HateBRXplain: A Benchmark Dataset with Human-Annotated Rationales for Explainable Hate Speech Detection in Brazilian Portuguese

Isadora Salles
Federal University of Minas Gerais
isadorasalles@dcc.ufmg.br

Francielle Vargas
University of São Paulo
francielleavargas@usp.br

Fabrício Benevenuto
Federal University of Minas Gerais
fabricio@dcc.ufmg.br

Abstract

Nowadays, hate speech technologies are surely relevant in Brazil. Nevertheless, the inability of these technologies to provide reasons (rationales) for their decisions is the limiting factor to their adoption since they comprise bias, which may perpetuate social inequalities when propagated at scale. This scenario highlights the urgency of proposing explainable technologies to address hate speech. However, explainable models heavily depend on data availability with human-annotated rationales, which are scarce, especially for low-resource languages. To fill this relevant gap, we introduce HateBRXplain[1], the first benchmark dataset for hate speech detection in Portuguese, with text span annotations capturing rationales. We evaluated our corpus using mBERT, BERTimbau, DistilBERTimbau, and PTT5 models, which outperformed the current baselines. We further assessed these models' explainability using model-agnostic explanation methods (LIME and SHAP). Results demonstrate plausible post-hoc explanations when compared to human annotations. However, the best-performing hate speech detection models failed to provide faithful rationales.

1 Introduction

Over the last few years, there has been an increased focus on hate speech as a result of the global role related to new technologies and platforms that facilitate orchestrated hateful behavior and incitement to discrimination (Wardle, 2024). Hate speech is defined as a type of offensive language that expresses violence, intolerance, prejudice, or discrimination against an individual or a group based on their social identity (Fortuna and Nunes, 2018; Zampieri et al., 2019), which may be implicitly or explicitly expressed (Poletto et al., 2021; Vargas et al., 2024a).

To mitigate this issue, hate speech detection systems have been developed as effective countermeasures to inhibit offensive and hateful language from being published or spread on the Web and social media. Nonetheless, while there was significant progress in the hate speech research area, for instance, new expert and comprehensive datasets (Vargas et al., 2024a; Guest et al., 2021; Fortuna et al., 2019b; Vargas et al., 2024b), the high performance of deep learning models (Zimmerman et al., 2018; Gambäck and Sikdar, 2017) and transformer architectures (Caselli et al., 2021), these recent models are becoming less interpretable (Tsvetkov et al., 2019) highlighting a lack of transparency posing unwanted risks, such as unintended biases that has recently been identified as a major concern in the area (May et al., 2019).

In modern Natural Language Processing (NLP), a prevalent approach to building hate speech classifiers consists of training on hate speech datasets using fine-tunning Large-Scale Language Models (LLMs), which, according to (Davani et al., 2023), leads to representational biases, such as preferring European American names over African American names (Caliskan et al., 2017), associating words with more negative sentiment against persons with disabilities (Hutchinson et al., 2020), or associating ethnic stereotypes between Hispanics and housekeepers and Asians with professors (Garg et al., 2018). Furthermore, a specific gap in neural hate speech classifiers consists of their over-sensitivity to group identifiers such as "Muslim", "gay", and "black", which are only hate speech according to offensive context (Dixon et al., 2018).

Recent NLP models are frequently recognized as "black boxes", meaning their decisions lack transparency and may be biased. Explainability methods can uncover potentially biased text features by applying a set of explanation techniques (Gongane et al., 2024). For instance, (Mathew et al., 2021) introduced the HateXplain dataset in English,

[1] The code and dataset are publicly available: https://github.com/isadorasalles/HateBRXplain

Proceedings of the 31st International Conference on Computational Linguistics, pages 6659–6669
January 19–24, 2025. ©2025 Association for Computational Linguistics

which includes human rationales to explain the hate speech labels assigned to the text. Their findings demonstrate that incorporating human rationales as an additional signal during training can reduce bias in hate speech detection models toward specific communities.

Specifically tailored to low-resource languages, the resources are still scarce. As a result, research and development of hate speech technologies for low-resource languages are less developed. For instance, five datasets were proposed for the Portuguese language. However, none of these datasets comprise human-annotated rationales, which is essential to building more interpretable and responsible hate speech detection models.

To fill this relevant gap, we introduce `HateBRXplain`, a benchmark dataset for explainable hate speech detection in Brazilian Portuguese. The `HateBRXplain` was manually annotated by two different annotators using human-based rationales. We evaluated our corpus using mBERT, BERTimbau, DistilBERTimbau, and PTT5 models, which reached high performance and outperformed the current baselines for Portuguese. Furthermore, to assess the explainability of these models, we employed two model-agnostic explainable methods (LIME and SHAP). We then evaluated the predicted rationales following standard metrics. Our contributions may be summarized as follows:

- We study an under-explored and relevant problem: explainable hate speech detection for low-resource languages.

- We provide the first hate speech dataset in Brazilian Portuguese with human-annotated rationales for explainable hate speech detection. The `HateBRXplain` consists of 7,000 Instagram comments derived from the HateBR dataset, in which 3,500 offensive comments were annotated with human-based rationales, i.e., text spans that support a particular class.

- We evaluate the `HateBRXplain` on mBERT, BERTimbau, DistilBERTimbau, and PTT5 models. The results overcame the current baseline for Portuguese, achieving an F1-score of 0.91 with BERTimbau.

