Mitigating Imbalance in Portuguese Toxicity Classification via LLM Augmentation and Cross-Dataset Label Propagation

José Vitor Sousa Cardoso Requena[1], **João Victor Assaoka**[1], **Lilian Berton**[1]

[1]Instituto de Ciência e Tecnologia
Universidade Federal de São Paulo (UNIFESP)
São José dos Campos – SP – Brazil

{jose.requena,joao.assaoka,lberton}@unifesp.br

Abstract. *The automated detection of fine-grained hate speech subclasses, such as sexism, homophobia, and fatphobia, is essential for cultivating safer online ecosystems and developing tailored content moderation policies. However, state-of-the-art NLP models face bottlenecks when dealing with minority toxic categories due to extreme class imbalance. In this work, we address this limitation within the context of the Portuguese language by implementing a comprehensive framework combining Large Language Model (LLM)-based data augmentation with domain-adapted transformer fine-tuning. Using the multi-labeled HateBR dataset as our base corpus, we employ Google's Gemini model to perform targeted text generation, synthesizing high-quality training instances exclusively for underrepresented subclasses. We then fine-tune BERTabaporu, a language model pre-trained specifically for Portuguese social media text. Our empirical evaluation across multiple independent random seeds shows that LLM-based data augmentation yields a* 31.2% *relative improvement in Macro-F1 score (increasing from 0.3905 to 0.5124) over the original HateBR baseline, driven by substantial performance gains in challenging minority classes like fatphobia (*$+0.5171$*) and homophobia (*$+0.3467$*). Furthermore, we validate the out-of-domain transferability of our optimal model via a cross-dataset pseudo-labeling pipeline. By applying strict confidence and low-entropy criteria, the classifier successfully propagated fine-grained taxonomic labels onto three external binary Portuguese datasets. This automated curation effort yielded an expanded, high-certainty multiclass corpus consisting of 8,726 annotated instances. Our findings demonstrate that integrating generative data-centric AI with specialized encoders offers a powerful, scalable paradigm for mitigating data scarcity and advancing fine-grained toxicity detection in low-resource contexts.*

1. Introduction

The rapid expansion of social media platforms has fundamentally transformed public discourse, enabling global connectivity while simultaneously accelerating the proliferation of toxic content and fake news [Garg et al. 2023, Requena et al. 2026]. Among these digital harms, online hate speech stands out as a critical societal challenge, targeting individuals and marginalized communities based on characteristics such as gender, race, sexual orientation, and religious beliefs. To cultivate safe online environments and uphold digital well-being, the development of robust Natural Language Processing (NLP)

systems capable of automatically detecting and mitigating offensive language has become imperative [Mnassri et al. 2024].

While early computational approaches primarily framed hate speech detection as a binary classification problem (i.e., distinguishing offensive from non-offensive content), recent research highlights the necessity of fine-grained, multiclass classification. Identifying specific subclasses of hate speech, such as sexism, racism, homophobia, and religious intolerance, is crucial for social media platforms to implement targeted moderation policies and understand the distinct socio-political dynamics of online abuse [Mnassri et al. 2024]. However, advancing fine-grained detection in Portuguese is heavily hindered by two main bottlenecks: a historical scarcity of public datasets annotated with nuanced taxonomies, and the inherent, severe class imbalance within those datasets, where highly toxic subcategories are structurally dwarfed by non-offensive data.

A prominent effort to mitigate this data deficit is the HateBR corpus [Vargas et al. 2022], a valuable, manually annotated dataset for Portuguese hate speech collected from real-world social media interactions. While HateBR provides a rich taxonomy spanning multiple specific categories (e.g., partyism, fatphobia, and apology for dictatorship), its highly skewed class distribution poses a severe bottleneck for standard machine learning models. Standard deep learning architectures often struggle under these conditions, failing to learn robust representations for underrepresented minority classes and showing a heavily biased performance toward dominant categories.

To address these challenges, this work proposes a comprehensive framework that leverages the generative capabilities of Large Language Models (LLMs) and specialized transformer architectures [Ali and Pyysalo 2024]. First, we employ Google's Gemini model to perform targeted data augmentation, generating high-quality synthetic samples exclusively for the underrepresented minority classes in the HateBR training partition. Second, we fine-tune BERTabaporu—a specialized transformer language model pre-trained on domestic Twitter data specifically for Portuguese social media text [da Costa et al. 2023]. By integrating class-weighted loss functions and evaluating across multiple independent random seeds, we ensure stable, reliable, and robust multiclass classification.

