A Change in Perspective: The Trade-Off Between Perspective API and Custom Models in Classifying Hate Speech in Portuguese

Arthur Buzelin[1], Yan Aquino[1], Pedro Bento[1], Samira Malaquias[1], Wagner Meira Jr[1], Gisele L. Pappa[1]

[1]Departamento de Ciência da Computação – Universidade Federal de Minas Gerais
Belo Horizonte – MG – Brazil

{arthurbuzelin, yanaquino, pedro.bento, samiramalaquias}@dcc.ufmg.br

{meira, glpappa}@dcc.ufmg.br

Abstract. *This paper examines the performance of the Perspective API, developed by Jigsaw, in detecting hate speech in Portuguese. Although the Perspective API supports multiple languages, its performance metrics are often aggregated, obscuring specific details. Our study reveals that the API's AUC-ROC score for Portuguese is significantly lower than for English (0.744 vs. 0.942). To address this, we developed a BERT classifier model trained on a Portuguese Twitter hate speech dataset. Our model, with just 100 messages in it's training set, outperformed the Perspective API. These findings highlight the need for more granular performance metrics and suggest that custom models may offer better solutions for specific languages.*

1. Introduction

Perspective API is a tool designed to identify and mitigate toxic language online [Lees et al. 2022]. Using advanced machine learning and Natural Language Processing (NLP) models, Perspective API analyzes textual content to detect various forms of harmful speech, including threats, insults, and hate speech. It is considered state-of-the-art for detecting toxicity, and used by multiple platforms, such as Reddit, The New York Times, The Wall Street Journal, and EL PAÍS.

Despite its widespread adoption and claimed multilingual support, including Portuguese, the actual performance of the Perspective API in different languages remains unclear. The official documentation and associated research papers often report performance metrics by aggregating data from multiple languages, within a multilingual dataset. This aggregation conceal the individual performance metrics for Portuguese, making it difficult to evaluate the API's effectiveness in this specific language. The lack of transparency in language-specific performance metrics raises concerns about the API's reliability when applied to non-English texts.

The widespread acceptance of Perspective API as a leading tool for hate speech detection combined with its claimed support for Portuguese, suggests that professionals may readily adopt it for work in Portuguese-speaking contexts. However, if the API's performance in Portuguese is not on par with its performance in English or the aggregated results, this could lead to inaccurate analyses and conclusions, particularly in fields like computational social sciences, where precise language detection is critical. The potential for misleading results is especially concerning when the tool's reliability in Portuguese is taken for granted based on its performance in other languages.

This paper addresses this gap by evaluating the performance of the Perspective API in detecting hate speech specifically in Portuguese. It also assesses the feasibility of developing a custom hate speech detection tool tailored for Portuguese. To guide our investigation, we formulated the following research questions:

RQ1: How well does the Perspective API perform when detecting hate speech in Portuguese?

RQ2: For Portuguese, is it more effective and efficient to use a custom-made tool rather than relying on existing solutions like the Perspective API? If so, how much effort would it take to build it?

To address these questions, we evaluated the Perspective API's metrics using a Portuguese Twitter hate speech dataset. We then compared it to metrics obtained in a similar English dataset regarding classification, date of collection, and content. Our findings revealed that the Perspective API's performance in Portuguese was significantly worse than in English. Based on this insight, we developed our version of a BERT classifier to detect hate speech in Portuguese. Remarkably, with just 100 messages, the BERT model outperformed the Perspective API in detecting hate speech in Portuguese. In contrast, the BERT model trained with the English dataset did not surpass the Perspective API's performance.

2. Related Works

This section reviews studies related to hate speech detection models and language-specific performance comparisons. oportunities

2.1. Model Comparisons

Multilingual transformer models, such as BERT and its variants, have gained significant attention in hate speech detection across various languages. For instance, [Roy et al. 2021] demonstrated the superiority of fine-tuned transformer models in handling multilingual data, showcasing their effectiveness compared to more generalized approaches like the Perspective API. This highlights the potential of specialized models to outperform broader, one-size-fits-all solutions.