- We assess the explainability of these classifiers by evaluating the quality of rationales predicted by two model-agnostic explanation methods (LIME and SHAP) following classical metrics, such as plausibility and faithfulness. Results for plausibility demonstrate an overall high agreement between human annotations and post-hoc explanations. However, while some models excel in classification performance, they cannot always provide faithful rationales for their decisions.

2 Related Work

2.1 Hate Speech Datasets in Portuguese

Most hate speech and offensive language corpora have been developed for the English language (Davidson et al., 2017; Zampieri et al., 2019; Fersini et al., 2018), leaving a resource gap for other languages. To address this gap for Portuguese, (de Pelle Pelle and Moreira, 2017) introduced the OFFCOMBR, a corpus containing 1,250 comments comprising two annotations: (i) “offensive” or “non-offensive”; and (ii) six hate speech categories (racism, sexism, homophobia, xenophobia, religious intolerance, and cursing). Further work by (Fortuna et al., 2019a) presents a dataset consisting of 5,668 tweets in both European and Brazilian Portuguese, comprising two annotation levels: (i) “hate” or “no-hate”; and (ii) multilabel hate speech hierarchical annotation schema with 81 hate categories. (Leite et al., 2020) introduced ToLD-Br (Toxic Language Dataset for Brazilian Portuguese), a dataset containing 21,000 tweets in Brazilian Portuguese manually annotated into one of seven categories: non-toxic, LGBTQ+phobia, obscene, insult, racism, misogyny, and xenophobia. More recently, (Trajano et al., 2023) introduced OLID-BR (Offensive Language Identification Dataset for Brazilian Portuguese) comprising 6,354 comments extracted from Twitter, YouTube, and other related datasets, annotated according to different categories: health, ideology, insult, LGBTQ+phobia, other lifestyle, physical aspects, profanity/obscene, racism, religious intolerance, sexism, and xenophobia. Finally, (Vargas et al., 2022, 2024a) proposed HateBR corpus, composed of 7,000 Instagram comments manually annotated by experts, comprising three levels: (i) “offensive” or “non-offensive”; (ii) offense levels (high, moderately, and slightly offensive); and (iii) nine hate group targets (xenophobia, racism, homophobia, sexism, religious intolerance, partyism, apology to dictatorship, antisemitism, and fatphobia). Table 1 summarizes the available data resources for the Portuguese language.

Authors	Data	# Instances	% Offensive	Platform	Models	F-score
(de Pelle Pelle and Moreira, 2017)	OFFCOMBR	1,250	32.50	Globo news comments	NB, SVM	0.80
(Fortuna et al., 2019a)	No-Name	5,668	31.50	Twitter	LSTM	0.78
(Leite et al., 2020)	ToLD-Br	21,000	44.07	Twitter	BERT	0.76
(Trajano et al., 2023)	OLID-BR	6,354	90.07	Twitter, YouTube	BERT	0.77
(Vargas et al., 2022, 2024a)	HateBR	7,000	50.00	Instagram	BERT	0.85

Table 1: Overview of Portuguese data resources for hate speech detection.

Authors	Data	Language	# Instances	Platform	Models	F-score
(Mathew et al., 2021)	HateXplain	English	20,148	Gab and Twitter	CNN-GRU, BiRNN, BERT	0.69
(Pavlopoulos et al., 2021)	No-Name	English	10,629	Civil Comments dataset	BERT	0.71
(Ravikiran and Annamalai, 2021)	DOSA	Tamil-English Kannada-English	4,786 1,097	YouTube	BERT	0.40
(Hoang et al., 2023)	ViHOS	Vietnamese	11,056	Facebook and YouTube	BiLSTM, BERT	0.78
(Delbari et al., 2024)	PHATE	Persian	7,000	Twitter	BERT, GPT	0.78

Table 2: Overview of data resources for hate speech with human rationales annotation.

2.2 Hate Speech Datasets with Human-Annotated Rationales

Hate speech detection has been extensively researched. However, there are limited studies specifically explaining the model's decision on what is classified as hate speech. (Zaidan et al., 2007) introduced the concept of exploiting human-annotated rationales during training of a learning method to boost performance. Using rationales may result in better models that reduce unintended bias toward target communities (Mathew et al., 2021). Moreover, it enhances transparency and accountability by supporting the development of interpretable models clarifying the reasoning behind each hate speech classification.

(Mathew et al., 2021) introduced the HateXplain dataset containing hate and offensive speech with span annotations that capture human rationales for 20,148 Gab and Twitter posts. (Pavlopoulos et al., 2021) provides a collection of 10,629 English posts derived from the Civil Comments dataset with human annotations for toxic spans. (Ravikiran and Annamalai, 2021) presents DOSA (Dravidian Offensive Span Identification Dataset), a dataset with annotated offensive spans for under-resourced language comments posted on YouTube. It comprises 4,786 Tamil-English comments and 1,097 Kannada-English comments. (Hoang et al., 2023) presents the ViHOS (Vietnamese Hate and Offensive Spans) dataset, which consists of 11,056 comments derived from Facebook and YouTube with span annotations that capture the human rationales for labeling a comment as hate or offensive. Finally, (Delbari et al., 2024) introduced PHATE, a dataset consisting of 7,000 Persian tweets tailored to multi-label hate speech detection. Each tweet was manually annotated with the targeted hate speech group and the rationale for the assigned label. Table 2 summarizes the available data resources for hate speech detection with rationales annotations.

