Unsupervised Evaluation of Explanations for Hate Speech Classification in Portuguese

Isabel Carvalho[1], Hugo Gonçalo Oliveira[1], Catarina Silva[1]

[1] University of Coimbra, CISUC/LASI – Centre for Informatics and Systems of the University of Coimbra, Department of Informatics Engineering

Correspondence: isabelc@dei.uc.pt, hroliv@dei.uc.pt, catarina@dei.uc.pt

Abstract

Top-performing Artificial Intelligence models often operate as black boxes. Explainable AI (XAI) can increase transparency, but its evaluation is currently hindered by a lack of annotated explanation data and agreed-upon validation standards. We propose a framework for evaluating the faithfulness of explanations in Portuguese hate speech detection. Our approach is based on the premise that a faithful explanation should identify features whose removal degrades a model's performance. We follow a three-step process: (i) prediction on the original input; (ii) identification and removal of explanatory keywords; and (iii), prediction on the modified input, with performance differences used as an evaluation signal. We conduct experiments using ensemble classifiers, multiple keyword selection strategies, and SHAP and LIME as XAI methods. In addition, Large Language Models (LLMs) are explored both as classifiers and as explainers. Results demonstrate that removing explanatory keywords degrades model performance more than random word removal, indicating explanation faithfulness. Notably, SHAP and LIME consistently provided more faithful explanations than LLM-generated or manual alternatives, although impact depends on the keyword selection strategy. These findings highlight the importance of standardised, unsupervised evaluation protocols for XAI and the faithfulness limitations of current generative LLM explanations.

1 Introduction

The widespread adoption of Machine Learning (ML), particularly deep neural networks and Large Language Models (LLMs), has brought remarkable advances across domains. However, the opacity of these models raises concerns across different applications, from health diagnosis, where transparency often outweighs performance (Watson et al., 2019); to facial recognition[1] or recidivism prediction[2], where bias has negatively affected people's lives; to hate speech detection, where AI can be used for prevention but also for proliferation[3]. To answer these concerns, Explainable AI (XAI) (Gunning, 2016) emerged as a new field, which aims to make AI decisions more transparent and trustworthy by providing insights into models' inner workings and predictions. However, there are no standardised way of explaining the previous widely-adopted models. The same happens for explanations, which are challenging to quantitatively evaluate. The ideal choice is to involve humans in the process, either by asking them to qualitatively evaluate explanations or to create annotated rationales. Yet, both of the previous options are time and resource-consuming (Arya et al., 2019; Herrewijnen et al., 2024). The lack of annotated data particularly affects the development and application of XAI to lower-resource languages such as Portuguese. In this work, we choose the case of hate speech detection for two reasons:

1. The increasing relevance of this task in society, as noted by Gongane et al. (2024) and Dietrich (2025). Both mention the reliance on AI for hate speech detection, with the former calling for explainability of ML models due to the multitude of dimensions involved in hate speech, and the latter reinforcing the risks of automated deletion, and presenting this as a high-risk application, which, based on the European Union's AI Act[4], requires interpretability, further motivating the application of XAI to this scenario;

[1] https://archive.nytimes.com/bits.blogs.nytimes.com/2015/07/01/google-photos-mistakenly-labels-black-people-gorillas/

[2] https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[3] https://stories.undp.org/hate-speech-in-the-age-of-ai

[4] https://artificialintelligenceact.eu/

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 1, pages 780–789
April 13-16, 2026

2. The existence of a dataset with human-annotated rationales in Portuguese, HateBRXplain (Salles et al., 2025), which allows for a better analysis and comparison of approaches.

Our goal is to develop unsupervised evaluation methods, which are not limited by the lack of human rationales or language-specific resources. Our proposal consists of three steps: (i) prediction of the original input; (ii) identification and removal of relevant keywords; and (iii), prediction of the edited input. We develop different methods to identify the keywords and, besides randomised baselines, rely on SHAP (Lundberg and Lee, 2017) and LIME (Ribeiro et al., 2016) as explainability methods due to their high popularity and agnostic nature. In line with current research trends and to analyse the faithfulness of explanations obtained via generative AI, two LLMs, LLaMa3 (Grattafiori et al., 2024) and Gemma3 (Team et al., 2025), are also integrated, both as classifiers and as explainers. We evaluate the impact of keyword-removal in the classifiers' performance as a proxy for explanation quality, present quantitative evaluation metrics, and compare the results with human rationales. We also analyse the keywords identified regarding their role as content or function words.

As such, the main contributions of this work are:

1. An unsupervised approach for the evaluation of explanations based on keyword-removal, with different keyword selection solutions;

2. The application to a Portuguese hate speech classification dataset;

3. A multi-layered analysis, including the impact of keyword-removal on performance, annotator evaluation, difference between rationales, keyword roles, and the comparison between different approaches.