Another noteworthy contribution is by [Kennedy et al. 2020], who introduced a hybrid approach that combines faceted Rasch measurement with multitasking deep learning. This methodology enhances both the interpretability and precision of hate speech detection by integrating traditional psychometric techniques with advanced deep learning models. Compared to the Perspective API, which relies on more generalized algorithms, Kennedy et al.'s approach offers a more nuanced understanding of linguistic variations and the intensity of hate speech.

2.2. Language-Specific Comparisons

In Perspective's introductory paper [Lees et al. 2022], developers reported AUC-ROC scores of 0.98 for English, 0.91 for Russian, and 0.87 for a group of ten other languages.

These results highlight a disparity in the API's effectiveness across languages, raising concerns about its applicability in non-English contexts.

Further studies have confirmed these concerns were relevant. For instance, [Nogara et al. 2024] analyzed the use of the Perspective API in German and found that the API tends to classify German texts as significantly more toxic than their English counterparts. This finding underscores the potential biases and inaccuracies that arise when applying the API to languages other than English, highlighting the need for further investigation into its multilingual capabilities.

The seminal study of the use of the Perspective API's in Portuguese was conducted by [Kobellarz and Silva 2022]. They compared identical texts in Portuguese and English using the API and concluded that it performs better when analyzing texts in their original language. This suggests that the Perspective API may be less effective in detecting nuances in translated or non-native language content.

Building upon this study, [Lima et al. 2024] developed a manually labeled dataset of toxic messages in Portuguese and evaluated the API against this dataset. Their findings revealed significant discrepancies, emphasizing the need for the API to undergo more focused training on Portuguese-language content to improve its accuracy and reliability in detecting hate speech.

Additionally, [Silva et al. 2023] proposed standardized datasets and benchmarks for sentiment analysis in English, specifically addressing the challenges of automating the development process. While their focus was on English, the methods and standards they advocate could provide valuable insights for improving the Perspective API's performance in other languages, including Portuguese.

2.3. Research Gap

Despite the widespread use and validation of the Perspective API for hate speech detection in various languages, a significant gap remains in its performance evaluation for less commonly studied languages like Portuguese. Previous research has shown the API's strong performance in English and other major languages, demonstrated by high AUC-ROC scores and robust metrics. However, detailed assessments for less-represented languages in its training datasets are lacking.

To address these gaps, we conducted focused evaluations of the Perspective API's performance for individual languages. Our study highlights the advantages of developing custom models tailored to specific languages, such as Portuguese, offering more accurate and reliable hate speech detection. This emphasizes the need to consider custom solutions alongside existing multilingual models to improve the effectiveness of hate speech detection across diverse languages.

3. Methodology

In this section, we discuss the dataset selection, Perspective API evaluation, and the BERThs models fine-tuning.

3.1. Dataset

Our analysis required a Portuguese hate speech dataset and a similar English dataset, for the purpose of an unbiased comparison. Instead of manually labeling messages, which can be costly and prone to errors, we opted to use two well-known Twitter hate speech datasets: the Hierarchically-Labeled Portuguese Hate Speech Dataset [Fortuna et al. 2019] and the Automated Hate Speech Detection and the Problem of Offensive Language dataset [Davidson et al. 2017].

Both datasets were created using the same methodology for classifying messages. This involved identifying and mining accounts likely to post hate speech-related tweets in 2017. The tweets were then classified as either containing hate speech or not, which matches the output of the Perspective API.

The original Portuguese and English datasets vary significantly in size and proportion of hate speech messages. The Portuguese dataset includes 5,934 non-toxic messages and 1,607 toxic messages, resulting in a ratio of approximately 3.7 non-toxic messages per toxic message. On the other hand, the English dataset initially consisted of 25,000 classified tweets, with 3,280 non-toxic messages and 21,720 toxic messages.