No existing hate speech dataset in Portuguese provides human rationales. To address this gap, we propose `HateBRXplain`, the first benchmark dataset in Portuguese for explainable hate speech detection.

3 HateBRXplain Corpus

`HateBRXplain` consists of an explainable version of HateBR corpus (Vargas et al., 2022, 2024a). The HateBR corpus comprises 7,000 comments collected from the comment section of Brazilian politicians' accounts on Instagram and manually annotated by specialists, reaching a high inter-annotator agreement (75% of kappa). Precisely, it consists of comments annotated according to three different layers: a binary classification (offensive versus non-offensive comments), offensiveness-level classification (highly, moderately, and slightly offensive), and nine hate speech groups (xenophobia, racism, homophobia, sexism, religious intolerance, partyism, apology for the dictatorship, antisemitism, and fatphobia). In addition, the HateBR corpus presents a balanced class distribution (3,500 offensive and 3,500 non-offensive), in contrast to the other Portuguese corpora presented in Table 1, in which the classes are unbalanced. Moreover, the baseline experiments on HateBR overcame the other existent baselines for the Portuguese language.

This paper presents `HateBRXplain`, the first human-annotated rationale corpus designed for explainable hate speech detection in Brazilian Portuguese. To develop this explainable version of HateBR, we employed the concept of rationales—specific text spans within a comment that supports its categorization.

Do you really believe what you're saying? If you don't, you're just a liar. If you do you're mentally ill.
Criminal! Scoundrel and racist! May justice be served!
That's what you want — people employed but just following orders like puppets.

Table 3: Examples of annotations from `HateBRXplain`. The highlighted text indicates the rationale provided by the annotator.

3.1 Human-Annotated Rationales

A detailed description of our annotation process approach is presented in this section. For annotating rationales, we focused exclusively on spans of text that indicate offensiveness. Annotators were instructed to highlight only the portions of text that supported the offensive label, resulting in rationales being provided exclusively for the 3,500 offensive comments. According to our guidelines, a *rationale* is defined as a set of text spans, with each span being the smallest text segment that conveys offensive meaning. Thus, a text span can be either a word or a phrase. Consequently, each comment may contain multiple text spans that constitute the rationale. Table 3 presents annotation examples from our dataset.

3.1.1 Profile of Annotators

To ensure the reliability of data annotation, two independent annotators performed the task. To minimize bias and enhance the robustness of the results, we diversified the annotators' profiles, as detailed in Table 4.

Profile	Description
Education	Master's and PhD candidates
Gender	Female
Color	Black and White
Brazilian States	São Paulo and Minas Gerais

Table 4: Annotators' profile.

3.1.2 Annotation Evaluation

The final step of the annotation process involves assessing the quality of the annotated data by evaluating the annotators' agreement. We used two metrics for this evaluation: the *Jaccard Index* and the *F1-score*, calculated at the human rationale annotations' span level. We treated partial overlaps between spans as valid matches to compute both measures. Specifically, we calculated the Intersection-over-Union (IoU) for each pair of spans, and if the IoU was 0.5 or higher, the spans were considered a match. The *Jaccard Index* metric measures the similarity between two sets of spans by calculating the Intersection-over-Union (IOU), as defined in Equation 1.

Metric	Human-annotations	Random
Jaccard Index	**0.6746**	0.4855
F1-score	**0.7168**	0.5735

Table 5: Evaluation of human-annotation rationales compared with random annotation.

$$\text{Span Jaccard index} = \frac{\text{Number of Matching Spans}}{\text{Total Unique Spans (Union)}} \tag{1}$$

In Equation 2, we define the *span-level F1-score*.

$$\begin{aligned} \text{Span F1-score} &= 2 \times \frac{P_i \times R_i}{P_i + R_i} \\ \text{where } P_i = \frac{|A_i^1 \cap A_i^2|}{|A_i^1|} &\text{ and } R_i = \frac{|A_i^1 \cap A_i^2|}{|A_i^2|} \end{aligned} \tag{2}$$

where A_i^1 and A_i^2 represent the spans for the $i-$th instance provided by Annotator 1 and Annotator 2, respectively. We averaged the Jaccard Index and F1-score across pairs of human rationale annotations and compared these results with randomly generated rationale annotations. The comparison is presented in Table 5. To generate the random rationale annotations, we maintained the same distribution of span sizes as Annotator 1 and Annotator 2, selecting an equivalent number of tokens for each instance and producing two random rationale sets. As expected, the metrics for human annotations outperformed those for random annotations, indicating a higher level of agreement among human annotators.

3.2 Corpus Statistics

Table 6 presents the statistics of the `HateBRXplain` corpus, categorized by label (offensive and non-offensive). As mentioned, the corpus is balanced, with 3,500 offensive comments and 3,500 non-offensive comments. Additionally, we observe minimal differences in both the average length of the comments and the distribution of part-of-speech tags between the two classes.