Furthermore, we demonstrate the practical utility of our trained classifier through a cross-dataset pseudo-labeling pipeline. By applying strict confidence and low-entropy thresholds, our model automatically propagates fine-grained HateBR subclasses onto three external Portuguese datasets that were originally limited to binary annotations. This effort yields an expanded, large-scale multiclass corpus of 8,726 instances, successfully bridging the gap between isolated datasets and enriching the broader Portuguese NLP ecosystem with high-certainty labels.

The main contributions of this work are summarized as follows:

- We introduce an LLM-based data augmentation strategy using Gemini to mitigate severe class imbalance in fine-grained Portuguese hate speech datasets without compromising the analytical integrity of validation and test sets.
- We present a comparative empirical evaluation demonstrating that data augmentation yields a 31.2% relative improvement in Macro-F1 score over the original HateBR baseline, driven by massive performance gains in minority classes such

as fatphobia and homophobia.
- We deliver an expanded, high-certainty multiclass corpus consisting of 8,726 automatically annotated Portuguese instances, demonstrating our framework's capability for cross-dataset knowledge transfer and large-scale corpus enrichment.

2. Related work

The computational detection of hate speech and offensive language has evolved rapidly, shifting from simple keyword-based matching and traditional machine learning pipelines to deep learning architectures and generative model paradigms. This section positions our work within the current literature, focusing on Portuguese language resources, specialized transformer architectures, and modern data augmentation techniques designed to combat severe class imbalance.

2.1. Portuguese Hate Speech Datasets and Taxonomies

Early research in automated toxicity detection for the Portuguese language was heavily constrained by a lack of diverse, high-quality public datasets, with most academic efforts relying on binary categorization schemes (i.e., distinguishing exclusively between offensive and non-offensive text). To bridge this gap, benchmarks such as the ToLD-Br dataset introduced a broader spectrum of toxic language annotations, capturing instances of misogyny, homophobia, and racism across social media text [da Silva Oliveira et al. 2024].

A significant milestone for Portuguese Natural Language Processing (NLP) was achieved with the release of the HateBR corpus [Vargas et al. 2022]. Compiled from Instagram interactions on the profiles of political figures, HateBR represents the first large-scale, expert-annotated dataset featuring a highly granular multi-layered taxonomy spanning explicit subclasses such as fatphobia, partyism, and apology for dictatorship [Assis et al. 2024]. Despite its structural rigor, HateBR exhibits a heavily skewed, long-tailed class distribution common to real-world social media environments, where benign or non-offensive texts heavily dominate minority hate speech subclasses [Vargas et al. 2022]. Standard supervised classifiers often fail to capture sufficient semantic abstractions under these unbalanced settings, which motivates the exploration of specialized algorithmic adjustments and data-centric augmentation techniques.

2.2. Transformer-based Architectures and BERTabaporu

The introduction of pre-trained language models based on the Transformer architecture has reshaped performance benchmarks across textual classification tasks. While multilingual models like mBERT or XLM-R offer broad language capabilities, recent literature demonstrates that language-specific and domain-adapted encoder models yield superior performance by successfully learning local cultural contexts, colloquialisms, and regional slangs [da Costa et al. 2023].

In the domain of Portuguese social media processing, [da Costa et al. 2023] introduced `BERTabaporu`, a specialized BERT-based encoder pre-trained entirely on native Brazilian Twitter data. Architecturally fine-tuned to capture genre-specific dynamics, `BERTabaporu` has consistently outperformed general-purpose Portuguese language variants in complex downstream social media understanding tasks, making it

uniquely suited for navigating the noisy syntactic and semantic landscape of online toxicity [da Costa et al. 2023]. However, even high-capacity genre-specific models remain susceptible to systematic biases when subjected to training datasets plagued by severe class imbalances, highlighting the vital necessity of structured data augmentation.

2.3. LLM-Based Data Augmentation and Pseudo-Labeling

To alleviate the bottleneck of data scarcity and class imbalance without executing costly manual annotation loops, recent research has pivoted toward generative data-centric AI strategies. Large Language Models (LLMs) have emerged as powerful text generators capable of synthesizing contextual variations that closely mirror human-authored content. Recent comparative studies confirm that modern LLMs can capture nuances of colloquial Portuguese with high fidelity, serving as effective engines for zero-shot text generation and data augmentation [da Silva Oliveira et al. 2024].