For a fair comparison of classification scores between the two datasets, we balanced their proportions by using the Portuguese dataset as the baseline, since this will be the main object of our study. By selecting a random sample of messages from the English dataset that reflected the same proportion, we leveraged a final English dataset consisting of 3,280 non-toxic messages and 886 toxic messages, with the same ratio of non-toxic to toxic messages of approximately 3.7.

3.2. Comparing Perspective API results

To compare the models of Perspective for English and Portuguese, we selected random samples of messages and analyzed them for toxicity using the Perspective API. We focused on the *Toxicity* attribute, which is widely used in literature due to its robustness and compatibility with both datasets under examination. The analysis was conducted in June 2024, and the Perspective API provided toxicity scores for each sample in both datasets.

Each message was assigned a toxicity score ranging from 0 to 1, where 0 represents a very low probability of toxicity and 1 indicates a very high probability. To ensure the most precise possible comparison, we optimized the threshold for toxicity classification by maximizing the F1 score for each dataset individually. The optimal threshold was determined to be 0.48 for the Portuguese dataset and 0.59 for the English dataset, reflecting the different calibrations needed by the two languages.

3.3. BERThs (BERT hate speech) Model

This section shows how we fine-tuned our own BERT classifier for hate speech detection, namely BERThs. BERThs was fine-tuned using both a Portuguese and an Englih dataset.

Initially, the goal of the model, particularly the Portuguese one, was not to achieve the highest possible accuracy, but to be easy to replicate. This will help us show whether a simple fine-tuned model may be more effective than the Perspective API in Portuguese.

For fine-tuning the BERThs-Pt, we used BERTimbau [Souza et al. 2020] as the base model, as it is pre-trained in Portuguese and better suited for our task. Given the small size of the annotated corpus, we fine-tuned and evaluated the model 30 times using different randomized non-overlapping stratified sets: training, validation, and test sets, comprising 80%, 10%, and 10% of the labeled dataset. Each split maintained the original class distribution of approximately 21.3% toxic messages and 78.7% non-toxic messages. The same test sets were used to evaluate the Perspective API. This approach ensured robustness and prevented issues such as training on an all-toxic set of messages, which could lead to unreliable results.

To determine the minimum number of messages needed for our classifier to outperform the Perspective API, initially, only 10 messages from the training set were used for fine-tuning BERTimbau. We incrementally added 10 more messages to the training set after each iteration, until BERThs achieved a better AUC score than Perspective. The AUC metric was chosen because it was the only metric reported for Portuguese in the Perspective API paper. After that, we added 200 new messages to the training set in each subsequent iteration, until all messages were included, highlighting the highest performance our model could achieve.

The fine-tuning was performed using the PyTorch library [Paszke et al. 2019], with the AdamW optimizer [Loshchilov and Hutter 2017] and a learning rate of 5×10^{-6}. The classification thresholds were established based on the output probabilities of the model, defined as the thresholds that yielded the best mean F1-score on our validation set.

For BERThs-En, we employed the BERT uncased model [Devlin et al. 2019], which is optimized for English language processing. The fine-tuning procedure followed the same general approach used for the Portuguese variant, but with specific modifications to account for the superior performance of the Perspective API on English texts. Specifically, instead of gradually increasing the training set by 10 messages and subsequently by 200 messages per iteration, we opted to directly increase the training set by 200 messages in each iteration.

4. Results

This section presents the Perspective API prediction metrics for the English and Portuguese datasets and compares them to BERThs. The models were fine-tuned on an NVIDIA RTX 4090 GPU. As the models were trained 30 times with different data samples, the results in this section present the mean followed by the standard deviation.