Two annotators provided rationale explanations for each instance in the dataset. Table 7 presents statistics for these human annotations. On average, Annotator 1 highlighted 7.72 tokens per post, compared to 5.14 tokens by Annotator 2, and identified more spans overall. Moreover, Figure 1 illustrates that the spans annotated by Annotator 1 are more

Description		Label	
		Offensive	Non-offensive
# Comments		3,500	3,500
# Sentences		4,871	4,674
# Words		53,455	42,891
Avg Sentences/Comment		1.3917	1.3354
Avg Words/Comment		15.2728	12.2546
Part-of-Speech (Avg)	Noun	3.4920	2.7994
	Verb	2.4043	1.8068
	Adjective	0.9880	0.8774
	Adverb	1.1263	0.8888
	Pronoun	1.1197	0.8900

Table 6: HateBRXplain data statistics per label.

Description		Rationales	
		Annotator 1	Annotator 2
# Spans		6,601	5,922
# Tokens		27,038	17,992
Avg spans/Comment		1.8860	1.6920
Avg tokens/Comment		7.7251	5.1406
Part-of-Speech (Avg)	Noun	2.0328	1.5097
	Verb	1.3183	0.7988
	Adjective	0.6194	0.4754
	Adverb	0.4991	0.2874
	Pronoun	0.4477	0.2614

Table 7: Human rationales statistics per annotator.

extended in terms of the number of characters, indicating that Annotator 1 needed more contextual information to identify offensiveness compared to Annotator 2. This observation is further supported by the analysis of part-of-speech tags, which reveals a higher presence of nouns and verbs in the rationales provided by Annotator 1.

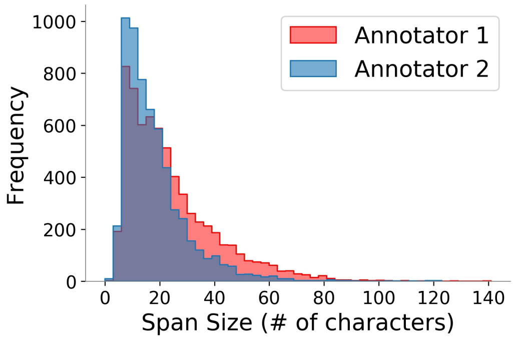

Figure 1: Histogram of annotated span size regarding the number of characters for both annotators.

4 Experimental Setup

In this section, we provide (i) details on the models used to evaluate the performance of hate speech detection using the proposed dataset and (ii) details of the post-hoc explanation methods used to provide automatic explanations over the sentences of our dataset to assess the explainability aspects of the proposed learning models.

4.1 Hate Speech Detection Models

To evaluate the generalization of the dataset, we used four pre-trained models as follows:

mBERT.[2] Developed by (Devlin et al., 2019), mBERT is a multilingual variation of the BERT model. Pre-trained in the top 104 languages with the largest Wikipedia, including Portuguese. The base model has 12 layers and 110M parameters.

BERTimbau.[3] Developed by (Souza et al., 2020), BERTimbau is a Brazilian Portuguese language model based on the BERT architecture (Devlin et al., 2019). Trained on over 3.5 million documents from BrWaC corpus (Brazilian Web as Corpus) (Wagner Filho et al., 2018). The base model has 12 layers and 110M parameters.

DistilBERTimbau.[4] A small, fast, and cheap transformer model trained by distilling BERTimbau base architecture (Silva Barbon and Akabane, 2022), having 66.4M parameters.

PTT5.[5] Developed by (do Carmo et al., 2020), PTT5 is also a Brazilian Portuguese language model pre-trained on the BrWaC corpus (Wagner Filho et al., 2018). PTT5 is based on T5 architecture (Raffel et al., 2020), an encoder-decoder transformer model. The base model has 220M parameters, and we used the T5ForSequenceClassification layer to predict each instance.

4.1.1 Hyperparameters

We fine-tuned these models on the HateBRXplain dataset, ensuring the Instagram comments were uniformly normalized using a Portuguese-specific tool as described by (Costa Bertaglia and Volpe Nunes, 2016). To compare the results, all these models were evaluated using the same train:validation:test split of 8:1:1, performing stratified split to maintain class balance. The results were obtained from the test set, while the validation set was used for hyperparameter tuning. For all models, we set the maximum sentence length to 512 and employed the AdamW optimizer. The learning rates were set as follows: 2e-5 for mBERT, 1e-5 for BERTim-

[2] google-bert/bert-base-multilingual-cased

[3] neuralmind/bert-base-portuguese-cased

[4] adalbertojunior/distilbert-portuguese-cased

[5] unicamp-dl/ptt5-base-portuguese-vocab

Rationales method	Text
Human Annotator	Who cares about what you think, corrupt communist? You should be in jail!
mBERT [LIME]	Who cares about what you think, corrupt communist? You should be in jail!
mBERT [SHAP]	Who cares about what you think, corrupt communist? You should be in jail!
BERTimbau [LIME]	Who cares about what you think, corrupt communist? You should be in jail!
BERTimbau [SHAP]	Who cares about what you think, corrupt communist? You should be in jail!
DistilBERTimbau [LIME]	Who cares about what you think, corrupt communist? You should be in jail!
DistilBERTimbau [SHAP]	Who cares about what you think, corrupt communist? You should be in jail!
PTT5 [LIME]	Who cares about what you think, corrupt communist? You should be in jail!
PTT5 [SHAP]	Who cares about what you think, corrupt communist? You should be in jail!