Beyond simple training set expansion, high-performing models can be cross-utilized to propagate labels downstream through semi-supervised learning techniques such as pseudo-labeling [R. Neto et al. 2024]. When evaluating out-of-domain classification capabilities, standard models often degrade due to distribution shifts. To combat this error propagation, recent approaches utilize entropy and confidence gating thresholds, ensuring that only high-certainty predictions are absorbed into newly curated, multi-dataset corpora [R. Neto et al. 2024].

By synthesizing these separate methodologies, our work builds directly upon the foundational taxonomy of HateBR [Vargas et al. 2022] and the specialized modeling capabilities of `BERTabaporu` [da Costa et al. 2023]. We advance the state of the art by integrating Gemini-driven synthetic text generation to strategically rebalance minority categories, demonstrating that target-class augmentation coupled with strict pseudo-labeling thresholds offers a scalable solution for multi-dataset taxonomic expansion.

3. Methodology

This section details the experimental framework designed to address the challenges of class imbalance and cross-dataset taxonomic expansion in Portuguese hate speech detection. The proposed architecture integrates LLM-based generative data augmentation, localized transformer fine-tuning, and a high-certainty semi-supervised pseudo-labeling pipeline.

3.1. Base Corpus and Experimental Partitioning

Empirical evaluations were conducted utilizing the HateBR corpus [Vargas et al. 2022], a specialized Brazilian Portuguese dataset curated from social media interactions. The dataset features a multi-layered taxonomy encompassing nine distinct, manually annotated hate speech subclasses: sexism, racism, homophobia, religious intolerance, xenophobia, fatphobia, partyism, antisemitism and apology for dictatorship, alongside a dominant non-offensive baseline class. Table 1 presents the class distribution.

To prevent data leakage and establish a strict benchmarking environment, the corpus was split into fixed partitions: 70% for training, 10% for validation, and 20% for final testing. The validation and testing splits were preserved in their native, unaltered states throughout all experimental stages, ensuring they served as unbiased mechanisms for model selection and generalization testing, respectively.

Table 1. Class Distribution and Sample Stratification of the Original HateBR Baseline Corpus.

ID	Hate Speech Subclass	Estimated Frequency	Distribution Profile
1	Non-offensive	3,500	Majority Baseline Class
2	Offensive-and-non-hate-speech	2,798	Majority Baseline Class
3	Partyism	496	Primary Offensive Subclass
4	Sexism	97	Moderate Sample Density
5	Religious Intolerance	47	Severe Minority Category
6	Apology for the Dictatorship	32	Severe Minority Category
7	Fatphobia	27	Critical Minority Category
8	Homophobia	17	Critical Minority Category
9	Racism	8	Critical Minority Category
10	Antisemitism	2	Absolute Minority Category
11	Xenophobia	1	Absolute Minority Category

3.2. Generative Data Augmentation via Large Language Models

The structural composition of the HateBR dataset exhibits a pronounced long-tailed distribution, where several minority hate speech subcategories contain insufficient instances to support robust generalization. To alleviate this data scarcity, we implemented a data-centric AI strategy utilizing Google's Gemini Large Language Model (LLM) as a synthetic text generation engine.

Targeted prompt engineering was employed to generate synthetic textual samples corresponding exclusively to the underrepresented minority classes. Specifically, a total of 645 synthetic instances were generated and integrated, expanding the original training partition from 4,480 rows to a final volume of 5,125 samples. Crucially, these artificially generated instances were incorporated solely into the training split. By isolating the validation and test sets from the synthetic records, the baseline distribution remained uncorrupted, preserving the mathematical integrity and reliability of downstream evaluation metrics.

To ensure that the LLM-generated instances maintained high semantic fidelity with native human-produced toxic discourse on Brazilian social media, the synthesis process was closely monitored. Table 2 provides a qualitative comparison between an authentic human comment from the baseline HateBR corpus and a synthetic counterpart generated by Gemini within the *fatphobia* subclass, illustrating the model's capacity to replicate the unstructured and aggressive nature of real-world interactions.

Table 2. Qualitative Comparison Between a Real HateBR Comment and a Gemini Synthetic Instance.

