SBC.

Soares, T. S., Costa, R. L. H., Soares, E., Calderon, I., Lunardi, G. M., Valle, P. H. D., Guedes, G. T., and Silva, W. (2025). Machine learning-assisted tools for user experience evaluation: A systematic mapping study. *Simpósio Brasileiro de Sistemas de Informaçao (SBSI)*, pages 379–388.

Sujana, Y. and Kao, H.-Y. (2023). Lida: Language-independent data augmentation for text classification. *IEEE Access*, 11:10933–10945.

Wankhade, M., Rao, A. C. S., and Kulkarni, C. (2022). A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*, 55(7):5731–5780.