Self-Explaining Hate Speech Detection with Moral Rationales

Francielle Vargas
University of Chile

Jackson Trager
University of Southern California

Diego Alves
Saarland University

Matteo Guida
University of Melbourne

Surendrabikram Thapa
Virginia Tech

Berk Atil
Pennsylvania State University

Daryna Dementieva
Technical University of Munich

Andrew Smart
Google Research

Ameeta Agrawal
Portland State University

Abstract

Existing hate speech detection models are often opaque and rely on surface-level lexical cues, which makes them vulnerable to spurious correlations and limits robustness, interpretability and cultural contextualization. We propose Supervised Moral Rationale Attention (SMRA)[1], the first self-explaining hate speech detection framework to incorporate moral rationales as direct supervision for attention alignment. Based on Moral Foundations Theory, SMRA aligns token-level attention with expert-annotated moral rationales, guiding models to attend to morally salient spans. Unlike prior rationale-supervised or post-hoc approaches, SMRA integrates moral rationale supervision directly into the training objective, producing inherently interpretable and contextualized explanations. To support our framework, we also introduce HateBRMoralXplain, a Brazilian Portuguese benchmark dataset annotated with hate labels, moral categories, token-level moral rationales, and socio-political metadata. Across binary hate speech detection and multi-label moral sentiment classification, SMRA consistently improves performance while enhancing both faithful and plausible explanations. Although explanations become more concise, sufficiency decreases, indicating more compact and informative rationales. Fairness remains stable, suggesting that improvements in explanation quality do not introduce significant bias trade-offs. [2]

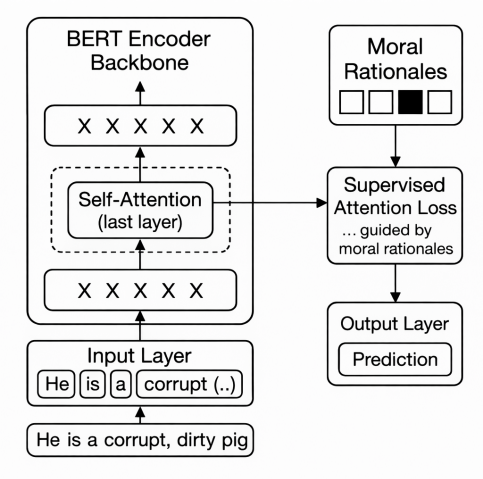

Figure 1: Supervised Moral Rationale Attention (SMRA) for Self-Explaining Hate Speech Detection.

1 Introduction

Despite significant advances in automatic Hate Speech (HS) detection, current approaches remain fundamentally limited. Existing models often encode biases originating from training data (Davidson et al., 2019; Wiegand et al., 2019), stereotypical associations learned by hate speech classifiers (Davani et al., 2023; Vargas et al., 2023), and annotation processes (Sap et al., 2019), which are frequently shaped by annotators' subjective judgments and sociocultural backgrounds (Davani et al., 2024a; Tonneau et al., 2024; Abercrombie et al., 2023; Sap et al., 2022; Vargas et al., 2022b; Poletto et al., 2021). Such biases can lead to systematic disparities in model behavior, resulting in the unfair treatment or over-targeting of marginalized groups and producing negative social consequences when deployed at scale (Blodgett et al., 2020; Wiegand et al., 2019; Sap et al., 2019; Davidson et al., 2019).

Given that these models may inherit biases from their training data, explainable hate speech detection models (Salles et al., 2025a; Eilertsen et al., 2025; Mathew et al., 2021) are crucial for ensuring transparency, diagnosing biases, and enabling trustworthy detection models across diverse languages and cultural contexts. Existing explainable HS detection approaches can be broadly categorized into two interpretability paradigms: post-hoc and self-explaining models. Post-hoc approaches generate explanations after the prediction, relying

[1]Dataset, annotator disagreements, and code are available at https://github.com/franciellevargas/SMRA

[2]**Warning**: This document contains offensive content.

Findings of the Association for Computational Linguistics: ACL 2026, pages 34109–34131

on external attribution or surrogate techniques to approximate the model's reasoning (Trager et al., 2025b; Salles et al., 2025a; Cercas Curry et al., 2023; Mathew et al., 2021; Kennedy et al., 2020). Moreover, several post-hoc explainability methods remain unfaithful and computationally expensive. In contrast, self-explaining models integrate explanation generation directly into the learning process through inherently interpretable mechanisms, e.g., supervision that aligns model attention with human-annotated rationales, making explanations an integral part of the prediction pipeline (Eilertsen et al., 2025; Kim et al., 2022; Mathew et al., 2021).

Furthermore, most existing hate speech detection approaches continue to rely primarily on surface-level lexical features, which are insufficient due to elevated false positive rates (Poletto et al., 2021; Davidson et al., 2019) and their limited ability to account for cultural and contextual factors (Lee et al., 2023; Zhou et al., 2023; Kim et al., 2024; Vargas et al., 2024; Wong, 2024). Even when contextual information is available (Vargas et al., 2021; Kennedy et al., 2020), judgments of offensiveness remain deeply culture-dependent and normatively grounded, challenging models that assume stable or universal interpretations of hate speech (Tonneau et al., 2024; Lee et al., 2023). More fundamentally, existing models fail to capture normative and moral reasoning, limiting their ability to distinguish culturally contingent hate speech from harmless language (Trager et al., 2025b; Tonneau et al., 2024; Davani et al., 2024a).

Recent studies provide empirical evidence that moral values constitute a transferable latent representation for hate speech across languages and cultures (Trager et al., 2025b; Davani et al., 2024a; Solovev and Pröllochs, 2023; Kennedy et al., 2023; Rezapour et al., 2021). Cross-lingual studies based on Moral Foundations Theory show that, despite variation in the lexical realization of hateful content across languages, hate speech targets a recurring set of moral violations and rarely occurs in the absence of moral sentiments (Trager et al., 2025b; Upadhyaya et al., 2023; Solovev and Pröllochs, 2022). A recent study by Davani et al. (2024a) further shows that disagreements in hate speech annotation are shaped by annotators' moral values rather than annotation noise. These findings suggest that standard label aggregation and opaque modeling approaches risk encoding dominant moral norms while obscuring minority perspectives.

To address these limitations, which risk amplifying existing social and cultural biases, **we introduce Supervised Moral Rationale Attention (SMRA), the first framework to align model attention with human-annotated moral rationales for self-explaining hate speech detection**. As illustrated in Figure 1, SMRA introduces a normative inductive regularization into the learning process through a supervised attention loss guided by moral rationales. While lexical markers of hate speech are highly language- and culture-dependent (Tonneau et al., 2024), moral categories and their rationales encode more stable normative structures (Atari et al., 2023; Kennedy et al., 2021), leading to reduced reliance on spurious correlations and improving robustness, interpretability, and cross-cultural generalization.

Experimental results show that SMRA consistently improves both predictive performance and explanation faithfulness and plausibility across binary hate speech detection and multi-label moral sentiment classification. In binary classification, SMRA improves Accuracy ($+0.0004$) and Macro F1 ($+0.0082$), while substantially enhancing rationale alignment, with gains in IoU F1 ($+0.0743$) and Token F1 ($+0.0503$). In addition, SMRA produces more concise explanations, with lower Comprehensiveness (-0.0203) and improved Sufficiency (-0.0231), indicating more compact yet still faithful rationales. In multi-label moral sentiment classification, SMRA further improves Macro F1 ($+0.015$), AUROC ($+0.002$), and Token F1 ($+0.241$), demonstrating stronger rationale coverage and interpretability without degrading predictive performance. Fairness remains stable, suggesting that improvements in explanation quality do not introduce significant bias trade-offs.