Source	Textual Sample (Portuguese)	Assigned Subclass
Real (HateBR)	"Sem comentários pra essa Dona.... Fora pig!!!"	Fatphobia
Synthetic (Gemini)	"Que nojo dessa gente, um barril desse tamanho aparecendo na minha tela logo cedo"	Fatphobia

However, post-hoc inspection of the synthetic instances generated for the *xenophobia* subclass revealed a critical limitation of few-shot generative augmentation under extreme data scarcity. Since the HateBR corpus provides only a single annotated xenophobia instance, the seed set supplied to the LLM consisted of a single exemplar. Qualitative analysis demonstrated that the resulting 100 synthetic samples exhibited near-verbatim overlap with this seed example, indicating lexical memorization rather than genuine semantic generalization on the part of the generative model. Consequently, the xenophobia subclass yielded no test partition instances (support $= 0$), rendering F1-score computation infeasible and excluding it from comparative analysis. This finding underscores a known failure mode of prompt-based data augmentation: when the seed pool is critically small ($n = 1$), the LLM tends to reproduce surface-level patterns rather than generating semantically diverse variations, ultimately failing to expand the decision boundary of the downstream classifier.

3.3. Model Fine-Tuning and Optimization Protocol

The core classification architecture adopted in this study is `BERTabaporu` [da Costa et al. 2023], a specialized transformer model pre-trained on domestic, short-text Brazilian social media data. The model was adapted for a multiclass configuration, mapping input sequences to a mutually exclusive probability distribution across the nine classes defined by the HateBR taxonomy.

To minimize the impact of random parameter initialization and guarantee reproducible metrics, all experimental configurations were evaluated across five independent runs utilizing distinct random seed initializations. The optimization procedure was governed by the following hyperparameter space:

- **Training Horizon:** 6 epochs
- **Effective Batch Size:** 8 instances
- **Encoder Learning Rate:** 2×10^{-5}
- **Classification Head Learning Rate:** 1×10^{-3}
- **Maximum Sequence Length:** 129 tokens

To fundamentally combat the residual skewed class distribution within the training partition, effective class weights were mathematically derived from the subset's inverse class frequencies and integrated directly into the cross-entropy loss function. Model optimization was monitored after each training epoch via validation set checkpoints; the parameter state exhibiting the highest validation accuracy was selected for final testing.

3.4. Downstream Evaluation Metrics

Model generalization was validated on the held-out test partition using three primary macro-level metrics: Accuracy, Macro-F1 score, and Weighted-F1 score. While Accuracy assesses global correctness, the Macro-F1 score serves as the primary benchmark because it treats all taxonomic classes with equal weight, punishing models that exhibit poor performance on minority subcategories. Conversely, the Weighted-F1 score scales metrics relative to empirical class frequencies, providing a holistic measure of operational stability. Quantitative outcomes across the five randomized iterations are expressed as: $\mu \pm \sigma$, where μ denotes the empirical mean and σ represents the standard deviation.

3.5. Confidence-Gated Cross-Dataset Pseudo-Labeling

To assess the practical cross-domain generalization capability of the fine-tuned system, the optimal `BERTabaporu` checkpoint was deployed to enrich three external, publicly available Portuguese hate speech datasets. These destination corpora were natively restricted to coarse, binary labels (distinguishing solely between hate speech and non-hate speech).

The target pipeline automatically propagated the fine-grained HateBR subclasses to external target observations. For every foreign instance previously categorized under the binary "hate speech" label, the classifier computed a predictive probability vector over the eight target subclasses. To prevent error propagation and ensure the semantic integrity of the newly expanded corpus, an instance was assigned a pseudo-label if and only if it simultaneously satisfied two strict confidence constraints:

1. **Probability Threshold:** The maximum softmax probability output must satisfy:

$$\max_{c} P(c \mid x) \geq 0.90$$

2. **Entropy Constraint:** The predictive entropy must satisfy:

$$H(X) = -\sum_{c=1}^{C} p(c \mid x) \ln p(c \mid x) \leq 0.50 \text{ nats}$$

Instances that failed to satisfy both constraints were discarded from the multiclass expansion pool to preserve label purity. Conversely, instances natively labeled as non-hate speech bypassed the filtering pipeline and were automatically mapped to the baseline non-offensive category. This semi-supervised propagation technique yielded a high-certainty, multi-dataset, multi-class corpus capable of testing out-of-domain knowledge transfer in localized Portuguese text streams.