Table 8: Example of model-predicted rationales compared to human annotations for a single instance. Green highlights indicate tokens both the model and human annotators found important for the prediction. In contrast, orange highlights indicate tokens considered important by the model but not by the human annotators.

bau, 2e-5 for DistilBERTimbau, and 3e-4 for PTT5, with a batch size of 8 for all models. The training was conducted for five epochs, and the reported test set results are based on the epoch that achieved the lowest loss on the validation set.

4.2 Post-hoc Explanation Methods

To evaluate the explainability of the proposed classification models, we employed two widely-used post-hoc explanation methods: LIME (Local Interpretable Model-agnostic Explanations) (Ribeiro et al., 2016) and SHAP (SHapley Additive exPlanations) (Lundberg, 2017). Both are model-agnostic methods, meaning they generate explanations without requiring access to the internal workings of the model, relying solely on input-output pairs. LIME provides local explanations by perturbing the input data and observing how these changes affect the model's predictions. In contrast, SHAP evaluates the contribution of each feature to the prediction by considering all possible feature combinations, offering both local and global interpretability.

For comparison purposes, we applied both methods to generate local explanations (rationales) for a randomly selected subset of 350 offensive comments (10% of the rationales annotated data) from `HateBRXplain`. Table 8 shows an example of model-predicted explanations.

5 Evaluation and Results

This section reports the evaluation of the four models for detecting hate speech. In addition, we present the evaluation of the explainability aspects of these learning models.

5.1 Evaluation of Models

We assessed the performance of the four classifiers in distinguishing between offensive and non-offensive speech, using accuracy and macro F1-score as evaluation metrics. The results are shown in Table 9. Additionally, we conducted a ROC curve analysis, as illustrated in Figure 2.

Model	Metric	
	Accuracy	macro F1
mBERT	0.8743	0.8737
BERTimbau	**0.9157**	**0.9153**
DistilBERTimbau	0.9000	0.8994
PTT5	0.9043	0.9035

Table 9: Evaluation of mBERT, BERTimbau, DistilBERTimbau, and PTT5 models on the test set.

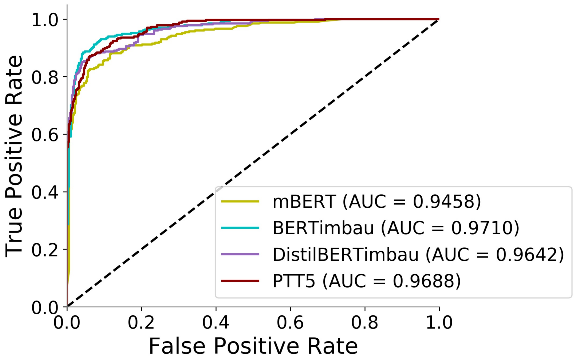

Figure 2: ROC curves for the four fine-tuned models.

We observed that models pre-trained exclusively on a Portuguese corpus outperformed the multilingual BERT. The highest performance was achieved with the BERTimbau model, which reached an F1-Score of 0.91 and an AUC of 0.97.

5.2 Evaluation of Explanations

We then employed classical metrics to evaluate the model-predicted rationales generated by the SHAP and LIME methods. This section describes these metrics and presents the results.

5.2.1 Metrics

Building on prior research (DeYoung et al., 2020; Jacovi and Goldberg, 2020; Mathew et al., 2021; Wang et al., 2022), we assessed the explainability of the proposed models with plausibility and faithfulness metrics. *Plausibility* evaluates how closely the model's explanations align with human judgment. In contrast, *faithfulness* measures how accurately these model-predicted rationales reflect the model's actual decision-making process (Jacovi and Goldberg, 2020).

Plausibility: To assess plausibility, we reported the IOU (Intersection-Over-Union) F1-score, along with token-level Precision, Recall, and F1-score metrics (DeYoung et al., 2020). These metrics are computed on the token level since the rationales given by the post-hoc explanations are individual tokens rather than text spans. Moreover, since each instance has two human-annotated rationales, we consider the human rationale with the highest agreement with the model-predicted rationale as the ground truth.

The IOU F1-score is defined in Equation 3, in which, for each instance, the IOU_i is given by the overlap between rationales tokens divided by the union of tokens. A model-predicted rationale matches a human-annotated rationale if its IOU_i equals or exceeds 0.5. We accounted for these partial matches to calculate the IOU F1-score.

$$\begin{gathered}\text{IOU-F1} = \frac{1}{N}\sum_{i=1}^{N}\text{Greater}(\text{IOU}_i, 0.5)\\ \text{where } \text{IOU}_i = \frac{|M_i \cap H_i|}{|M_i \cup H_i|}\end{gathered} \tag{3}$$

where M_i and H_i represent the rationales of the i-th instance provided by the model and human, respectively; N is the number of instances. Following these definitions, we also measured token-level Precision and Recall and used these to derive token-level F1-score as defined in Equation 4.

$$\begin{gathered}\text{Token-F1} = \frac{1}{N}\sum_{i=1}^{N}(2 \times \frac{P_i \times R_i}{P_i + R_i})\\ \text{where } P_i = \frac{|M_i \cap H_i|}{|M_i|} \text{ and } R_i = \frac{|M_i \cap H_i|}{|H_i|}\end{gathered} \tag{4}$$