4.1. Perspective Performance

Table 1 shows the performance metrics of the Perspective API in the English and Portuguese datasets. It highlights a significant disparity in the model's effectiveness between the two languages, with the English dataset consistently achieving higher scores across all metrics. Notably, the accuracy, precision, recall, F1 score, and AUC-ROC are considerably lower for the Portuguese dataset, suggesting that the model's capability to accurately classify toxic content is compromised in Portuguese.

Note that the F1 score – which serves as a balanced measure of a model's precision and recall in classification tasks – is almost 35 percentage points lower in Portuguese. On top of that, typically, there is a trade-off between the precision and recall metrics; adjusting the threshold to improve one often causes the deterioration of the other. However, in this case, precision and recall are significantly lower for the Portuguese dataset, indicating an overall performance issue. Low precision usually implies in a high number of false positives, while low recall indicates many false negatives.

Table 1. Metrics for the Perspective API model in English and Portuguese.

Metric	Perspective API(En)	Perspective API(Pt)	Difference
Accuracy	**0.901**	0.779	0.122
Precision	**0.813**	0.477	0.336
Recall	**0.744**	0.404	0.340
F1 Score	**0.777**	0.438	0.339
AUC-ROC	**0.942**	0.743	0.199

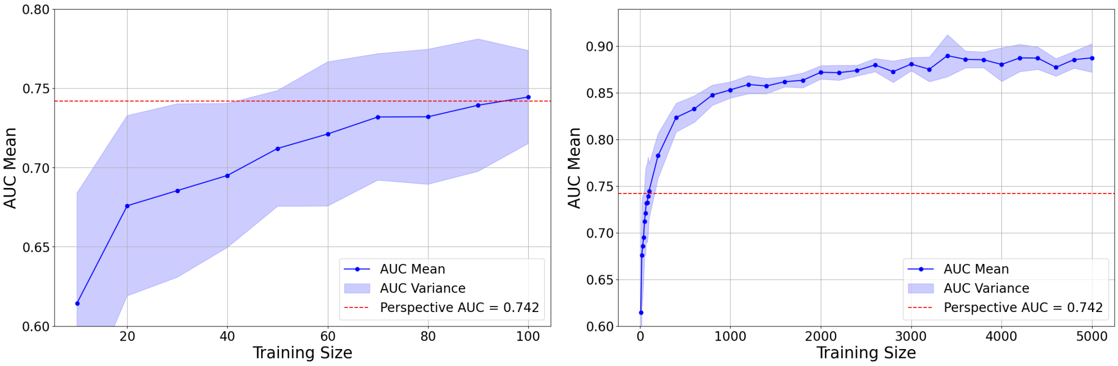

Figure 1. Graph displaying the mean AUC across varying training sizes of the BERThs-Pt model, with the left panel covering up to 100 Twitter posts to assess early performance, and the right panel extending to 5000 posts to evaluate the model's full training potential.

The most concerning results are in the AUC-ROC score, which measures the classification abilities of the Perspective API in its official paper. The Portuguese dataset scores 20 percentage points lower than the English dataset in AUC-ROC. This is both surprising and alarming, given that the multilingual Perspective API is reported to have an AUC-ROC of 0.877, only slightly lower than the English counterpart in the official documentation.

These findings suggest that, while the Perspective API claims to support multiple languages, its performance in Portuguese is substantially lower than in English. This underscores the importance of evaluating multilingual models on a per-language basis to ensure their effectiveness and reliability across different linguistic contexts.

4.2. Evaluating BERThs

BERThs-Pt was evaluated by incrementally increasing the training set by 10 messages at a time. Figure 1 illustrates the model's performance as the number of training messages increased. The red line represents the average performance of the Perspective API on the test set. Observe that our model surpassed the Perspective API in AUC-ROC score with only 100 training messages.

On the other hand, the BERThs-En dataset showed a different result. Even with an extensive training data, leaving aside a small portion for testing and validation, the fine-tuned BERT model still performed worse than the Perspective API, achieving an average AUC-ROC of 0.934 compared to Perspective's 0.942.