The remainder of the paper is structured as follows: Section 2 discusses other works involving XAI and the evaluation of explanations; Section 3 introduces the general approach proposed, whereas Section 4 includes details on the dataset, models used, approaches for keyword selection, and implementation details; Section 5 presents the results and insights gathered from the application of the approach to hate speech detection; Section 6 concludes the paper and provides cues for future work. The paper ends with an analysis of its limitations.

2 Related Work

XAI is increasingly relevant in a society brimming with AI due to its role in auditing, debugging, improving models, and for discovering new insights (Adadi and Berrada, 2018). However, despite its growing relevance, there is no standardised form of explaining or evaluating explanations (Wiratsin and Ragkhitwetsagul, 2025).

Several works have discussed this issue and proposed explanation evaluation criteria (Molnar, 2020; Ghanvatkar and Rajan, 2024), e.g., accuracy, fidelity, faithfulness, and user usefulness, but many are difficult to measure, especially automatically. Faithfulness and plausibility have also been explored in LLMs, with Agarwal et al. (2024) expressing concerns about their faithfulness and lack of reliability. Ideally, explanations should involve humans (Herrewijnen et al., 2024; Adadi and Berrada, 2018; Wiratsin and Ragkhitwetsagul, 2025), but this is resource-intensive, and even supervised methods to select rationales can lead to weak or misaligned explanations (Hu and Yu, 2024).

There have been some works exploring XAI in Portuguese for different tasks, e.g., humor recognition (Inácio et al., 2023), proficiency and readability modeling (Ribeiro-Flucht et al., 2024), and punctuation restoration (de Lima et al., 2024). None of these works evaluated the explanations obtained. Salles et al. (2025), the creators of HateBRXplain, the hate speech detection corpus used in this work, apply model-agnostic explanation methods like SHAP (Lundberg and Lee, 2017) and LIME (Ribeiro et al., 2016) to identify and analyse rationales, which are common XAI methods for the task of hate speech detection (Ngueajio et al., 2025). The authors evaluate the plausibility and faithfulness of these approaches in encoder and encoder-decoder models, e.g., BERTimbau (Souza et al., 2020) and PTT5 (Carmo et al., 2020), concluding that the best-performing models fail to provide faithful rationales. Decoder-based models have also been explored for the same task, namely LLaMa 2[5], which achieved similar results to the proprietary GPT 3.5. Concise and direct prompts have been shown to be more effective (Kumarage et al., 2024), possibly due to performance drops when relevant information is positioned in the middle of long prompts (Liu et al., 2024). The authors of the corpus applied XAI to a random subset of

[5] https://huggingface.co/meta-llama/Llama-2-7b-chat-hf

10% of the annotated data and their analysis revolves around plausibility and faithfulness.

In another evaluation approach (Madsen et al., 2024), the main idea is that if a set of words is important for making a prediction, then it should not be able to make its prediction without these words. This is called Deletion Check (Nauta et al., 2023) and has been used for the evaluation of explanations. To determine the features for removal, some approaches include training models to assign relevance scores or relying on XAI approaches (Selvaraju et al., 2017) for feature identification and then selecting the top-k words. LLMs have also been used to identify relevant keywords, but with the goal of evaluating faithfulness and human alignment (Fayyaz et al., 2024).

While some forms of evaluating explanations have been explored, there is no standardised manner of doing this, and there is room for further research, namely on the identification of keywords and using the complete HateBRXplain dataset, making use of all its human-annotated rationales.

3 Proposed Approach

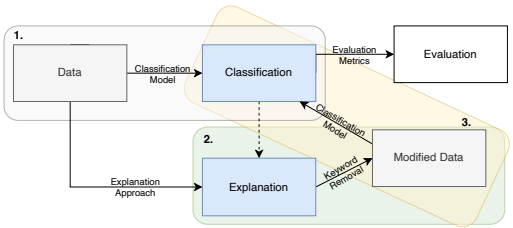

Figure 1: General approach for the evaluation of explanations, signalling each step of the proposal.

We propose an unsupervised framework for validating explanation faithfulness based on the assumption that a faithful explanation identifies features that are important for a model's prediction. Under this assumption, removing such features should lead to a measurable degradation in classification performance, whereas the removal of irrelevant features should have a smaller effect. Figure 1 illustrates the proposed framework, which consists of three main steps, detailed below with examples:

1. **Original prediction:** the classifier predicts a label for the original input, and its baseline performance is recorded.
    - **Input:** They are repulsive human beings.
    - **Predicted label:** Offensive/Hate speech.
2. **Explanation and keyword-removal:** an explanation method is applied to identify the most relevant keywords, which are then removed from the original input according to a predefined selection strategy.
    - **Explanatory keyword:** repulsive.
    - **Edited Input:** They are human beings.
3. **Modified-input prediction:** the classifier predicts a label for the modified input, and its performance is re-evaluated.
    - **Edited Input:** They are human beings.
    - **Predicted label:** Not offensive/neutral.