Faithfulness: To measure faithfulness, we reported the comprehensiveness and sufficiency (DeYoung et al., 2020). *Comprehensiveness*, defined in Equation 5, measures whether the tokens necessary to make a prediction were selected. For each instance, x_i, let $m(x_i)_j$ be the prediction a model m provides for class j and r_i be the predicted rationales. We then define $m(x_i \backslash r_i)_j$ as the model m predicted probability of x_i without the predicted rationales r_i. A high comprehensiveness value implies that the rationales are influential in the prediction. Otherwise, *sufficiency* measures the degree to which the predicted rationales are adequate for making a prediction. In Equation 6, let $m(r_i)_j$ be the prediction probability of giving only the model-predicted rationales r_i to a model m for class j. A low sufficiency value implies that the rationales are sufficient to make a prediction.

$$\text{Comp.} = \frac{1}{N}\sum_{i=1}^{N}(m(x_i)_j - m(x_i \backslash r_i)_j) \tag{5}$$

$$\text{Suff.} = \frac{1}{N}\sum_{i=1}^{N}(m(x_i)_j - m(r_i)_j) \tag{6}$$

5.2.2 Results

Table 10 presents the evaluation of the post-hoc explanations given by LIME and SHAP regarding plausibility and faithfulness. Notably, while the BERTimbau model achieved the best performance metrics in the classification task (shown in Section 5.1), it did not rank among the top models for explainability metrics. Regarding plausibility, our evaluation revealed that **PTT5 [SHAP]** model achieved the top scores for IOU F1, Token Recall, and Token F1. Apart from the DistilBERTimbau models, the post-hoc method SHAP performed better on plausibility metrics than LIME. However, the number of tokens returned by each method varies significantly. SHAP typically returned more tokens than LIME, as shown in Table 11. Since the plausibility metrics are evaluated by comparing model-predicted rationales to human-annotated ones—which are often more complex and contextually rich (e.g., complete phrases)—the overlap between LIME's tokens and human annotations is generally smaller than that of SHAP, leading to higher metric scores for SHAP.

Additionally, among the methods, the scores achieved for the Token F1 metric are often higher than the ones achieved by IOU-F1. However, IOU-F1 is less precise, as it counts a prediction as a match only if the predicted rationale overlapping with the human annotation is at least 0.5. Moreover, Token Precision is frequently higher than Token Recall. The explanation for this relies on the way we compute these metrics. As outlined in Equation 4, the precision metric relies on the model-predicted rationales, whereas the recall metric relies on the human-annotated. According to Table 11, the average number of tokens in model-predicted rationales is generally lower than the average tokens provided by Annotator 1, which is 7.72, as shown in Table 7. The mBERT [SHAP] and the PTT5 [SHAP] are the only models where Token Recall exceeded Token

	Plausibility				Faithfulness	
Moodel [XAI method]	IOU F1 ↑	Token Precision ↑	Token Recall ↑	Token F1 ↑	Comp. ↑	Suff. ↓
mBERT [LIME]	0.5828	0.7458	0.6936	0.6701	0.8809	0.0134
mBERT [SHAP]	0.6628	0.7143	0.7520	0.6897	0.9324	0.0172
BERTimbau [LIME]	0.5857	0.7557	0.6848	0.6698	0.5904	0.0237
BERTimbau [SHAP]	0.6600	0.7489	0.7099	0.6831	0.6458	0.0215
DistilBERTimbau [LIME]	0.6457	**0.7614**	0.7276	0.7003	0.9407	0.0115
DistilBERTimbau [SHAP]	0.6200	0.7543	0.6862	0.6720	**0.9475**	0.0114
PTT5 [LIME]	0.6057	0.7487	0.6978	0.6776	0.5654	**0.0016**
PTT5 [SHAP]	**0.7400**	0.7177	**0.8378**	**0.7362**	0.6160	0.0083

Table 10: Evaluation of explanations predicted by LIME and SHAP post-hoc explainability methods.

Model [XAI method]	Avg Tokens
mBERT [LIME]	5.3114
mBERT [SHAP]	8.9571
BERTimbau [LIME]	5.3171
BERTimbau [SHAP]	7.6400
DistilBERTimbau [LIME]	5.5686
DistilBERTimbau [SHAP]	7.2800
PTT5 [LIME]	5.4086
PTT5 [SHAP]	10.0428

Table 11: Average number of tokens predicted by the models.

Precision; these models have a higher average of predicted tokens.

Regarding the faithfulness evaluation, we reported the comprehensiveness and sufficiency scores. In terms of comprehensiveness, the models with the highest scores were **mBERT** and **DistilBERTimbau**, which were the ones that presented the lowest F1-scores on the classification task (as shown in Table 9). Although the PTT5 [SHAP] model achieved the highest plausibility metrics, it did not perform as well on the comprehensiveness metric. In contrast, the sufficiency metric presented a low score across all models, with the best value for **PTT5 [LIME]**, indicating that the models-predicted rationales are generally adequate for predicting whether a comment is offensive. However, some methods missed the prediction of essential rationales, which impacted their comprehensiveness, as this metric measures whether the model has included all necessary tokens for making a prediction, ensuring that no essential tokens are omitted.

Finally, Figure 3 displays the Jaccard Index (Equation 1) as a measure of similarity for the top 50 most important tokens identified by each model. The figure reveals that tokens predicted by the same classifier using different explainability methods tend to have a higher similarity. Notably, the PTT5 [SHAP] model, which achieved the highest plausibility metrics and predicted the largest average number of tokens, showed the lowest similarity with other models.