In addition, we evaluate the effectiveness of large language models (LLMs) for hate speech and moral sentiment classification. Our findings show that incorporating moral rationales leads to notable improvements in hate speech classification performance ($\approx$ 2–3%), although LLMs perform poorly on moral sentiment classification, with a maximum F1 of 0.38. Finally, we also introduce HateBRMoralXplain, the third version of the HateBR benchmark dataset for low-resource Brazilian Portuguese (Vargas et al., 2022b; Salles et al., 2025a). **To the best of our knowledge, HateBRMoralXplain is the first large-scale expert-annotated corpus of its kind that enriches hate speech data with both moral categories and human-annotated moral rationales**.

Our contributions can be summarized as follows:

1. We introduce Supervised Moral Rationale Attention (SMRA), the first self-explaining framework that aligns model attention with expert-annotated moral rationales to improve interpretability, robustness, and contextualization in hate speech detection;

2. We release HateBRMoralXplain, an extended version of a widely used Brazilian Portuguese hate speech benchmark, augmenting prior datasets with moral categories, moral rationales, and socio-political metadata; the dataset, models, inter-annotator disagreements, and code are publicly available to facilitate future research;

3. We evaluate multiple LLMs on HateBRMoralXplain corpus for hate speech and moral sentiment classification using prompts with different informational components, demonstrating that the inclusion of moral rationales provide statistically significant improvements in hate speech detection performance.

2 Related Work

Self-Explaining Hate Speech Detection. Moving toward intrinsic explainability, Supervised Rational Attention (SRA) (Eilertsen et al., 2025) aligns transformer attention with human rationales and substantially improves rationale faithfulness; however, its reliance on lexical rationales limits robustness, cross-cultural generalization, and the ability to capture deeper moral reasoning. Together, these limitations motivate self-explaining frameworks that explicitly incorporate moral context beyond surface-level cues. Nirmal et al. (2024) present SHIELD for interpretable hate speech detection by incorporating GPT-3.5–generated textual rationales as additional input. The text and rationales are encoded separately using HateBERT and a frozen BERT, respectively, and their representations are concatenated for classification. This design promotes interpretability by grounding predictions in explicit rationale representations. Calabrese et al. (2022) formulate Intent Classification and Slot Filling (ICSF) task. The approach uses a two-stage BART-based sequence-to-sequence model that first produces a semantic sketch specifying relevant slots (e.g., Target, ProtectedCharacteristic), and then populates these slots with the corresponding text spans. The abuse intent (e.g., Dehumanization, Derogation) is deterministically derived from the filled slots, making the prediction explicitly grounded in extracted evidence. This structured output provides an intrinsic, human-interpretable explanation, rather than a post-hoc rationale. Kim et al. (2022) introduces an intermediate token-level task in which portions of the rationale embeddings are masked, and the model is trained to reconstruct the masked rationale labels using contextual information. This rationale prediction objective, optimized via cross-entropy loss on masked tokens, encourages context-aware reasoning before fine-tuning the model for hate speech classification.

Hate Speech with Moral Rationales. Recent work underscores the role of moral rationales in advancing hate speech detection beyond surface-level cues. MFTCXplain (Trager et al., 2025b) introduces a multilingual benchmark with expert annotations of hate labels, Moral Foundations Theory (MFT) categories, and span-level moral rationales, showing that while state-of-the-art LLMs perform reasonably in classification, they struggle to predict moral categories and generate faithful explanations. Complementarily, Davani et al. (2024a) demonstrate that although the lexical realization of moral values varies across languages, moral categories provide consistent predictive signals that improve cross-lingual robustness and transfer. Similarly, TWISTED (Upadhyaya et al., 2023) jointly models toxicity, moral values, and speech acts, thus producing consistent gains across benchmarks. Together, these works indicate that explicitly modeling moral dimensions supports more robust and generalizable hate speech detection while reducing reliance on spurious lexical correlations. Further literature on the relationship between hate and morality is discussed in Appendix B.

3 HateBRMoralXplain Corpus

We introduce HateBRMoralXplain, the third version of the HateBR corpus (Vargas et al., 2022b, 2024), extending both the original HateBR dataset and its explainable successor, HateBRXplain (Salles et al., 2025a). As in prior versions, the corpus consists of 7,000 Brazilian Portuguese Instagram comments extracted from public political accounts, annotated by three different experts. The dataset includes 3,500 offensive and 3,500 non-offensive comments, and hate speech rationales.

In the HateBRMoralXplain corpus, comments are annotated for moral categories grounded in Moral Foundations Theory (MFT) (Graham et al., 2013). We annotate all five core moral foundations, *Care/Harm*, *Fairness/Cheating*, *Loyalty/Betrayal*, *Authority/Subversion*, and *Purity/Degradation*, capturing both virtue and vice oriented moral expressions within a unified framework. Each comment may be assigned between one and three moral labels and their rationales, ordered by annotators according to their perceived salience in the text, thus representing primary, secondary, and tertiary moral rationales.

We selected two expert annotators with diverse backgrounds and perspectives in Brazil[3], an important design choice given that hate speech and moral judgments are inherently subjective and shaped by annotators' beliefs, identities, and sociocultural contexts (Prabhakaran et al., 2021; Sap et al., 2022). Inter-annotator agreement is measured using Cohen's weighted Kappa with quadratic weights, computed separately for each moral category to account for the multi-label nature of the annotation scheme (i.e., Class A refers to the first moral label, Class B the second, and Class C the third). Results are shown in Appendix A.1. Class A shows substantial to almost perfect agreement ($\kappa = 0.811$), while Classes B and C exhibit moderate to substantial agreement ($\kappa = 0.671$ and $\kappa = 0.612$, respectively), reflecting higher subjectivity in these categories.

Beyond categorical labels, HateBRMoralXplain includes human-annotated rationales for moral judgments. Rationales are defined as minimal text spans that justify why a given moral label was assigned, highlighting the specific linguistic evidence supporting the annotation. Moral rationale annotation follows the same span-based protocol adopted in MFTCXplain (Trager et al., 2025b). These rationales enable explainability, allowing models to ground their predictions in interpretable textual moral evidence rather than relying solely on surface-level hateful lexical cues. Detailed rationale annotation guidelines and examples are provided in Appendix A.1. Finally, the corpus includes rich socio-political metadata to support contextual and demographic analyses. Each comment is linked to its parent Instagram post and associated metadata indicating the political party and gender of the politician who authored the original post. This structure enables analyses of how hate speech and moral framing vary across political affiliation, gender, and discourse context, while preserving the original data collection constraints and privacy safeguards established in HateBR (Vargas et al., 2022a). An example of HateBRMoralXplain is shown in Appendix A.1.

[3]Annotators completed a pre-annotation survey measuring demographics, political ideology, moral values, personality, and cultural orientations, to increase transparency and support analyses of subjective decisions, See Appendix A.1

4 Supervised Moral Rationale Attention

Supervised Moral Rationale Attention (SMRA) is the first self-explaining framework that aligns neural attention with moral rationales. SMRA enhances standard transformer-based text classifiers by explicitly aligning model attention with expert-annotated moral rationales based on Moral Foundation Theory for hate speech detection, following the attention alignment framework of Eilertsen et al. (2025). Our SMRA framework is formally described as follows.