These findings suggest that while the Perspective API is an excellent tool, it is not suited for Portuguese. This means creating a custom model can easily surpass the Perspective API's performance with a relatively small amount of labeled data. Therefore, a classifier tailored to the specific linguistic and contextual nuances of Portuguese is better suited for the detection of hate speech.

4.3. Qualitative Analysis

Table 2 shows a comparative analysis of the classification outcomes for the Perspective API and BERThs-Pt on a Portuguese hate speech dataset. The most notable observation from this analysis is that BERThs-Pt misclassified significantly fewer messages (14.3%) when compared to the Perspective API(20.9%), indicating that the BERT model is generally more accurate in discerning the nuances of the text. This superiority is particularly evident in its handling of subtle and context-dependent instances of hate speech, where the Perspective API often struggles. The analysis further reveals that while both models perform well with clear and unambiguous content, they encounter challenges with ambiguous language and contextually rich messages, cases where the BERT model shows a better overall ability to navigate these complexities.

Table 2. Comparison of misclassified and correctly classified messages from the Perspective API and BERThs-Pt. Four random examples from each quadrant are included. In quadrants where only one model missed the true class, the message labels refer to the model that made the mistake. "FP" refers to False Positive, "FN" to False Negative, "TP" to True Positive, and "TN" to True Negative.

	BERThs-Pt was correct	**BERThs-Pt missed**
Perspective was correct	**72.2% of Messages** **(TP) "**Que mulher burra do cacete" **(TP)** "gorda e feia" **(TN)** "Boa semana para todos!" **(TN)** "Não vou orar, sou ateu"	**6.9% of Messages** **(FP)**"Nossa, mas feminismo necessário hoje em dia?" **(FN)** "Se vc bate nessa mulher, além de covarde, com certeza vc gosta de _" **(FP)** "quem é playboy safado fortalece no RT" **(FP)** "Isso sim é tratar gay com indiferença..."
Perspective missed	**13.5% of Messages** **(FN)** "Vai também ser lançado um manual de boas maneiras para lidar com fufas, gays e transsexuais, os chamados LGTB" **(FN)** "as pessoas não entendem que no meio dos refugiados tem inúmeros terroristas, é uma coisa tão óbvia" **(FN)** "E traveco mesmo , mó pirocão" **(FN)** "Você é cheinha, NÃO é gostosa."	**7.4% of Messages** **(FP)** "Pra mim BBB sempre foi uma merda." **(FN)** "feliz dia do não tenho roupa pra sair" **(FN)** "Isso é injusto!" **(FN)** "meritocracia: existe"

Having established that BERThs-Pt generally outperforms the Perspective API, we conducted a quadrant-specific analysis to explore these differences further. The first quadrant represents messages that both methods classified correctly, accounting for 72.2% of the messages. This indicates their effectiveness in handling unambiguous content, as shown in Table 2. The high success rate highlights the capability of both models to manage straightforward cases of hate speech or benign content where linguistic ambiguity is low. However, real-world scenarios often involve more nuanced language, where model differences become more evident.

The second quadrant covers the 6.9% of messages that Perspective correctly classified but BERThs-Pt misclassified. A random sample of four of these messages reveals that they are somewhat ambiguous, making it difficult to determine with certainty whether they were wrongly classified. These cases highlight the challenges of accurately categorizing nuanced and context-dependent language.

The third quadrant, which includes 13.5% of the messages that Perspective misclassified and BERThs-Pt correctly classified. It becomes apparent that Perspective struggles with more complex contexts, particularly when the hate speech is not explicit. The model has particular difficulty with slang or coded language, such as derogatory terms targeting LGBTQ+ individuals. Perspective's limitations in understanding such indirect insults become evident here, suggesting that its generalized training may not sufficiently capture the nuances of the Portuguese language.