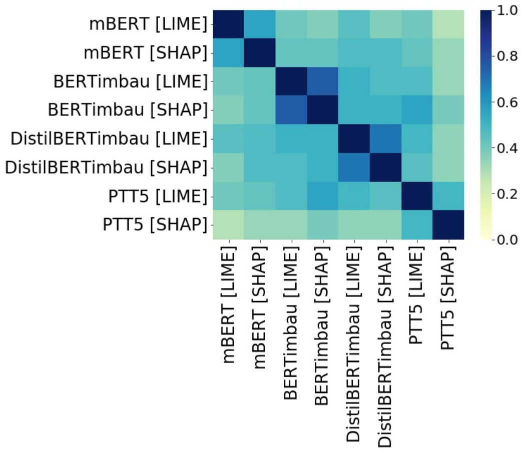

Figure 3: Similarity analysis of the top 50 most important tokens identified by each model.

6 Conclusion

This paper introduces HateBRXplain, a novel benchmark dataset designed for explainable hate speech detection in Brazilian Portuguese. The dataset contains 7,000 Instagram comments, evenly split between 3,500 offensive and 3,500 non-offensive entries. Two annotators further annotate each offensive comment with rationales—text spans that justify the offensive label. We evaluated several state-of-the-art hate speech detection models on this dataset, including variations of BERT and T5. Furthermore, we assessed the models' explainability through model-agnostic explanation methods, evaluating plausibility and faithfulness metrics. Our findings reveal that while some models excel in classification performance, they do not necessarily achieve the highest explainability metrics, highlighting the ongoing need for transparency and reliability in developing hate speech detection models. We believe this work will significantly impact the field, paving the way for further research in explainable hate speech detection in Portuguese, with HateBRXplain as a valuable resource for advancing model performance and interpretability.

Limitations

HateBRXplain is an extension of the HateBR dataset, meaning it inherits any limitation present in HateBR construction. For instance, HateBR was collected from the comment section of Brazilian politicians' accounts on Instagram, so the dataset may not represent hate speech on other platforms and domains. However, among all available datasets for Brazilian Portuguese, HateBR stands out as the only one annotated by experts, balanced in terms of offensive and non-offensive comments, and presented baseline evaluations, achieving the highest F1-score in the hate speech detection task among all Portuguese datasets available.

Additionally, the rationales annotation process relied on human annotators, which inherently introduces potential subjective biases and inconsistencies. We attempted to mitigate this by diversifying the annotator's profile and providing clear annotation guidelines, but some subjectivity is unavoidable.

Ethical Considerations

This paper's data resources and artifacts are open-source and anonymized. Furthermore, an expert in responsible AI supervised the entire annotation process to ensure that the labels did not raise ethical concerns. Finally, the annotators selected for this task represent diverse cultural and demographic backgrounds, including members of affected groups, to ensure both fairness and representativeness.

Acknowledgements

We kindly thank Google, CAPES, and CNPq for supporting this project.

References

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2021. HateBERT: Retraining BERT for abusive language detection in English. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 17–25, Online. Association for Computational Linguistics.

Thales Felipe Costa Bertaglia and Maria das Graças Volpe Nunes. 2016. Exploring word embeddings for unsupervised textual user-generated content normalization. In *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*, pages 112–120, Osaka, Japan. The COLING 2016 Organizing Committee.

Aida Mostafazadeh Davani, Mohammad Atari, Brendan Kennedy, and Morteza Dehghani. 2023. Hate speech classifiers learn normative social stereotypes. *Transactions of the Association for Computational Linguistics*, 11:300–319.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 512–515.

Rogers Prates de Pelle Pelle and Viviane P Moreira Moreira. 2017. Offensive comments in the brazilian web: a dataset and baseline results. In *Congresso da Sociedade Brasileira de Computação-CSBC*.

Zahra Delbari, Nafise Sadat Moosavi, and Mohammad Taher Pilehvar. 2024. Spanning the spectrum of hatred detection: a persian multi-label hate speech dataset with annotator rationales. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17889–17897.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458, Online. Association for Computational Linguistics.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, page 67–73, New York, NY, USA. Association for Computing Machinery.

Diedre Santos do Carmo, Marcos Piau, Israel Campiotti, Rodrigo Nogueira, and Roberto de Alencar Lotufo. 2020. Ptt5: Pretraining and validating the t5 model on brazilian portuguese data. *ArXiv*, abs/2008.09144.

Elisabetta Fersini, Paolo Rosso, Maria Anzovino, et al. 2018. Overview of the task on automatic misogyny identification at ibereval 2018. *Ibereval@ sepln*, 2150:214–228.

Paula Fortuna, Joao Rocha da Silva, Leo Wanner, Sérgio Nunes, et al. 2019a. A hierarchically-labeled portuguese hate speech dataset. In *Proceedings of the third workshop on abusive language online*, pages 94–104.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4):1–30.

Paula Fortuna, João Rocha da Silva, Juan Soler-Company, Leo Wanner, and Sérgio Nunes. 2019b. A hierarchically-labeled Portuguese hate speech dataset. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 94–104, Florence, Italy. Association for Computational Linguistics.

Björn Gambäck and Utpal Kumar Sikdar. 2017. Using convolutional neural networks to classify hate-speech. In *Proceedings of the First Workshop on Abusive Language Online*, pages 85–90, Vancouver, BC, Canada. Association for Computational Linguistics.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.