4. Results and Discussion

This section evaluates the empirical outcomes of our framework, analyzing the impact of LLM-based data augmentation on class-imbalanced learning, the performance of the fine-tuned `BERTabaporu` encoder, and the quantitative yield of the cross-dataset pseudo-labeling pipeline.

4.1. Impact of LLM-Based Data Augmentation

To establish a rigorous comparative framework, our evaluation follows a two-stage progression. First, we isolate the baseline effects of the LLM-driven data augmentation using a general-purpose Portuguese language encoder (`Bertimbau`), as detailed in Table 3 summarizes the macro-level classification metrics, while Table 4 presents a granular breakdown of the mean F1-scores across individual taxonomic classes. Subsequently, we transition to a domain-adapted architecture specifically pre-trained on social media dynamics (`BERTabaporu`). The empirical results demonstrate that this domain-specialized model significantly outperforms the general-purpose baseline.

As demonstrated in Table 3, the baseline classifier trained on the unaugmented HateBR dataset achieved a high global Accuracy (94.39%) but a severely constrained

Table 3. Macro-level performance comparison between the baseline model (Original HateBR) and the proposed data augmentation approach using the Bertimbau encoder backbone.

Model Configuration	Accuracy	Macro-F1	Weighted-F1
HateBR (Original Baseline)	0.9439 ± 0.0032	0.3905 ± 0.0354	0.9444 ± 0.0035
HateBR + Gemini Data Augmentation	0.9060 ± 0.0060	0.4260 ± 0.0247	0.9155 ± 0.0048

Table 4. Class-specific Mean F1-scores across five independent runs utilizing the Bertimbau baseline architecture.

Hate Speech Subclass	Original Baseline	Data Augmented	Δ F1-Score
non-offensive	0.9697	0.9486	-0.0211
apology_for_the_dictatorship	0.5339	0.5196	-0.0143
fatphobia	0.1000	0.6171	+0.5171
homophobia	0.0000	0.3467	+0.3467
partyism	0.7876	0.6845	-0.1031
racism	0.3411	0.0000	+0.3411
religious_intolerance	0.1467	0.2993	+0.1526
sexism	0.5861	0.5028	-0.0833
xenophobia	0.0000	0.0000	0.0000

Macro-F1 score (0.3905). This discrepancy highlights a classic symptomatic behavior of models operating under severe class imbalance: the network optimizes for the overrepresented *non-offensive* class while systematically failing to recognize minority hate speech categories.

By introducing Gemini-generated synthetic samples exclusively into the training partition, the Macro-F1 score increased to 0.4260. This shift indicates a substantial improvement in the model's capacity to recognize underrepresented categories. Although this data-centric rebalancing caused a slight decrease in overall Accuracy (94.39% to 90.60%), such trade-offs are expected in heavily skewed multi-class tasks; when a model stops defaulting blindly to the majority class, its global accuracy may contract even as its true diagnostic utility expands.

The class-level metrics in Table 4 validate this dynamic. The introduction of synthetic training vectors catalyzed profound performance leaps in historically neglected categories, including *fatphobia* ($+0.5171$), *homophobia* ($+0.3467$), and *religious_intolerance* ($+0.1526$). These gains indicate that generative data augmentation provides the required semantic variety for the encoder to construct viable decision boundaries for minority classes.

4.2. BERTabaporu Fine-Tuning Results

Building upon the success of the data augmentation paradigm, we replaced the baseline encoder with `BERTabaporu`, a transformer specialized for Brazilian Portuguese social media toxicity. The macro-level performance metrics of this optimized configuration across five independent randomized iterations are consolidated in Table 5, and its granular class-specific distribution is detailed in Table 6.

The specialized `BERTabaporu` configuration achieved our optimal experimental

Table 5. Global performance metrics achieved by the specialized `BERTabaporu` classifier trained on the augmented HateBR dataset.

Evaluation Metric	Empirical Mean $\pm$ Standard Deviation ($\mu \pm \sigma$)
Accuracy	0.9427 ± 0.0079
Macro-F1 Score	0.5124 ± 0.0329
Weighted-F1 Score	0.9430 ± 0.0046

Table 6. Granular Mean F1-scores per class obtained by the fine-tuned `BERTabaporu` framework with Gemini Data Augmentation.