4.1 Task Definition

We address two related tasks: *binary hate speech classification* and *multi-label moral sentiment classification*. Given an input text, the first task consists of predicting whether the content constitutes hate speech. The second task aims to identify which moral foundations, as defined by Moral Foundations Theory, are expressed in the text. Supervised Moral Rationale Attention (SMRA) introduces normative inductive regularization into the learning process through a supervised attention loss guided by moral rationales. The problem may be formalized as follows: Let $x = (w_1, \ldots, w_L)$ denote a tokenized input sequence of length L, and let $y \in \{0, 1, \ldots, C-1\}$ denote its class label, where C is the number of moral categories. In the HateBRMoralXplain benchmark, instances are annotated with one or more moral categories defined as $\texttt{LABELS} = \{\text{NN}, \text{HN}, \text{FN}, \text{PN}, \text{AN}, \text{LN}\}$, where NN denotes *Non-Morality*, HN *Harm/Care*, FN *Fairness/Cheating*, PN *Purity/Degradation*, AN *Authority/Subversion*, and LN *Loyalty/Betrayal*. Thus, $C = 6$. For a subset of the training instances, we additionally provide a binary *rationale mask* $r = (r_1, \ldots, r_L)$, where $r_i \in \{0, 1\}$ indicates whether token w_i is part of the human-annotated moral rationale associated with label y.

4.2 Model Architecture and Attention Mechanism

We use a pre-trained transformer encoder f_θ (e.g., `BERTimbau` and `mBERT`) to compute contextual representations $\mathbf{h}_i$ for each input token w_i. A classification head predicts the moral category based on the [CLS] representation:

$$\hat{y} = \arg \max_{c \in \{0,1,\ldots,C-1\}} \text{softmax}(\mathbf{W}_c \mathbf{h}_{\text{[CLS]}} + \mathbf{b}_c),$$

where $\mathbf{h}_{\text{[CLS]}} \in \mathbb{R}^d$ denotes the contextual embedding of the [CLS] token, and $\mathbf{W}_c \in \mathbb{R}^{1 \times d}$ and $\mathbf{b}_c \in \mathbb{R}$ are the learned parameters for class c.

Moral Attention Alignment Loss. For samples with available moral rationales, we encourage the model to attend to tokens identified as morally salient by minimizing the Mean Squared Error (MSE) between the normalized attention distribution $\mathbf{a}$ and the rationale mask r.

The final training objective combines a cross-entropy loss ℓ_{CE} for classification with a supervised attention alignment loss ℓ_{MSE}, which serves as an auxiliary loss. The latter is computed as the mean squared error between model attention and human-annotated moral rationales and is applied only to samples with available moral rationales. The overall objective is given by ℓ_{total}:

$$\ell_{\text{total}} = \ell_{\text{CE}} + \alpha \cdot \ell_{\text{MSE}}. \quad (1)$$

where α controls the strength of attention supervision. The moral attention alignment loss is applied only to instances associated with moral categories and for which human-annotated moral rationales are available.

Token-Level Binary Masks for Moral Rationales. The HateBRMoralXplain dataset provides *text rationale spans* for each moral sentiment category. Rationale masks r are constructed by mapping character-level spans to token indices using the tokenizer's offset mappings and are aligned with the tokenization of $x = (w_1, \ldots, w_L)$, with truncation or padding applied to match the model's maximum input length.

5 Experimental Setup

5.1 Model Architecture and Settings

We evaluate three model setups:

- standard fine-tuning of mBERT[4] and BERTimbau[5] (Souza et al., 2020),
- the same models fine-tuned with supervised attention using moral rationales (SMRA), and
- prompting LLMs with a wide range of hateful and moral information.

Fine-Tuning mBERT and BERTimbau: We used `HuggingFace Transformers` and `PyTorch` to fine-tune two BERT models. In our setup, the model embeddings were fed into a linear layer for prediction. We used a batch size of 16, a learning rate of 2e-5, and 128 maximum sequence length. Training was performed for 20 epochs. The dataset was split into training (80%), validation (10%), and test (10%) sets. Optimization was performed using the AdamW optimizer (Loshchilov and Hutter, 2017) with cross-entropy loss. The ground truth for moral rationales corresponds to the annotation provided by the first annotator.

Supervised Moral Rationale Attention: Our SMRA framework was implemented using `HuggingFace Transformers` and `PyTorch`, fine-tuning pre-trained BERT-based models (BERTimbau and mBERT) with attention supervision using BertTokenizer[6]. During training, [CLS]-to-token attention weights from the last encoder layer were normalized and aligned with human-annotated moral rationale masks via a mean squared error (MSE) loss, combined with standard cross-entropy loss using a weighting hyperparameter α, lambda attention and the attention weight 0.001. Attention supervision was applied only to examples with moral content, and rationale masks were padded and masked to handle variable-length sequences. Optimization used `AdamW` with weight decay, training on GPU in mini-batches.

LLMs: For both GPT-4o-mini (through OpenAI API) (OpenAI, 2024) and Llama3.1-70b (on local RTX A6000 GPUs) (Grattafiori et al., 2024), we use a temperature of 0 for reproducibility. All prompts are provided in Appendix C. For our experiments that use translated versions of comments, we use the Google Translate API. For these experiments, we test the effects of different information: hate speech definition, data collection context, and

[4] https://huggingface.co/google-bert/bert-base-multilingual-cased

[5] https://huggingface.co/neuralmind/bert-base-portuguese-cased

[6] https://huggingface.co/docs/transformers/model_doc/bert

predicting moral and hate speech together vs separately. We evaluate several prompt configurations that differ in task scope and contextual information: The `hate` setting performs only hate speech classification without providing definitions or moral rationales, whereas `hate w/ definition` includes the definition of hate speech. The `hate moral` configuration jointly requests hate speech classification, moral value classification, and an explanation explicitly describing the MFT. Contextual information about data collection is incorporated in `hate w/ context`, which focuses solely on hate speech classification, and in `hate moral w/ context`, which extends `hate moral` by additionally providing dataset context. Finally, moral-only settings are explored through `moral`, which requests moral sentiment classification and their rationales, `moral w/ definition`, which further includes the hate speech definition, and `moral w/ context`.

5.2 Evaluation Metrics

Following state-of-the-art evaluation metrics for hate speech classification, explainability, and fairness/bias (Mathew et al., 2021; Attanasio et al., 2022), we evaluate our models using the metrics described below.

Classification: For both LLM and fine-tuned mBERT and BERTimbau experiments, we report the *Macro F1* score for hate speech and moral sentiment classification. Let $y_i \subseteq \mathcal{Y}$ denote the set of human-annotated moral labels for instance i, and let $\hat{y}_i \in \mathcal{Y}$ be the model's predicted label. For moral value classification, we define an *adapted correctness function*:

$$\text{correct}(\hat{y}_i, y_i) = \begin{cases} 1 & \text{if } \hat{y}_i \in y_i, \\ 0 & \text{otherwise,} \end{cases}$$

Explainability: We evaluate SMRA's explanations using established metrics from the interpretability literature (DeYoung et al., 2020): *plausibility*, which measures human-alignment of generated explanations, and *faithfulness*, which assesses whether the explanation accurately reflects the model's decision process.

Plausibility is measured via token-level IOU F1 and Token-F1 (DeYoung et al., 2020), comparing model rationales M_i with human-annotated moral rationales H_i for instance i:

$$\text{IOU-F1} = \frac{1}{N}\sum_{i=1}^{N} \text{Greater}\left(\frac{|M_i \cap H_i|}{|M_i \cup H_i|}, 0.5\right) \quad (2)$$

$$\text{Token-F1} = \frac{1}{N}\sum_{i=1}^{N} \frac{2|M_i \cap H_i|}{|M_i| + |H_i|} \quad (3)$$

Faithfulness is assessed through *comprehensiveness* and *sufficiency* (DeYoung et al., 2020).