Steps 1 and 3 rely on ML models to classify each example. Step 2 involves the generation of explanations, which can be produced either by formal XAI models or by prompting the same LLMs used for classification. Keyword-removal can follow different strategies, e.g., selecting a number of top-ranked keywords or using score-based thresholds.

After this process, it is possible to evaluate the quality of the explanation based on its impact on performance, but also to analyse keyword roles (functional versus content), or compare rationales.

This approach was inspired by related work (Salles et al., 2025; Fayyaz et al., 2024) and differs in the explainer and classification models used, the keyword-selection techniques, and the metrics used, both regarding impact on performance and linguistic analysis.

In the next section, we detail the components and the decisions involved, including the choice of classifiers and LLM prompts, the techniques for keyword selection, and other technical details, including text tokenisation and representation.

4 Experimentation

We apply the general proposal presented earlier to the HateBRXplain dataset (Salles et al., 2025) and describe each step of the process in detail, namely, regarding data, models, statistical analyses, keyword selection, and implementation details.

4.1 Dataset Description

Our study is on a hate speech detection scenario, using HateBRXplain (Salles et al., 2025), a dataset derived from HateBR (Vargas et al., 2022), containing 7,000 Instagram comments in Portuguese, 3,500 labelled as offensive (1) and 3,500 as not

offensive (0), and human-annotated rationales, promoting explainable hate speech detection and enabling further explainability-related research. The rationales are composed of text spans that justify the attribution of the offensive label (1) to a given example, each annotated by two different people. There are no rationales for the non-offensive class. Included comments have, on average, 12 to 15 words, depending on the label, and one sentence per comment, as described by the authors. Regarding the rationales, the annotators identify, on average, five to seven tokens and one to two spans per comment. Table 1 presents examples of comments and annotated rationales. We select the rationales of annotator two, which tend to be shorter, as per the analysis performed by the authors, making them more suitable as examples.

Comment	Rationale
Quem tem pena é galinha, mas ela é uma VACA LOUCA.	ela é uma VACA LOUCA
Sem noção sobre o Brasil e uma corrupta	corrupta
Cai fora o Brasil não precisa de você	cai fora

Table 1: Examples from the HateBRXplain dataset, presenting hate speech comments and the rationales annotated by annotator two.

4.2 Classification Models

We chose ensemble methods and two LLMs for hate speech detection, motivated by their popularity, performance, and complexity, as they are seen as high-performing options with low explainability (Gongane et al., 2024). This makes them ideal for our work as we aim to analyse the drop in performance when relevant features are removed.

The ensembles chosen are Random Forest (RF) and Extreme Gradient Boosting (XGBoost). Regarding LLMs, we use two decoder-based multilingual models: LLaMa3-8B with Q4_0 quantisation and Gemma3-12B with Q4_K_M quantisation. RF and XGBoost were used in a traditional supervised approach, whereas the LLMs were prompted with the zero-shot prompt in Figure 2[6].

4.3 Explanation Approaches

Regarding XAI, we used the most popular model-agnostic options available, SHAP (Lundberg and Lee, 2017) and LIME (Ribeiro et al., 2016), to identify the most relevant features for a given prediction or overall. Since we explore LLMs as classifiers, we also use them as explainers, providing a fully generative option for our approach. We also include two baseline approaches, "Random" and "Unimportant", to be detailed next, as well as the human-annotated rationales which identify the relevant keywords for two annotators.

We use the prompt in Figure 3 to generate an explanation[7]. The symbol ";" is reinforced to make processing of the keywords easier and uniform. In a different version of the prompt, besides the identification of words, we request also expressions or spans. We simply add this in the prompt, e.g., where it reads "Identifica as palavras" (Identify the words) we change it to "Identifica as palavras e expressões" (Identify the words and expressions). It is expected that identifying only words will more easily compare with the XAI models, which do not consider spans, and that including expressions will more easily compare with the human-annotated rationales, which include both forms.

```
Classifica o seguinte texto como pertencente a
uma das seguintes categorias: Ofensivo ou
Não-Ofensivo. Indica apenas a categoria.
Texto: <TEXT>
```

Figure 2: Prompt template for classification, where <TEXT> is replaced by the example to be classified.

```
Identifica as palavras presentes no texto que
justificam a sua classificação como
<CATEGORY>, por ordem decrescente de
importância para esta classificação. Indica
apenas essas palavras, cada uma seguida do
símbolo ";". Caso seja apenas detetada uma
palavra, deve igualmente estar seguida do
símbolo ";". Omite qualquer texto extra ou
explicação da seleçao das palavras. Texto:
<TEXT>
```

Figure 3: Prompt template for generating an explanation, where <TEXT> is replaced by the dataset example being analysed and <CATEGORY> is replaced by the model prediction for that example.

[6]The prompt translates to "Classify the following text as belonging to one of these categories: Offensive or Not-Offensive. Indicate only the category. Text: <TEXT>".