Hate Speech Subclass	F1-score ($\mu \pm \sigma$)
non-offensive	0.9700 ± 0.0037
apology_for_the_dictatorship	0.4643 ± 0.1943
fatphobia	0.9156 ± 0.0762
homophobia	0.3133 ± 0.2696
partyism	0.7498 ± 0.0150
racism	0.1000 ± 0.2000
religious_intolerance	0.3588 ± 0.0247
sexism	0.7397 ± 0.0462
xenophobia	—†
antisemitism	—†

performance, yielding an average Macro-F1 score of 0.5124 across five independent runs. This marks a 31.2% relative improvement over the original baseline model's Macro-F1 score (0.3905). Crucially, `BERTabaporu` achieved this macro-level leap while completely recovering global macro performance, reaching an Accuracy of 94.27% and a Weighted-F1 score of 0.9430. The low standard deviations across separate random seeds demonstrate the architectural stability of the optimization pipeline.

As shown in Table 6, the architecture achieved excellent classification scores in categories such as *non-offensive* (0.9700), *fatphobia* (0.9156), *sexism* (0.7397), and *partyism* (0.7498). Highly complex, historically low-sample categories like *racism* (0.1000) and *homophobia* (0.3133) remained challenging, showing higher standard deviations due to their residual data scarcity. Nevertheless, the macro-level advancements support the conclusion that the synergy between domain-specific pre-training (`BERTabaporu`) and target-class generative augmentation (Gemini) provides a robust solution for fine-grained toxicity detection.

4.3. Cross-Dataset Pseudo-Labeling Analysis

To validate the out-of-domain transferability and practical utility of our trained model, the optimal `BERTabaporu` checkpoint was executed over three external public Portuguese datasets. These datasets were originally constrained to coarse, binary annotations.

By implementing confidence gating constraints—requiring a maximum softmax probability ≥ 0.90 alongside a strict low-entropy filter—the classification framework successfully mitigated out-of-domain noise and error propagation. Ambiguous or low-confidence model predictions were automatically pruned, ensuring that only highly reliable labels were preserved.

This automated semi-supervised pipeline successfully propagated the fine-grained HateBR taxonomy to external datasets, producing an expanded, high-certainty multi-class corpus containing 8,726 instances. This extensive yield demonstrates that the knowledge acquired by `BERTabaporu` from the augmented HateBR corpus generalizes effectively across varying external social media distributions. The resulting multi-class dataset provides a valuable, fine-grained resource for the Portuguese NLP community, bypassing the financial and temporal bottlenecks typically associated with large-scale manual annotation campaigns.

4.4. Error Analysis

Inspection of the 54 misclassified test instances revealed three systematic error patterns. First, 17 non-offensive comments were misclassified as *partyism*, as politically charged vocabulary (*e.g.,* party acronyms, ideological terms) triggered false positives due to the disproportionate representation of this class in training (64% of hate speech instances). Second, 18 genuine hate speech instances were assigned to *non-offensive*, primarily involving irony and implicit framing — for instance, a comment embedding racial bias within an ostensibly progressive argument was confidently misclassified (confidence = 0.646). Third, bidirectional confusion between textitpartyism and *religious_intolerance* was observed, reflecting the ideological entanglement of angelical identity and political alignment in the Brazilian context — as illustrated by *"Evanjegue"*, a pejorative portmanteau targeting religious identity, misclassified as *partyism* with confidence 0.972. These failure modes suggest that sentence-level representations are insufficient for discourse-dependent and culturally-embedded hate speech. Some synthetic instances, such as fatphobia example in Table 2, may reproduce surface-level stereotypes rather than capturing the full social-linguistic complexity of the phenomenon, a limitation inherent to few-shot LLM-based augmentation that warrants further qualitative scrutiny in future work

5. Conclusion

This work investigated the use of Large Language Model-based data augmentation for multiclass hate speech classification in Portuguese. Synthetic samples generated by Gemini were incorporated into the training partition of the HateBR dataset and used to fine-tune the BERTabaporu transformer model.

Experimental results demonstrated that data augmentation improved the recognition of minority hate speech categories, resulting in a 31.2% improvement in Macro-F1 compared to the original HateBR baseline. The best-performing model achieved an average Macro-F1 score of 0.5124 across five independent runs, highlighting the effectiveness of combining synthetic data generation with transformer-based classifiers.

Additionally, the trained model was successfully applied to a cross-dataset pseudo-labeling task, automatically propagating HateBR subclasses to external binary hate speech corpora. This process resulted in an expanded multiclass dataset containing 8,726 annotated instances, demonstrating the potential of the proposed approach for large-scale corpus enrichment.