Comprehensiveness evaluates the influence of predicted moral rationales r_i by measuring the drop in predicted probability after removing them:

$$\text{Comp} = \frac{1}{N}\sum_{i=1}^{N} \left(m(x_i)_j - m(x_i \backslash r_i)_j\right) \quad (4)$$

Sufficiency measures whether rationales alone are sufficient for prediction:

$$\text{Suff} = \frac{1}{N}\sum_{i=1}^{N} \left(m(x_i)_j - m(r_i)_j\right) \quad (5)$$

Fairness/Bias: We used the fairness/bias metrics (Borkan et al., 2019; Dixon et al., 2018) to compute a specific identity-term (subgroup distribution) and the rest (background distribution). The three per-term bias scores are: **GMB-Sub**, which measures the model's ability to distinguish hateful from non-hateful comments within a specific identity subgroup, with low values indicating poor subgroup performance; **GMB-BPSN** (*Background Positive, Subgroup Negative*) assesses whether non-hateful comments mentioning the identity are incorrectly predicted as hateful compared to hateful comments in the background; while **GMB-BNSP** (*Background Negative, Subgroup Positive*) measures whether hateful comments mentioning the identity are confused with non-hateful background comments. Low scores in any metric indicate potential bias or unfair treatment of the corresponding identity subgroup. We define identity groups across four dimensions: *gender*, *race*, *politics* and *religion*. For gender, we include the terms "mulher", "mulheres", "homem", "homens", "feminista", "gay", "lésbica", "trans", "viado", and "sapatão" (translated as "woman", "women", "man", "men", "feminist", "gay", "lesbian", "trans", "faggot", and "dyke"). For race, the terms are "negro", "negra", "preto", "preta", "branco", "branca", "índio", "indígena", and "macaco" (translated as "black", "white", "indigenous", and "monkey", with gender variations). For politics, we consider "bolsonaro", "lula", "petista", "comunista", "fascista", "esquerdista", and "direitista" (translated as "Bolsonaro", "Lula", "Workers' Party supporter", "communist", "fascist", "leftist",

and "right-wing"). Finally, for religion, we include "cristão", "evangélico", "católico", "ateu", "macumbeiro", and "crente" (translated as "Christian", "evangelical", "Catholic", "atheist", "practitioner of Afro-Brazilian religions", and "religious believer").

6 Results and Discussion

Supervised Attention with Moral Rationales As shown in Table 1, SMRA consistently improves over the baseline models across both binary hate speech classification and multi-label moral sentiment classification. For binary classification, SMRA increases Accuracy (+0.0004) and Macro F1 (+0.0082) and boosts plausibility, with higher IoU F1 (+0.0743) and Token F1 (+0.0503), indicating better alignment with human-annotated moral rationales. Comprehensiveness slightly decreases (−0.0203), while Sufficiency improves (−0.0231), reflecting more concise yet still faithful explanations. In multi-label moral sentiment classification, SMRA also outperforms the baseline model, achieving higher Macro F1 (+0.015), AUROC (+0.002), and Token F1 (+0.241), highlighting superior rationale coverage and improved interpretability without sacrificing predictive performance. These results show that SMRA improves performance while providing more faithful and plausible explanations. Across both tasks, we observe that the choice of backbone model (BERTimbau vs. mBERT) has a larger impact on overall performance than the use of supervised attention, highlighting the importance of language-specific pretraining. Additionally, improvements in rationale alignment do not consistently translate into gains in fairness.

Supervised Attention with Moral vs. Hate Rationales We also compare our model with an additional binary hate speech classification baseline shown in Table 2. In contrast to our approach, SRA supervises attention using hate speech rationales rather than moral rationales. When comparing SMRA with SRA, we observe that SMRA achieves substantially higher plausibility, with gains in both IoU F1 (+0.1195) and Token F1 (+0.1508), indicating stronger alignment with human-annotated rationales. Although SRA attains higher Comprehensiveness, this comes at the cost of unstable faithfulness behavior, as evidenced by negative Sufficiency values[7] (−0.0360), suggesting inconsistencies in the causal relationship between selected rationales and model predictions. In contrast, SMRA yields more balanced and reliable explanations, maintaining positive Sufficiency (0.0426) while providing broader rationale coverage. Overall, these results indicate that SMRA not only improves alignment with human explanations but also produces more stable and interpretable rationale representations compared to prior supervised attention approaches such as SRA, highlighting the benefits of leveraging moral rationales for explanation learning. Finally, we also compared deep learning models on binary hate speech and multi-label moral classification, as shown in Table 3. For binary hate speech classification, CNN achieves the best overall performance across all metrics, while BiRNN with attention (SMRA) performs worst. In multi-label moral classification, results are mixed: SMRA attains the highest Accuracy, Bag-of-Words the best Macro F1, and CNN the highest AUROC, indicating no clear overall winner.

Ablation Study We also analyze the effect of explicit reasoning guidance and details of the Moral Foundations Theory. Hence, we omit those parts and prompt the models again (the exact prompt can be found in Appendix C). For hate speech, the performance increased 2% compared to providing moral foundations theory and explicit guidance for Gpt4o-mini but it decreased 6% for Llama70b. On the other hand, there is a huge performance drop (34 % for Gpt4o-mini and 36% for Llama70b) for moral value classification. This indicates that explicit guidance is more required for the moral sentiment classification which is more nuanced than binary hate speech classification.

Effect of English Translation Given that LLMs generally perform better in English, we translated the Portuguese prompts into English and applied the same prompting techniques. Table 4 shows that translation to English leads to lower performance for both models across both tasks, particularly for hate speech classification. This drop may be due to subtle meaning loss during translation, and moral values may not transfer seamlessly across cultures, potentially causing the ground-truth labels to shift.

[7] Negative sufficiency score indicates that the model becomes more confident when conditioned only on the selected rationale than on the full input, suggesting inconsistencies in the causal alignment between explanations and model predictions.

Model	Classification			Plausibility			Faithfulness		Fairness/Bias (AUC)		
	Acc.↑	Macro F1↑	AUROC↑	IOU F1↑	Token F1↑	AUPRC↑	Comp.↑	Suff.↓	Sub.↑	BPSN.↑	BNSP.↑
Hate classification											
mBERT-base	0.5343	0.5013	0.5640	0.8186	0.8832	0.8829	0.0113	0.1381	0.4653	0.5234	0.5341
mBERT-smra	0.5414	0.5037	0.5588	0.8366	0.8953	0.9163	0.0235	0.1376	0.4236	0.4529	0.5343
BERTimbau-base	0.9029	0.9028	**0.9651**	0.7612	0.8455	0.8307	**0.1733**	0.0657	**0.9378**	**0.9721**	**0.9365**
BERTimbau-smra	**0.9114**	**0.9110**	0.9648	**0.8355**	**0.8958**	**0.9273**	0.1530	**0.0426**	0.9299	0.9694	0.9197
Moral classification											
mBERT-base	0.2700	0.1966	0.5158	0.7270	0.9659	0.9712	0.6422	0.0848	0.4112	0.5836	0.5316
mBERT-smra	0.2114	0.1698	0.5539	0.7402	0.9659	0.9712	0.6301	0.0995	0.5362	0.5912	0.6123
BERTimbau-base	0.7200	0.7570	0.9250	0.2350	0.7250	0.9150	0.8452	0.0549	0.9139	**0.9604**	**0.9650**
BERTimbau-smra	**0.7230**	**0.7720**	**0.9270**	**0.2396**	**0.9660**	**0.9710**	**0.9062**	**0.0562**	**0.9212**	0.9503	0.9491

Table 1: Results across transformer-based models for hate speech and moral sentiment classification in Portuguese using mBERT and BERTimbau under two evaluation settings: binary classification (LABELS = [`Hate`, `Non-Hate`]) and multi-label classification (LABELS = ['NN', 'HN', 'FN', 'PN', 'AN', 'LN'], where NN = Non-Morality, HN = Harm/Care, FN = Fairness/Cheating, and PN = Purity/Degradation). The *base* variant refers to models without supervised attention, while *smra* denotes our supervised attention approach guided by moral rationales. Classification metrics capture predictive performance, plausibility metrics assess rationale quality compared with human annotations, faithfulness metrics evaluate the causal alignment between rationales and model predictions, bias metrics (Sub, BPSN, BNSP) quantify group-based fairness, and accuracy corresponds to *Exact Match Accuracy*.