[7]The prompt translates to "Identify the words present in the text that justify its classification as <CATEGORY>, in descending order of importance for this classification. Indicate only these words, followed by the symbol ";". In case only one word is detected, it should also be followed by the symbol ";". Omit any extra text or explanation of the selection of words. Text: <TEXT>".

4.4 Keyword Selection

The selection of keywords for subsequent removal is one of the most crucial aspects of our methodology. Before delving into the explored options, two baselines were created: "Random" and "Unimportant". Both randomly pick K words of each example, that will then be removed and analysed. The difference between them is that "Unimportant" selects only from a list of not-relevant keywords, as it accesses the human-annotated rationales for the offensive class and removes the human-identified keywords from the list of options. Regarding these rationales, a combination of the keywords or spans selected by each annotator was created for further analysis, by performing the union of both groups.

We developed three keyword selection methods:

1. **Top-K**, in which the K most relevant keywords are identified for removal. We leverage the existing statistical analysis of the rationale annotations (Salles et al., 2025) to define the range of K, [1,3,5,7]. When not available, the range can be defined by exploring the distribution of the examples' length;

2. **Top-W**, in which we define the maximum cumulative weight value, W, across the selected words. This only applies to SHAP and LIME, as they provide scores for each feature. Any number of words can be selected, as long as the sum of their weights does not surpass W. The range of values differs based on the classifier and XAI model used, and we will analyse this further ahead;

3. **Min-W**, in which we define the minimum weight value, W, that a word must have to be selected for removal. Any number of words can be selected, as long as their weight is at least W. Similarly to the previous option, this only applies to SHAP and LIME, and the range of values differs between combinations of classifier and XAI model.

To define the range of values for each classifier and explanation models, a brief statistical analysis was performed on the ten most relevant features identified, which can be seen in Table 2. Due to the selection of ten features, each of which has a mean or median score, results are displayed as a range.

Whereas with SHAP we retrieve the global scores of each feature, with LIME, a local approach, all the examples need to be processed and the features and their scores aggregated. Then, the data is sorted and the ten largest values are selected.

Model	XAI	Mean	Std	Median	Max	Min
RF	LIME	0.24–0.33	0.20	0.22–0.32	0.70	0.05
RF	SHAP	0.01–0.02	0.02	0.01–0.02	0.46	0.00
XGB	LIME	0.26–0.43	0.16	0.29–0.49	0.66	0.04
XGB	SHAP	0.06–0.21	0.40	0.04–0.15	4.00	0.00

Table 2: Statistical analysis of the selected features.

The analysis of the statistical distribution of each experiment shows that the scores change greatly based on the XAI model, and not so much based on the classifier, except for SHAP, where the range of values differs greatly from RF to XGB. In fact, the latter has large standard deviation and maximum values, so larger thresholds for Top-W will be used. The resulting range of values for each experiment can be seen in Table 3.

Model	XAI	Top-K	Top-W	Min-W
RF	LIME	[1, 3, 5, 7]	[0.3, 0.5, 0.7]	[0.01, 0.05, 0.10, 0.20]
RF	SHAP	[1, 3, 5, 7]	[0.1, 0.3, 0.5]	[0.01, 0.02, 0.05, 0.10]
XGB	LIME	[1, 3, 5, 7]	[0.3, 0.5, 0.7, 1.0]	[0.10, 0.20, 0.50, 1.00]
XGB	SHAP	[1, 3, 5, 7]	[0.7, 1.0, 3.0, 5.0]	[0.10, 0.20, 0.50, 1.00]

Table 3: Selected range of values for each keyword selection method, based on the classifier and XAI model.

4.5 Implementation Details

In this section, we describe the technical specifics and implementation decisions made. The packages used include *scikit-learn* for RF, TF-IDF and evaluation (with a dataset split of 80-20); *xgboost* for XGBoost; *lime* and *SHAP* (TreeShap) for XAI; *NLTK* for detokenisation; *regex* for tokenisation and word removal; *numpy* and *pandas* for data structures and operations; *matplotlib* for the creation of graphs. When applicable, a random seed of 42 was set.

Regarding ensembles, the following hyperparameters were used: (1) for RF: *n_estimators* was 200; (2) for XGBoost: *eta* was 0.2, *n_estimators* was 300, and *objective* was binary:logistic. For LLMs, we used Llangchain and set a temperature of 0.0 to maximise determinism.

For word tokenisation and search, regex was used for coherence with LIME's tokeniser (\w+).

Regarding word representation, when necessary, TF-IDF was applied due to its interpretable nature. No hyperparameters were set as to not remove any keyword that could be relevant for the model.

There were some scenarios that required consideration and adjustments, e.g., when a keyword appears more than once in the example or when keyword-removal would lead to an empty string. Regarding the former, our solution was to count each occurrence of the word for the top-k, except if this was the last word selected, in which case we overrule the top-k value and remove all occurrences of this word, as it is not certain which is the most relevant. Regarding the former, we chose to classify examples with insufficient words as 0, which can be considered the default state, rather than removing them, in order to preserve comparable experiments, i.e., with the same examples. This constraint should be considered when analysing the results in the next section.