Vaishali U Gongane, Mousami V Munot, and Alwin D Anuse. 2024. A survey of explainable ai techniques for detection of fake news and hate speech on social media platforms. *Journal of Computational Social Science*, pages 1–37.

Ella Guest, Bertie Vidgen, Alexandros Mittos, Nishanth Sastry, Gareth Tyson, and Helen Margetts. 2021. An expert annotated dataset for the detection of online misogyny. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1336–1350, Online. Association for Computational Linguistics.

Phu Gia Hoang, Canh Duc Luu, Khanh Quoc Tran, Kiet Van Nguyen, and Ngan Luu-Thuy Nguyen. 2023. ViHOS: Hate speech spans detection for Vietnamese. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 652–669, Dubrovnik, Croatia. Association for Computational Linguistics.

Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. 2020. Social biases in NLP models as barriers for persons with disabilities. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5491–5501, Online. Association for Computational Linguistics.

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205, Online. Association for Computational Linguistics.

João Augusto Leite, Diego Silva, Kalina Bontcheva, and Carolina Scarton. 2020. Toxic language detection in social media for Brazilian Portuguese: New dataset and multilingual analysis. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 914–924, Suzhou, China. Association for Computational Linguistics.

Scott Lundberg. 2017. A unified approach to interpreting model predictions. *arXiv preprint arXiv:1705.07874*.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 14867–14875.

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. 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)*, pages 622–628, Minneapolis, Minnesota. Association for Computational Linguistics.

John Pavlopoulos, Jeffrey Sorensen, Léo Laugier, and Ion Androutsopoulos. 2021. Semeval-2021 task 5: Toxic spans detection. In *Proceedings of the 15th international workshop on semantic evaluation (SemEval-2021)*, pages 59–69.

Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. 2021. Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, 55(3):477–523.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Manikandan Ravikiran and Subbiah Annamalai. 2021. Dosa: Dravidian code-mixed offensive span identification dataset. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 10–17.

Marco Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. “Why should I trust you?”: Explaining the predictions of any classifier. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 97–101, San Diego, California. Association for Computational Linguistics.

Rafael Silva Barbon and Ademar Takeo Akabane. 2022. Towards transfer learning techniques—bert, distilbert, bertimbau, and distilbertimbau for automatic text classification from different languages: A case study. *Sensors*, 22(21):8184.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. Bertimbau: pretrained bert models for brazilian portuguese. In *Intelligent Systems: 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, October 20–23, 2020, Proceedings, Part I 9*, pages 403–417. Springer.

Douglas Trajano, Rafael H Bordini, and Renata Vieira. 2023. Olid-br: offensive language identification dataset for brazilian portuguese. *Language Resources and Evaluation*, pages 1–27.

Yulia Tsvetkov, Vinodkumar Prabhakaran, and Rob Voigt. 2019. Socially responsible natural language processing. In *Companion Proceedings of The 2019 World Wide Web Conference*, WWW '19, page 1326, New York, NY, USA. Association for Computing Machinery.

Francielle Vargas, Isabelle Carvalho, Thiago A. S. Pardo, and Fabrício Benevenuto. 2024a. Context-aware and expert data resources for brazilian portuguese hate speech detection. *Natural Language Processing*, page 1–22.

Francielle Vargas, Isabelle Carvalho, Fabiana Rodrigues de Góes, Thiago Pardo, and Fabrício Benevenuto. 2022. HateBR: A large expert annotated corpus of Brazilian Instagram comments for offensive language and hate speech detection. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7174–7183, Marseille, France. European Language Resources Association.

Francielle Vargas, Samuel Guimarães, Shamsuddeen Hassan Muhammad, Diego Alves, Ibrahim Said Ahmad, Idris Abdulmumin, Diallo Mohamed, Thiago Pardo, and Fabrício Benevenuto. 2024b. HausaHate: An expert annotated corpus for Hausa hate speech detection. In *Proceedings of the 8th Workshop on Online Abuse and Harms (WOAH 2024)*, pages 52–58, Mexico City, Mexico. Association for Computational Linguistics.

Jorge A Wagner Filho, Rodrigo Wilkens, Marco Idiart, and Aline Villavicencio. 2018. The brwac corpus: A new open resource for brazilian portuguese. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*.

Lijie Wang, Yaozong Shen, Shuyuan Peng, Shuai Zhang, Xinyan Xiao, Hao Liu, Hongxuan Tang, Ying Chen, Hua Wu, and Haifeng Wang. 2022. A fine-grained interpretability evaluation benchmark for neural NLP. In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, pages 70–84, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Claire Wardle. 2024. *A Conceptual Analysis of the Overlaps and Differences between Hate Speech, Misinformation and Disinformation*. Department of Peace Operations (DPO). Office of the Special Adviser on the Prevention of Genocide (OSAPG). United Nations.

Omar Zaidan, Jason Eisner, and Christine Piatko. 2007. Using "annotator rationales" to improve machine learning for text categorization. In *Human language technologies 2007: The conference of the North American chapter of the association for computational linguistics; proceedings of the main conference*, pages 260–267.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the type and target of offensive posts in social media. 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)*, pages 1415–1420, Minneapolis, Minnesota. Association for Computational Linguistics.

Steven Zimmerman, Udo Kruschwitz, and Chris Fox. 2018. Improving hate speech detection with deep learning ensembles. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).