	Plausibility ↑				Faithfulness	
Model	IoU F1 ↑	Token Prec ↑	Token Rec ↑	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	0.9094	0.0237
BERTimbau [SHAP]	0.6600	0.7489	0.7099	0.6831	0.8458	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
PTTS [LIME]	0.6057	0.7487	0.6978	0.6776	0.5654	0.0016
PTTS [SHAP]	0.7400	0.7177	0.8378	0.7362	0.6160	0.0083
SRA ($\alpha = 10$)	0.7160	**0.9350**	0.6680	**0.7450**	0.4540	**-0.0360**

Table 2: SRA (Eilertsen et al., 2025) results on the HateBRXplain benchmark (Salles et al., 2025a). Plausibility metrics evaluate alignment with human rationales, while faithfulness metrics assess the impact of rationales on model predictions.

Model	Accuracy↑	Macro F1↑	AUROC↑
Binary hate speech classification			
Bag-of-Words	0.8157	0.8157	0.8991
CNN	**0.8214**	**0.8180**	**0.9089**
BiRNN + MaxPool	0.8114	0.8114	0.9065
BiRNN + Attention (smra)	0.8014	0.8014	0.8903
Multi-label moral sentiment classification			
Bag-of-Words	0.1371	**0.1982**	0.5683
CNN	0.3263	0.1748	**0.7283**
BiRNN + MaxPool	0.4314	0.1190	0.6086
BiRNN + Attention (smra)	**0.4386**	0.1586	0.5855

Table 3: Comparison of classic deep learning models on binary hate speech classification and multi-label moral sentiment classification.

In addition, as shown in Table 1, the multilingual mBERT fine-tuned model performs poorly on our Portuguese dataset.

LLMs We compare the effect of different types of information in prompts on hate speech and moral value classification (Abdurahman et al., 2024). Table 4 reports the F1 scores for each task and model. Including the hate speech definition improves performance for both models on hate speech detection and for Llama70B on moral value classification. In contrast, providing data collection context tends to degrade performance. We also explore jointly predicting hate speech and moral labels, which benefits both tasks, with a stronger improvement for moral classification. Adding the hate speech definition to this multi-task prompting further enhances hate speech performance. Overall, the best prompts combine hate speech classification with moral rationales, showing that multi-hop moral explanations enhance both tasks and align the model more closely with human judgments. Lastly, we also compare the moral ratio-

nale predicted by an LLMs and humans to understand if LLMs can provide an plausible explanation for moral sentiment classification. We compute BERTScore (Zhang et al., 2019) and Jaccard (Jaccard, 1901) similarity between human and LLM rationales. Table 5 shows that both LLMs achieve reasonable BERTScore values, but exhibit low Jaccard similarity. We find that the average number of words identified by humans is around 5, whereas it is 1.93 for LLMs, indicating that LLMs tend to identify correct words but miss relevant portions of the rationale, resulting in lower recall. However, further studies are still required to better understand this phenomenon more effectively.

Model and Prompt	Hate Speech (F1)		Morality (F1)	
	pt	en	pt	en
Gpt4o-m hate only	0.864	0.739	–	–
Gpt4o-m hate w/ definition	0.893	0.797	–	–
Gpt4o-m hate w/ context	0.822	0.708	–	–
Gpt4o-m hate moral	0.870	0.745	**0.369**	0.378
Gpt4o-m hate moral w/ definition	**0.897**	0.790	0.366	**0.396**
Gpt4o-m hate moral w/ context	0.893	**0.805**	0.359	0.374
Gpt4o-m moral	–	–	0.351	0.334
Gpt4o-m moral w/ definition	–	–	0.346	0.358
Gpt4o-m moral w/ context	–	–	0.308	0.295
Llama70B hate only	0.822	0.740	–	–
Llama70B hate w/ definition	**0.901**	**0.871**	–	–
Llama70B hate w/ context	0.818	0.732	–	–
Llama70B hate moral	0.826	0.725	0.370	**0.360**
Llama70B hate moral w/ definition	0.893	0.835	0.379	0.393
Llama70B hate moral w/ context	0.807	0.712	**0.383**	0.341
Llama70B moral	–	–	0.270	0.265
Llama70B moral w/ definition	–	–	0.318	0.306
Llama70B moral w/ context	–	–	0.245	0.219

Table 4: F1 scores for hate speech detection and moral sentiment classification for datasets in Portuguese and their English translations. Results for GPT-4o-mini and LLaMA-70B under different prompting strategies.

Model	BertScore↑	Jaccard$_{sim}$↑
GPT4o-m hate moral w/ definition	0.76	0.11
Llama70b moral hate w/ definition	0.71	0.11

Table 5: Similarity between the moral rationale predicted by LLMs and human rationales.

Qualitative Analysis We extracted Instagram comments from HateBR-MoralXplain containing sarcasm or irony, representing challenging cases where offensive or non-offensive meaning depends on cultural knowledge. Table 6 shows examples illustrating these challenges. Comments 1 and 2 contain sarcasm and implicit hate that simple models may miss, while comments 3 and 4 rely on political context (e.g., "myth" refers to former president Bolsonaro, and "dumb" is preceded by negation). These cases highlight the need for models that can robustly distinguish subtle, context-dependent expressions while remaining culturally aware and transparent. Current models, which primarily rely on surface-level lexical features, are insufficient for capturing such nuanced hateful content. In total, 235 offensive comments[8] were extracted, either containing sarcasm/irony or highly dependent on cultural context. Moreover, adding prompts with moral framing ("hate moral w definition", see Appendix C) enabled Llama-70B to correctly label these nuanced cases, whereas prompts without moral framing ("hate w definition") misclassified them, suggesting that moral guidance can improve robust hate speech detection.

Instagram Comments	Labels
(1) Se jogar um carteira de trabalho não sobra um **Translation:** If you throw away a work permit, there won't be one left.	Hate
(2) Pepa só está dando tiro no pé **Translation:** Pepa is just shooting herself in the foot... lol	Hate
(3) O povo não é mais burro. Você não está ouvindo seus eleitores. **Translation:** The people aren't stupid anymore. You're not listening to your voters.	Non-Hate
(4) Chic esses filhos do mito. **Translation:** Chic, these kids of the 'Myth'.	Non-Hate

Table 6: Examples extracted from the HateBR dataset containing irony/sarcasm and culturally context-dependent expressions.

7 Conclusion

This paper introduces Supervised Moral Rationale Attention (SMRA), the first self-explaining framework for hate speech detection that aligns neural attention with expert-annotated moral rationales. To support this approach, we release HateBRMoralXplain, a Brazilian Portuguese benchmark dataset enriched with moral categories, their corresponding rationales, and socio-political metadata. Experimental results show that SMRA enhances explanation faithfulness and plausibility while slightly improving predictive performance. Fairness remains stable, suggesting that improvements in explanation quality do not introduce significant bias trade-offs. Overall, our findings highlight the importance of integrating moral and context-aware reasoning into hate speech detection, contributing toward more interpretable and culturally aware approaches, as well as advancing research in language model interpretability.