Finally, the fourth quadrant, comprising 7.4% of messages, involves cases where both models failed. These messages typically lack sufficient context, making accurate classification challenging. The shared difficulty in this category underscores the challenges of detecting hate speech when language is ambiguous or context is missing.

5. Conclusions

This study assessed the performance of the Perspective API in detecting hate speech in Portuguese, comparing it to English and exploring the potential of custom-trained models. The results show a significant performance gap, with the API achieving an AUC-ROC score of 94.2 in English but only 74.4 in Portuguese. This drop illustrates the limitations of using a generalized multilingual tool for specific languages.

Relying on a model that supports Portuguese yet delivers subpar results poses two main issues. First, research conducted using such a tool may produce inaccurate or misleading outcomes, undermining the validity of the study. Second, researchers from non-English-speaking regions, may feel compelled to conduct their research in English contexts to leverage the more reliable performance of tools like the Perspective API, potentially overlooking important linguistic and cultural nuances.

While the Perspective API excels in English, our study shows it may not be the best choice for Portuguese. A custom BERT model we developed using BERTimbau outperformed the API with only 100 training messages, suggesting that fine-tuning models for specific languages can yield better results in hate speech detection.

In conclusion, while the Perspective API offers robust performance for English, its efficacy in Portuguese is limited. Researchers and practitioners should consider developing custom models tailored to their specific linguistic contexts to achieve more accurate and reliable results.

Acknowledgments

This work was partially funded by CNPq, CAPES, FAPEMIG, and IAIA - INCT on AI.

References

Davidson, T., Warmsley, D., Macy, M., and Weber, I. (2017). Automated hate speech detection and the problem of offensive language. In *Proceedings of the 11th International AAAI Conference on Web and Social Media (ICWSM)*. AAAI.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding.

Fortuna, P., Nunes, S., Soler-Company, J., and Wanner, L. (2019). A hierarchically-labeled portuguese hate speech dataset. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 94–104. Association for Computational Linguistics.

Kennedy, C., Bacon, G., Sahn, A., and Vacano, C. (2020). Constructing interval variables via faceted rasch measurement and multitask deep learning: a hate speech application.

Kobellarz, J. K. and Silva, T. H. (2022). Should we translate? evaluating toxicity in online comments when translating from portuguese to english. In *Anais do Simpósio Brasileiro de Sistemas Multimídia e Web (WebMedia)*, pages 95–104, Porto Alegre, Brazil. Sociedade Brasileira de Computação. In: 28th Simpósio Brasileiro de Sistemas Multimídia e Web (WebMedia), 2022, Curitiba.

Lees, A., Tran, V. Q., Tay, Y., Sorensen, J., Gupta, J., Metzler, D., and Vasserman, L. (2022). A new generation of perspective api: Efficient multilingual character-level transformers.

Lima, Q. L. H., Pagano, S. A., and da Silva, A. (2024). Toxic content detection in online social networks: A new dataset from brazilian reddit communities. In *16th International Conference on Computational Processing of Portuguese (PROPOR 2024)*.

Loshchilov, I. and Hutter, F. (2017). Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Nogara, G., Pierri, F., Cresci, S., Luceri, L., Törnberg, P., and Giordano, S. (2024). Toxic bias: Perspective api misreads german as more toxic.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. (2019). Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Roy, S. G., Narayan, U., Raha, T., Abid, Z., and Varma, V. (2021). Leveraging multilingual transformers for hate speech detection. *ArXiv*, abs/2101.03207.

Silva, M., de Oliveira, V., and Pardo, T. (2023). A sentiment analysis benchmark for automated machine learning applications and a proof of concept in hate speech detection. In *Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 199–206, Porto Alegre, RS, Brasil. SBC.

Souza, F., Nogueira, R., and Lotufo, R. (2020). Bertimbau: Pretrained bert models for brazilian portuguese. In *Proceedings of the 9th Brazilian Conference on Intelligent Systems (BRACIS)*, pages 403–417. IEEE.