Whereas to define the range of values for the Top-W and Min-W keyword selection methods, SHAP and LIME (aggregated) were used for a global analysis, when selecting and removing keywords from a given sample, they were defined as local models, looking at the scores of each feature for that specific instance.

5 Results and Discussion

In this section, we describe the metrics and experiments developed and discuss the results. The following aspects were analysed: (i) evaluation of explanations based on the impact of keyword-removal on performance; (ii) keyword-removal methods; (iii) keyword roles; (iv) similarity with human-annotated rationales and annotator impact. When human rationales are used, they correspond to the union of the keywords selected by both annotators.

5.1 Evaluation of Explanations

Our evaluation of explanations is grounded on the assumption that features identified by a faithful explanation should have a measurable impact on the classifier's performance when removed from the input. To quantify this, we introduce two metrics: PerfDiff, which measures the difference between the original performance and the performance after keyword removal, and RandomDiff, which measures the distance between the post-removal performance and random behaviour (50% accuracy for a binary task). While a faithful explanation is expected to drive performance closer to random, PerfDiff is also necessary to contextualize the magnitude of the performance drop.

Figure 4 shows the lowest performance obtained for each combination of explanation and classification approach. Across most settings, it is clear that there is a large drop in performance upon keyword-removal, often close to 50%. SHAP and LIME have the highest impact, suggesting that they provide faithful explanations. For LLM-based explanations, we compare two prompting strategies, one requesting only word identification (LLM1), and another including expressions (LLM2). Results show minimal differences between these prompts in terms of performance impact. As expected, the removal of unimportant words is the baseline with the least impact. However, it should be noted that only keywords for the hate speech class were considered, as they were the only human-annotated examples, so the real impact should be even smaller.

Table 4 presents the metrics for each of the approaches followed, allowing us to verify once again that LIME and SHAP provide good explanations, as they show the highest PerfDiff and the lowest RandomDiff. The LLM experiments show higher RandomDiff than the random baseline, where keyword selection is randomised, casting doubts on their adoption as explainers.

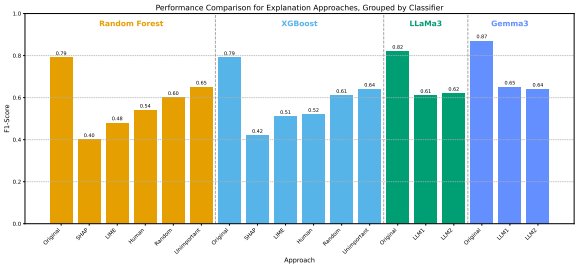

Figure 4: F1-Score comparison for the best explanation approaches for each classifier.

Approach	Original	PerfDiff	RandomDiff
RF-Random (Top-K 7)	0.79	-19pp	10pp
RF-Unimp. (Top-K 7)	0.79	-14pp	15pp
RF-LIME (Top-K 7)	0.79	-28pp	**1pp**
RF-SHAP (Top-W 0.5)	0.79	**-39pp**	10pp
RF-Human	0.79	-25pp	4pp
XGB-Random (Top-K 7)	0.79	-18pp	11pp
XGB-Unimp. (Top-K 7)	0.79	-15pp	14pp
XGB-LIME (Top-W 1.0)	0.79	-28pp	**1pp**
XGB-SHAP (Min-W 0.1)	0.79	**-37pp**	8pp
XGB-Human	0.79	-27pp	2pp
LLaMa3-LLM1	0.82	-21pp	11pp
LLaMa3-LLM2	0.82	-20pp	12pp
Gemma3-LLM1	0.87	-22pp	15pp
Gemma3-LLM2	0.87	-23pp	14pp

Table 4: Impact analysis for each combination of approaches, regarding their F1-Score. When applicable, the keyword-removal method is described.

5.2 Keyword-removal Methods

Previously, only the lowest-score results were presented, but several approaches were evaluated, especially regarding the keyword-removal methods: Top-K, Top-W, and Min-W, with the two latter applying only to XAI and ensemble models.

Figure 5 illustrates the effect of Top-K on performance, showcasing the difference between the random baselines and the use of XAI models like SHAP and LIME. There is a clear drop as the value of K increases, but with only the removal of three keywords, it is possible to see a large impact on performance, especially for XGBoost.

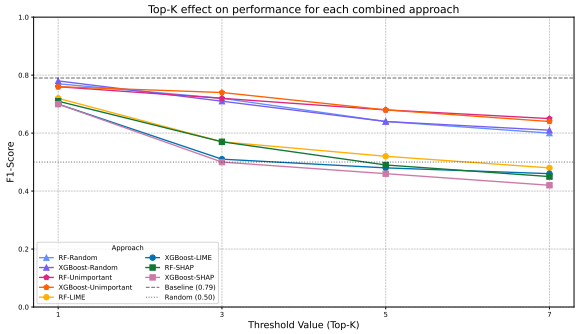

Figure 5: F1-Score comparison for each model, explanation approach, and top-k threshold value.