[8] https://github.com/franciellevargas/SMRA

Limitations

While SMRA improves interpretability and robustness in hate speech detection, several limitations remain. First, our approach relies on high-quality human-annotated moral rationales, which are costly and time-consuming to produce, potentially limiting scalability to new languages or domains. Second, SMRA has been evaluated primarily on Brazilian Portuguese social media data; further studies are needed to assess generalization across languages, platforms, and different cultural settings. Finally, integrating moral reasoning does not eliminate all biases in training data, and careful consideration of minority perspectives remains essential to avoid unintended harms. Finally, while we adopt the standard Moral Foundations Theory taxonomy, we do not distinguish between subdimensions of the Fairness foundation (e.g., equality vs. proportionality) (Atari et al., 2023), which have been shown to capture meaningful variation in moral reasoning across cultures and political contexts and may differentiate between language used to identify versus justify harm (Trager and Atari, 2025; Guo et al., 2024).

Ethics Statements

The dataset will be released under an open-source license CC BY-NC 4.0 (Attribution–NonCommercial), with full transparency regarding data collection and annotation procedures. In addition to the annotated labels, we also release detailed per-annotator information, enabling further analysis of annotation reliability and subjectivity. Particular care was taken to ensure fair and respectful treatment of annotators throughout the process.

We emphasize that this work is intended solely for research purposes; any downstream deployment of automated content moderation systems should be approached with additional safeguards, rigorous evaluation, and responsible governance. We truly believe that our research will make content moderation more fair and transparent.

Acknowledgements

Part of this research was conducted during the first author's Ph.D. at the University of São Paulo (USP) and further developed during her time as a visiting researcher at the University of Southern California (USC). The authors thank Dr. Morteza Dehghani and the Morality and Language Lab at the University of Southern California for their early guidance and support on this project. We also thank Isadora Salles for her assistance with the annotation of HateBRMoralXplain. Dr. Daryna Dementieva's work was supported by Prof Alexander Fraser's chair by the German Research Foundation (DFG; grant FR 2829/7-1). This research was partially supported by Google. Corresponding author: franciellealvargas@gmail.com.

References

Suhaib Abdurahman, Mohammad Atari, Farzan Karimi-Malekabadi, Mona J Xue, Jackson Trager, Peter S Park, Preni Golazizian, Ali Omrani, and Morteza Dehghani. 2024. Perils and opportunities in using large language models in psychological research. *PNAS nexus*, 3(7):pgae245.

Gavin Abercrombie, Dirk Hovy, and Vinodkumar Prabhakaran. 2023. Temporal and second language influence on intra-annotator agreement and stability in hate speech labelling. In *Proceedings of the 17th Linguistic Annotation Workshop (LAW-XVII)*, pages 96–103, Toronto, Canada. Association for Computational Linguistics.

Mohammad Atari, Jonathan Haidt, Jesse Graham, Sena Koleva, Sean T Stevens, and Morteza Dehghani. 2023. Morality beyond the weird: How the nomological network of morality varies across cultures. *Journal of Personality and Social Psychology*, 125(5):1157.

Giuseppe Attanasio, Debora Nozza, Dirk Hovy, and Elena Baralis. 2022. c. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1105–1119, Dublin, Ireland. Association for Computational Linguistics.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Held Online.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion Proceedings of The 2019 World Wide Web Conference*, WWW '19, page 491–500, New York, NY, USA. Association for Computing Machinery.

William J Brady, Killian L McLoughlin, Mark P Torres, Kara F Luo, Maria Gendron, and MJ Crockett. 2023. Overperception of moral outrage in online social networks inflates beliefs about intergroup hostility. *Nature human behaviour*, 7(6):917–927.

William J Brady, Julian A Wills, John T Jost, Joshua A Tucker, and Jay J Van Bavel. 2017. Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences*, 114(28):7313–7318.

Agostina Calabrese, Björn Ross, and Mirella Lapata. 2022. Explainable abuse detection as intent classification and slot filling. *Transactions of the Association for Computational Linguistics*, 10:1440–1454.

Amanda Cercas Curry, Giuseppe Attanasio, Debora Nozza, and Dirk Hovy. 2023. MilaNLP at SemEval-2023 task 10: Ensembling domain-adapted and regularized pretrained language models for robust sexism detection. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2067–2074, Toronto, Canada. Association for Computational Linguistics.

Aida Davani, Mark Díaz, Dylan Baker, and Vinodkumar Prabhakaran. 2024a. D3CODE: Disentangling disagreements in data across cultures on offensiveness detection and evaluation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18511–18526, Miami, Florida, USA. Association for Computational Linguistics.

Aida Davani, Mark Díaz, Dylan Baker, and Vinodkumar Prabhakaran. 2024b. Disentangling perceptions of offensiveness: Cultural and moral correlates. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 2007–2021, Rio de Janeiro, Brazil.

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, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial bias in hate speech and abusive language detection datasets. In *Proceedings of the 3rd Workshop on Abusive Language Online*, pages 25–35, Florence, Italy.

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.

Brage Eilertsen, Røskva Bjørgfinsdóttir, Francielle Vargas, and Ali Ramezani-Kebrya. 2025. Aligning attention with human rationales for self-explaining hate speech detection. In *Proceedings of the 40th AAAI Conference on Artificial Intelligence. Alignment Track.*, pages 1–15, Singapore, Singapore.

Jim AC Everett. 2013. The 12 item social and economic conservatism scale (secs). *PloS one*, 8(12):e82131.

Paula Fortuna, João Rocha da Silva, Juan Soler-Company, Leo Wanner, and Sérgio Nunes. 2019. 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.

Michele J Gelfand, Jana L Raver, Lisa Nishii, Lisa M Leslie, Janetta Lun, Beng Chong Lim, Lili Duan, Assaf Almaliach, Soon Ang, Jakobina Arnadottir, and 1 others. 2011. Differences between tight and loose cultures: A 33-nation study. *science*, 332(6033):1100–1104.

Jesse Graham, Jonathan Haidt, Sena Koleva, Matt Motyl, Ravi Iyer, Sean P Wojcik, and Peter H Ditto. 2013. Moral foundations theory: The pragmatic validity of moral pluralism. In *Advances in experimental social psychology*, volume 47, pages 55–130. Elsevier.

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.

Rongchen Guo, Isar Nejadgholi, Hillary Dawkins, Kathleen C Fraser, and Svetlana Kiritchenko. 2024. Adaptable moral stances of large language models on sexist content: Implications for society and gender discourse. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19548–19564.

Kobi Hackenburg, William J Brady, and Manos Tsakiris. 2023. Mapping moral language on us presidential primary campaigns reveals rhetorical networks of political division and unity. *PNAS nexus*, 2(6):pgad189.

Arnold K Ho, Jim Sidanius, Nour Kteily, Jennifer Sheehy-Skeffington, Felicia Pratto, Kristin E Henkel, Rob Foels, and Andrew L Stewart. 2015. The nature of social dominance orientation: Theorizing and measuring preferences for intergroup inequality using the new sdo7 scale. *Journal of personality and social psychology*, 109(6):1003.

Paul Jaccard. 1901. Étude comparative de la distribution florale dans une portion des alpes et des jura. *Bulletin de la Société Vaudoise des Sciences Naturelles*, 37:547–579.

Peter K Jonason and Gregory D Webster. 2010. The dirty dozen: a concise measure of the dark triad. *Psychological assessment*, 22(2):420.

Farzan Karimi-Malekabadi, Suhaib Abdurahman, Zhivar Sourati, Jackson Trager, and Morteza Dehghani. 2026. Theory trace card for llm socio-cognitive evaluation. Manuscript. Preprint.

Brendan Kennedy, Mohammad Atari, Aida Mostafazadeh Davani, Joe Hoover, Ali Omrani, Jesse Graham, and Morteza Dehghani. 2021. Moral concerns are differentially observable in language. *Cognition*, 212:104696.

Brendan Kennedy, Preni Golazizian, Jackson Trager, Mohammad Atari, Joe Hoover, Aida Mostafazadeh Davani, and Morteza Dehghani. 2023. The (moral) language of hate. *PNAS nexus*, 2(7):pgad210.