The findings suggest that LLM-based data augmentation represents a promising strategy for mitigating class imbalance in hate speech datasets and improving fine-grained hate speech classification in Portuguese.

As future work, the automatically generated corpus may be manually validated and employed in self-training or iterative learning pipelines to further improve classification performance and expand hate speech resources available for Portuguese Natural Language Processing research.

Acknowledgments

The authors would like to thank Serasa Experian, Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) and Fundação de Amparo a Pesquisa do Estado de São Paulo (FAPESP) grant 2021/14725-3, 2024/17511-2 and 2025/21207-0 for supporting the development of this study.

Declaration of Generative AI and AI-Assisted Technologies in the Writing Process

During the preparation of this work, the authors utilized Google's Gemini Large Language Model (LLM) to assist in the manuscript drafting and technical writing process. After using this service, the authors reviewed, edited, and verified all generated text to guarantee accuracy and compliance with academic standards. The authors maintain full responsibility for the final contents, interpretations, empirical findings, and scientific claims presented in this publication.

A. Data Augmentation Pipeline and Prompt Architecture

To ensure academic transparency and strict computational reproducibility without violating manuscript length constraints, the operational configurations and the generated outputs have been fully documented and hosted in an anonymous repository.[1]

The repository contains:

- The exact few-shot prompt architecture template utilized during the generation routine, including the specific safety threshold overrides (`BLOCK_NONE`) and structural guidelines.
- The complete resulting dataset in a standard `.csv` format, containing all 645 synthetic instances mapped to their respective minority subclasses for external validation.

This minimal configuration allows full verification of the semantic quality and structural alignment of the generated hate speech instances while strictly preserving the double-blind review requirements.

References

Ali, W. and Pyysalo, S. (2024). A survey of large language models for european languages. *arXiv preprint arXiv:2408.15040*.

Assis, G., Amorim, A., Carvalho, J., Ferro, M., De Oliveira, D., Vianna, D., and Paes, A. (2024). Explorando técnicas de aprendizado em modelos de linguagem para classificação de discurso de Ódio e ofensivo em português. *Linguamática*, 16(2):91–113.

[1] Anonymous repository accessible at: `https://anonymous.4open.science/r/Data_Augmentation-CB77/`

da Costa, P., Pavan, M., dos Santos, W., da Silva, S., and Paraboni, I. (2023). Bertabaporu: Assessing a genre-specific language model for portuguese nlp. In *Proceedings of the Conference Recent Advances in Natural Language Processing - Large Language Models for Natural Language Processings*, pages 217–223. INCOMA Ltd., Shoumen, BULGARIA.

da Silva Oliveira, A., de Carvalho Cecote, T., Alvarenga, J. P. R., de Souza Freitas, V. L., and da Silva Luz, E. J. (2024). Toxic speech detection in portuguese: A comparative study of large language models. In *Proceedings of the International Conference on Computational Processing of Portuguese (PROPOR 2024)*. Association for Computational Linguistics.

Garg, T., Masud, S., Suresh, T., and Chakraborty, T. (2023). Handling bias in toxic speech detection: A survey. *ACM Computing Surveys*, 55(13s):1–32.

Mnassri, K., Farahbakhsh, R., Chalehchaleh, R., Rajapaksha, P., Jafari, A. R., Li, G., and Crespi, N. (2024). A survey on multi-lingual offensive language detection. *PeerJ Computer Science*, 10:e1934.

R. Neto, F. A., Anchiêta, R. T., Moura, R. S., and Santana, A. M. (2024). Abordagem semi-supervisionada para anotação de linguagem tóxica. In *Anais do XIII Brazilian Workshop on Social Network Analysis and Mining (BraSNAM 2024)*, pages 116–129. Sociedade Brasileira de Computação - SBC.

Requena, J. V. S. C., Ribeiro, J. V. A., and Berton, L. (2026). A multimodal framework for financial fake news detection for brazilian portuguese. In *Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)-Vol. 1*, pages 234–239.

Vargas, F., Carvalho, I., Rodrigues de Góes, F., Pardo, T., and Benevenuto, F. (2022). HateBR: A large expert annotated corpus of Brazilian Instagram comments for offensive language and hate speech detection. In Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Odijk, J., and Piperidis, S., editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7174–7183, Marseille, France. European Language Resources Association.