Figure 6 shows the impact of Top-W on performance, with SHAP presenting the highest effect but also higher divergence regarding values of W, contrary to LIME, where, for both ensembles, the range of values is similar.

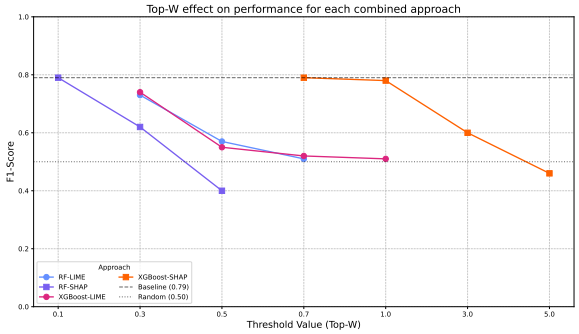

Figure 6: F1-Score comparison for each model, explanation approach, and top-w threshold value.

Figure 7 presents the impact of Min-W on performance, where the trend is contrary to the previous methods, as a higher Min-W means fewer words are selected for removal, hence less impact. Again, SHAP presents the lowest performances and the value range is very diverse, as opposed to LIME.

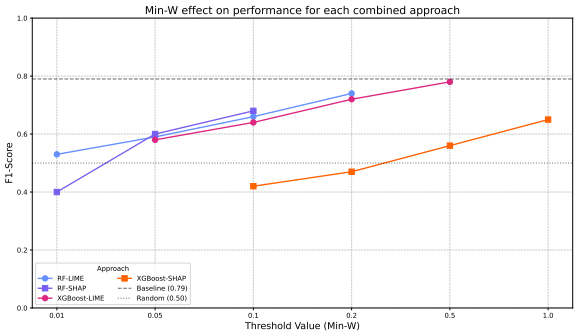

Figure 7: F1-Score comparison for each model, explanation approach, and min-w threshold value.

Whereas the largest impact is frequently observed via the removal of the Top-K words, for SHAP a weight-based approach is more impactful.

5.3 Keyword Role Analysis

We analyse the linguistic roles of the keywords identified for the explanations, namely regarding the purpose of function, e.g., *e* (and), *esse* (that), or content, e.g., *lindo* (beautiful), *traidores* (traitors).

Table 5 presents the average number of words with function (FRole) and content (CRole) roles. Three aspects should be noted: (i) in the human annotations, only one class is represented; (ii) Gemma3 required some post-processing as it repeatedly identified certain keywords, affecting the statistical analysis, so only unique keywords were considered; and (iii), the role analysis is based on an approximation using NLTK's Portuguese stopwords as a proxy for the functional role.

Human-annotated explanations include more content than functional words, which makes sense in a hate speech scenario, where stopwords are less likely to be relevant. However, these words are still selected across all experiments, but less so when LLMs are involved. Furthermore, SHAP tends to select more keywords in comparison with the other approaches, as confirmed in related work (Salles et al., 2025), especially content words. LIME seems much closer to human-annotated explanations regarding these linguistic roles.

5.4 Human Similarity and Annotator Impact

Finally, we analyse the impact of individual annotators and the similarity between human rationales and automatically generated explanations.

Table 6 shows the impact on performance when the keywords removed are the ones identified by each annotator and the combination (union) of

Approach	FRole	CRole
RF-Random	1.76±1.73	2.43±2.20
RF-Unimp.	1.56±1.66	2.21±2.13
RF-LIME	1.77±1.73	2.42±2.20
RF-SHAP	4.35±4.38	7.50±7.85
RF-Human	1.69±2.87	3.78±4.23
XGB-Random	1.75±1.71	2.44±2.20
XGB-Unimp.	1.59±1.68	2.18±2.10
XGB-LIME	1.55±1.54	2.64±2.34
XGB-SHAP	1.16±1.27	3.03±2.62
XGB-Human	1.69±2.87	3.78±4.23
LLaMa3-LLM1	0.20±0.55	2.26±1.29
LLaMa3-LLM2	0.51±1.37	2.98±3.09
Gemma3-LLM1	0.25±0.73	3.69±2.32
Gemma3-LLM2	0.40±0.87	3.80±2.40

Table 5: Analysis of keyword roles, function (FRole) and content (CRole), for each approach.

both. Regardless of classification model, annotator one has a five-point difference regarding annotator two, suggesting that they better identified relevant and faithful keywords. However, this does not mean that for a human reading the explanations this would remain the case, as we are only analysing the impact on ML models. Related work (Salles et al., 2025) has also shown that annotator one considers a larger number of words and also more nouns and verbs, compared to annotator two, which may have an impact on this result.