Brendan Kennedy, Xisen Jin, Aida Mostafazadeh Davani, Morteza Dehghani, and Xiang Ren. 2020. Contextualizing hate speech classifiers with post-hoc explanation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5435–5442, Online. Association for Computational Linguistics.

Eunsu Kim, Juyoung Suk, Philhoon Oh, Haneul Yoo, James Thorne, and Alice Oh. 2024. CLIcK: A benchmark dataset of cultural and linguistic intelligence in Korean. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3335–3346, Torino, Italia. ELRA and ICCL.

Jiyun Kim, Byounghan Lee, and Kyung-Ah Sohn. 2022. Why is it hate speech? masked rationale prediction for explainable hate speech detection. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6644–6655, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Nayeon Lee, Chani Jung, and Alice Oh. 2023. Hate speech classifiers are culturally insensitive. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 35–46, Dubrovnik, Croatia. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Marian Marchal, Merel Scholman, Frances Yung, and Vera Demberg. 2022. Establishing annotation quality in multi-label annotations. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3659–3668, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

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 35th AAAI Conference on Artificial Intelligence*, volume 35, pages 14867–14875, Held Online.

Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282.

Jari-Mikko Meriläinen. 2024. The role of gender in hate speech targeting politicians: Evidence from finnish twitter. *International Journal of Politics, Culture, and Society*, pages 1–27.

Anirudh Mittal, Pranav Jeevan P, Prerak Gandhi, Diptesh Kanojia, and Pushpak Bhattacharyya. 2021. "so you think you're funny?": Rating the humour quotient in standup comedy. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10073–10079, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ayushi Nirmal, Amrita Bhattacharjee, Paras Sheth, and Huan Liu. 2024. Towards interpretable hate speech detection using large language model-extracted rationales. In *Proceedings of the 8th Workshop on Online Abuse and Harms (WOAH 2024)*, pages 223–233, Mexico City, Mexico. Association for Computational Linguistics.

OpenAI. 2024. Gpt-4o mini: advancing cost-efficient intelligence.

Niels Bjørn Grund Petersen, Rasmus Tue Pedersen, and Mads Thau. 2025. Citizens' perceptions of online abuse directed at politicians: Evidence from a survey experiment. *European Journal of Political Research*, 64(2):790–809.

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.

Vinodkumar Prabhakaran, Aida Mostafazadeh Davani, and Mark Diaz. 2021. On releasing annotator-level labels and information in datasets. In *Proceedings of the Joint 15th Linguistic Annotation Workshop (LAW) and 3rd Designing Meaning Representations (DMR) Workshop*, pages 133–138, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Rezvaneh Rezapour, Ly Dinh, and Jana Diesner. 2021. Incorporating the measurement of moral foundations theory into analyzing stances on controversial topics. In *Proceedings of the 32nd ACM Conference on Hypertext and Social Media*, page 177–188, New York, NY, USA. Association for Computing Machinery.

Ludovic Rheault, Erica Rayment, and Andreea Musulan. 2019. Politicians in the line of fire: Incivility and the treatment of women on social media. *Research & Politics*, 6(1):2053168018816228.

Isadora Salles, Francielle Vargas, and Fabrício Benevenuto. 2025a. 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.

Isadora Salles, Francielle Vargas, and Fabrício Benevenuto. 2025b. 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.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5884–5906, Seattle, United States. Association for Computational Linguistics.

Kristina Bakkær Simonsen and Tobias Widmann. 2025. When do political parties moralize?: A cross-national study of the use of moral language in political communication on immigration. *British Journal of Political Science*, 55:e33.

Theodore M Singelis, Harry C Triandis, Dharm PS Bhawuk, and Michele J Gelfand. 1995. Horizontal and vertical dimensions of individualism and collectivism: A theoretical and measurement refinement. *Cross-cultural research*, 29(3):240–275.

Kirill Solovev and Nicolas Pröllochs. 2023. Moralized language predicts hate speech on social media. *PNAS nexus*, 2(1):pgac281.

Kirill Solovev and Nicolas Pröllochs. 2022. Moralized language predicts hate speech on social media. *PNAS Nexus*, 2(1):pgac281.

Christopher J Soto and Oliver P John. 2017. Short and extra-short forms of the big five inventory–2: The bfi-2-s and bfi-2-xs. *Journal of Research in Personality*, 68:69–81.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. BERTimbau: pretrained BERT models for Brazilian Portuguese. In *9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23 (to appear)*.

Manuel Tonneau, Diyi Liu, Samuel Fraiberger, Ralph Schroeder, Scott A. Hale, and Paul Röttger. 2024. From languages to geographies: Towards evaluating cultural bias in hate speech datasets. In *Proceedings of the 8th Workshop on Online Abuse and Harms (WOAH 2024)*, pages 283–311, Mexico City, Mexico. Association for Computational Linguistics.

Jackson Trager and Mohammad Atari. 2025. The immorality of too much money. *PNAS nexus*, 4(6):pgaf158.

Jackson Trager, Farzan Karimi-Malekabadi, Suhaib Abdurahman, and Morteza Dehghani. 2025a. Hate is justified when values are threatened: Evidence from ideologically threatening tweets and real world events. *arxiv*.

Jackson Trager, Francielle Vargas, Diego Alves, Matteo Guida, Mikel K Ngueajio, Ameeta Agrawal, Yalda Daryani, Farzan Karimi-Malekabadi, and Flor Miriam Plaza-del Arco. 2025b. Mftcxplain: A multilingual benchmark dataset for evaluating the moral reasoning of llms through multi-hop hate speech explanation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 15709–15740, Suzhou, China.

Jackson Trager, Alireza S Ziabari, Aida Mostafazadeh Davani, Preni Golazizian, Farzan Karimi-Malekabadi, Ali Omrani, Zhihe Li, Brendan Kennedy, Nils Karl Reimer, Melissa Reyes, and 1 others. 2022. The moral foundations reddit corpus. *arXiv preprint arXiv:2208.05545*.

Apoorva Upadhyaya, Marco Fisichella, and Wolfgang Nejdl. 2023. Toxicity, morality, and speech act guided stance detection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4464–4478, Singapore. Association for Computational Linguistics.

Francielle Vargas, Isabelle Carvalho, Fabiana Rodrigues de Góes, Thiago Pardo, and Fabrício Benevenuto. 2022a. 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.

Francielle Vargas, Isabelle Carvalho, Ali Hürriyetoğlu, Thiago Pardo, and Fabrício Benevenuto. 2023. Socially responsible hate speech detection: Can classifiers reflect social stereotypes? In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 1187–1196, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Francielle Vargas, Isabelle Carvalho, Thiago A. S. Pardo, and Fabrício Benevenuto. 2024. Context-aware and expert data resources for brazilian portuguese hate speech detection. *Natural Language Processing*, 31(2):435–456.

Francielle Vargas, Isabelle Carvalho, Fabiana Rodrigues de Góes, Thiago Pardo, and Fabrício Benevenuto. 2022b. 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, Fabiana Rodrigues de Góes, Isabelle Carvalho, Fabrício Benevenuto, and Thiago Pardo. 2021. Contextual-lexicon approach for abusive language detection. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1438–1447, Held Online. INCOMA Ltd.

Sze-Yuh Nina Wang and Yoel Inbar. 2021. Moral-language use by us political elites. *Psychological Science*, 32(1):14–26.

Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. 2019. Detection of Abusive Language: the Problem of Biased Datasets. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 602–608, Minneapolis, United States.

Sidney Wong. 2024. Sociocultural considerations in monitoring anti-LGBTQ+ content on social media. In *Proceedings of the 2nd Workshop on Cross-Cultural Considerations in NLP*, pages 84–97, Bangkok, Thailand. Association for Computational Linguistics.