Model	Annotator	Score
RF	Annotator1	**0.54**
RF	Annotator2	0.59
RF	Combined	0.55
XGB	Annotator1	**0.52**
XGB	Annotator2	0.57
XGB	Combined	0.53

Table 6: Post-removal effect on performance (F1-Score) by annotator.

Table 7 shows the average Jaccard Similarity scores for each approach. It is clear that the unimportant baseline barely has any similarity with the human-annotated rationales, as expected, since we deliberately exclude their rationales from the possible keywords to be identified for class one. Nevertheless, the vast majority of approaches shows weak similarity with the human rationales, with the exception of XGB-LIME and Gemma3-LLM2, which present moderate similarity. However, in most cases, the standard deviation is high and could lead to moderate-to-high similarity in some examples. It should be noted that this analysis was done on the token-level, despite human-annotators including spans in their annotations, as for most approaches, only tokens are identified as keywords.

Approach	Similarity
RF-Random	0.21±0.25
RF-Unimp.	0.09±0.13
RF-LIME	0.22±0.25
RF-SHAP	0.25±0.22
XGB-Random	0.21±0.25
XGB-Unimp.	0.09±0.13
XGB-LIME	0.40±0.31
XGB-SHAP	0.12±0.14
LLaMa3-LLM1	0.23±0.24
LLaMa3-LLM2	0.25±0.24
Gemma3-LLM1	0.28±0.25
Gemma3-LLM2	0.30±0.26

Table 7: Jaccard Similarity regarding the combined human-annotated rationales for the hate speech class.

6 Conclusion and Future Work

We presented an unsupervised, performance-based approach for evaluating explanation faithfulness in Portuguese hate speech classification. Results demonstrate that performance degradation following the removal of explanatory keywords provides a meaningful proxy for faithfulness, particularly when compared against random-removal baselines.

Experiments on the HateBRXplain dataset demonstrated that formal XAI methods, e.g. SHAP and LIME, consistently identify more faithful features than manually annotated rationales or LLM-generated explanations and simple keyword-removal strategies like Top-K and weight-based approaches proved effective.

Our analysis highlights a divergence between plausibility and faithfulness, as high-impact approaches did not consistently correlate with human rationales. In addition, LLM-generated explanations tended to focus on content words and exhibited limited faithfulness under performance-based evaluation, suggesting unreliability as explainers.

Overall, this work underscores the need for standardised, unsupervised evaluation protocols for explanation faithfulness and contributes empirical evidence on the limitations and strengths of current explanation methods.

Future work includes: (i) exploring more XAI models; (ii) comparing proprietary versus open-source LLMs of varying scales; (iii) testing keyword selection through semantic, e.g., antonym replacement; and (iv), integrating Part-of-Speech tagging to enhance the linguistic analysis.

Limitations

This work presents some limitations, part of which are derived from the data available, and its format, reinforcing the importance of developing more resources for XAI research, particularly in Portuguese.

The comments in the data are short with approximately 1 sentence and 12-15 words per comment, which inherently makes each word more important. Moreover, there are only human-annotated rationales for one of the classes, which limits some of the analysis performed, but does not affect our approach, due to its unsupervised nature.

Whereas the methodology is generalisable, the experiments are conducted on a single task and dataset. To make the findings more robust, more tasks and domains should be explored. Furthermore, to ensure a rich explanation quality analysis, other metrics can be added, e.g., consistency, completeness, as explainability is a multi-faceted concept.

Acknowledgments

This work was financed by the Portuguese Recovery and Resilience Plan (PRR), through project C645008882-00000055 – Center for Responsible AI.

This work was also supported by FCT – Foundation for Science and Technology, I.P., within the scope of the research unit UID/00326 - Centre for Informatics and Systems of the University of Coimbra. Isabel Carvalho was supported by FCT – Foundation for Science and Technology, I.P. through the PhD scholarship with reference 2023.04883.BD.

References

Amina Adadi and Mohammed Berrada. 2018. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, 6:52138–52160.

Chirag Agarwal, Sree Harsha Tanneru, and Himabindu Lakkaraju. 2024. Faithfulness vs. Plausibility: On the (Un)Reliability of Explanations from Large Language Models. *Preprint*, arXiv:2402.04614.

Vijay Arya, Rachel K. E. Bellamy, Pin-Yu Chen, Amit Dhurandhar, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Q. Vera Liao, Ronny Luss, Aleksandra Mojsilović, Sami Mourad, Pablo Pedemonte, Ramya Raghavendra, John Richards, Prasanna Sattigeri, Karthikeyan Shanmugam, Moninder Singh, Kush R. Varshney, Dennis Wei, and Yunfeng Zhang. 2019. One Explanation Does Not Fit All: A Toolkit and Taxonomy of AI Explainability Techniques. *Preprint*, arXiv:1909.03012.

Diedre Carmo, Marcos Piau, Israel Campiotti, Rodrigo Nogueira, and Roberto Lotufo. 2020. PTT5: Pretraining and validating the T5 model on Brazilian Portuguese data. *Preprint*, arXiv:2008.09144.

Tiago Barbosa de Lima, Vitor Rolim, André C. A. Nascimento, Péricles Miranda, Valmir Macario, Luiz Rodrigues, Elyda Freitas, Dragan Gašević, and Rafael Ferreira Mello. 2024. Towards explainable automatic punctuation restoration for Portuguese using transformers. *Expert Systems with Applications*, 257:125097.

Frank Dietrich. 2025. AI-based removal of hate speech from digital social networks: Chances and risks for freedom of expression. *AI and Ethics*, 5(3):2943–2953.

Mohsen Fayyaz, Fan Yin, Jiao Sun, and Nanyun Peng. 2024. Evaluating Human Alignment and Model Faithfulness of LLM Rationale. *Preprint*, arXiv:2407.00219.

Suparna Ghanvatkar and Vaibhav Rajan. 2024. Evaluating Explanations From AI Algorithms for Clinical Decision-Making: A Social Science-Based Approach. *IEEE Journal of Biomedical and Health Informatics*, 28(7):4269–4280.

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*, 7(1):587–623.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The Llama 3 Herd of Models. *Preprint*, arXiv:2407.21783.

David Gunning. 2016. Broad agency announcement explainable artificial intelligence (XAI).

Elize Herrewijnen, Dong Nguyen, Floris Bex, and Kees van Deemter. 2024. Human-annotated rationales and explainable text classification: A survey. *Frontiers in Artificial Intelligence*, 7.

Shuaibo Hu and Kui Yu. 2024. Learning Robust Rationales for Model Explainability: A Guidance-Based Approach. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):18243–18251.

Marcio Inácio, Gabriela Wick-pedro, and Hugo Gonçalo Oliveira. 2023. What do humor classifiers learn? An attempt to explain humor recognition models. In *Procs of 7th Joint SIGHUM Workshop on*

Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature, pages 88–98, Dubrovnik, Croatia. ACL.

Tharindu Kumarage, Amrita Bhattacharjee, and Joshua Garland. 2024. Harnessing Artificial Intelligence to Combat Online Hate: Exploring the Challenges and Opportunities of Large Language Models in Hate Speech Detection.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the Middle: How Language Models Use Long Contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Scott M Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems*, volume 30, pages 4765–4774, New York, NY, USA. Curran Associates, Inc.

Andreas Madsen, Sarath Chandar, and Siva Reddy. 2024. Are self-explanations from Large Language Models faithful? In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 295–337, Bangkok, Thailand. Association for Computational Linguistics.

Christoph Molnar. 2020. *Interpretable Machine Learning*. Lulu.com.

Meike Nauta, Jan Trienes, Shreyasi Pathak, Elisa Nguyen, Michelle Peters, Yasmin Schmitt, Jörg Schlötterer, Maurice van Keulen, and Christin Seifert. 2023. From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI. *ACM Comput. Surv.*, 55(13s):295:1–295:42.

Mikel K. Ngueajio, Saurav Aryal, Marcellin Atemkeng, Gloria Washington, and Danda Rawat. 2025. Decoding Fake News and Hate Speech: A Survey of Explainable AI Techniques. *ACM Comput. Surv.*, 57(7):169:1–169:37.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 1135–1144, New York, NY, USA. Association for Computing Machinery.

Luisa Ribeiro-Flucht, Xiaobin Chen, and Detmar Meurers. 2024. Explainable AI in Language Learning: Linking Empirical Evidence and Theoretical Concepts in Proficiency and Readability Modeling of Portuguese. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 199–209, Mexico City, Mexico. Association for Computational Linguistics.

Isadora Salles, Francielle Vargas, and Fabrício Benevenuto. 2025. HateBRXplain: A Benchmark Dataset with Human-Annotated Rationales for Explainable Hate Speech Detection in Brazilian Portuguese. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6659–6669, Abu Dhabi, UAE. Association for Computational Linguistics.

Ramprasaath R. Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. 2017. Grad-CAM: Why did you say that? *Preprint*, arXiv:1611.07450.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. BERTimbau: Pretrained BERT Models for Brazilian Portuguese. In *Intelligent Systems*, pages 403–417, Cham. Springer International Publishing.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean-bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 Technical Report. https://arxiv.org/abs/2503.19786v1.

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 13th Conference on Language Resources and Evaluation (LREC 2022)*, pages 7174–7183, Marseille, France. European Language Resources Association.

David S. Watson, Jenny Krutzinna, Ian N. Bruce, Christopher EM Griffiths, Iain B. McInnes, Michael R. Barnes, and Luciano Floridi. 2019. Clinical applications of machine learning algorithms: Beyond the black box. *BMJ*, 364:l886.

In-On Wiratsin and Chaiyong Ragkhitwetsagul. 2025. Effectiveness of Explainable Artificial Intelligence (XAI) Techniques for Improving Human Trust in Machine Learning Models: A Systematic Literature Review. *IEEE Access*, 13:121326–121350.