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 17th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1415–1420, Minnesota, USA.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Li Zhou, Antonia Karamolegkou, Wenyu Chen, and Daniel Hershcovich. 2023. Cultural compass: Predicting transfer learning success in offensive language detection with cultural features. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12684–12702, Singapore. Association for Computational Linguistics.

A Appendix

A.1 HateBRMoralXplain Corpus

We introduce HateBRMoralXplain, the third version of the HateBR corpus (Vargas et al., 2022b, 2024), which extends both the original HateBR dataset and its explainable successor, HateBRXplain (Salles et al., 2025a). As in previous versions, the corpus comprises 7,000 Brazilian Portuguese Instagram comments collected from public political accounts and annotated by three expert annotators. The dataset includes 3,500 offensive and 3,500 non-offensive comments, spanning nine distinct hate speech targets.

A.1.1 Data Collection

HateBRMoralXplain builds upon the HateBR corpus (Vargas et al., 2022a), a large-scale expert-annotated dataset of Brazilian Portuguese Instagram comments collected from public political accounts. The political domain was selected due to its high prevalence of offensive and hateful language targeting social groups, as well as its relevance for studying ideological polarization, moral framing, and identity-based attacks. The original HateBR corpus consists of 7,000 comments collected from the comment sections of Brazilian politicians' Instagram posts published during the second half of 2019. Data were extracted using the Instagram API, following a controlled sampling strategy that balanced political ideology and gender of the account holders. Specifically, comments were collected from six public political accounts, evenly distributed across left- and right-leaning parties and including both male and female politicians. Only public posts and comments were included, and no personally identifiable information beyond publicly available metadata was retained. Prior to annotation, the data underwent a cleaning process to remove noise such as URLs, isolated mentions, and comments containing only emojis or laughter tokens. Hashtags and expressive markers were preserved, as they often convey pragmatic or affective meaning relevant to hate speech and moral expression. This procedure follows the data collection and preprocessing pipeline established in HateBR and HateBRXplain (Vargas et al., 2022a; Salles et al., 2025b). HateBRMoralXplain reuses the same 7,000 comments from HateBR and HateBRXplain, ensuring continuity across dataset versions while enabling direct comparison between hate speech labels, hate rationales, and the newly introduced moral annotations.

A.1.2 Annotation Process

The annotation process in HateBRMoralXplain extends the multi-layer expert annotation framework introduced in HateBR and HateBRXplain by

incorporating moral labels and moral rationales grounded in Moral Foundations Theory (MFT). The resulting annotation scheme enables the joint study of hate speech detection, explainability, and moral reasoning within a single unified corpus. Annotations were conducted by expert annotators with backgrounds in linguistics, hate speech research, and computational social science. One annotator participated in prior versions of the dataset, ensuring continuity and calibration across versions, while a second annotator was newly introduced and trained using the same guidelines. As in prior work, annotation followed an iterative process with guideline refinement and discussion to ensure conceptual alignment, following the annotator-in-the-loop paradigm described in MFTCXplain (Trager et al., 2025b). Tables 7 and 8 show an example of an Instagram comment extracted from our HateBR-MoralXplain dataset, including moral labels, moral rationales annotations, and metadata.

A.1.3 Hate Speech Definition

We adopt the same hate speech definition used in HateBR (Vargas et al., 2022b) and HateBRXplain (Salles et al., 2025b) to maintain consistency across dataset versions. Hate speech is defined as a form of offensive language that expresses violence, hostility, intolerance, prejudice, or discrimination. This definition captures both explicit and implicit forms of offensive language, acknowledging that hateful intent may be conveyed through indirect language, sarcasm, or moralized justifications rather than overt slurs.

A.1.4 Moral Categories

To classify expressions of moral sentiment, we rely on the framework provided by Moral Foundations Theory (MFT; Graham et al., 2013). MFT proposes that human moral judgment draws upon a set of core psychological systems that are widely observed across societies (Atari et al., 2023). These systems are typically organized along five foundational domains, each of which can be expressed through both moral virtues (positive adherence) and moral violations (negative transgressions). The framework has become a central resource for computational investigations of morality, including prior work in automated analysis of moral and hateful discourse (e.g., Trager et al., 2025b). In our study, we adopt these five foundational domains to guide our annotation of moral expression. Brief descriptions of each domain are provided below:

Care vs. Harm: (HN) Concerns the protection and welfare of others. Moral language in this domain emphasizes empathy, compassion, and safeguarding, while violations involve inflicting harm, dismissing suffering, or expressing cruelty.

Fairness vs. Cheating: (FN) Focuses on equitable treatment, rights, and reciprocity. Virtuous expressions include justice, honesty, and impartiality, whereas violations highlight exploitation, deceit, or rule-breaking.

Loyalty vs. Betrayal: (LN) Centers on group cohesion and social allegiances. Loyalty indicates devotion to one's group or allies, while betrayal refers to abandonment, disloyalty, or undermining the collective.

Authority vs. Subversion (AN): Involves respect for legitimate leadership, social order, and tradition. Endorsing authority reflects deference and duty to existing structures; subversion entails resistance, disrespect, or challenges to hierarchy.

Purity vs. Degradation (PN): Pertains to cultural norms surrounding cleanliness, modesty, and sanctity. Purity involves maintaining physical or moral integrity, whereas degradation is signaled through impurity, disgust, or perceived contamination.

A.1.5 Moral Rationales

For each assigned moral label, annotators identified moral rationales by highlighting the specific spans of text that expressed or justified the corresponding moral judgment. Similar to hate rationales, moral rationales were defined as the smallest set of words or phrases sufficient to convey the moral meaning underlying the label. Annotators were instructed to select distinct rationales for each moral label and to avoid reusing the same span across multiple labels whenever possible. This constraint encourages fine-grained differentiation between moral foundations and supports multi-hop explanation settings, where models must connect hate speech detection to moral reasoning through intermediate justifications.

A.1.6 Annotator's Profile

The updated version of the corpus was annotated by the same two specialists as in the previous release. Given the complexity of offensive language and hate speech detection in a highly politicized domain, both annotators have advanced academic training (PhD or MSc). To minimize bias and its potential impact on the results, the annotators were intentionally diversified across key demographic

ID	Instagram Comment	HS Label	MFT Label	Rationales	MFT Label	Rationales	MFT Label	Rationales
Portuguese	Celebrar a morte de milhões de judeus? Estúpido, nojento.	Hate	AN	Celebrar a morte de milhões de judeus?	PN	nojento	HN	estúpido
Translation	Celebrating the death of millions of Jews? Stupid, disgusting.	Hate	AN	Celebrating the death of millions of Jews?	PN	disgusting	HN	stupid

Table 7: Example of an Instagram comment from HateBRMoralXplain with moral labels and moral rationales.

Instagram Comment	Summary Post	Themes Post	Politician Gender	Politician Party	Link Post
O povo ODEIA vcs, bando de marginais. **Translation**: The people HATE you, you bunch of criminals.	The post announces an event in Brasília held in defense of national and popular sovereignty, shared by Fernando Haddad (Brazilian politician)	National sovereignty; Popular sovereignty; Political mobilization	male	right	https://www.instagram.com/p/B1_m4F5lO_d/

Table 8: Example of a Brazilian political Instagram comment extracted from our HateBRMoralXplain corpus, with metadata annotated for post themes, politician gender and party, and post link.

dimensions: both are women, one identifying as liberal and the other as conservative; one is white and the other Black; and they come from different Brazilian regions (North and Southeast).

Unlike prior versions of this corpus, the current release includes detailed pre-annotation survey data collected from annotators alongside standard demographic information. Annotators completed an extensive pre-annotation survey designed to capture psychological, cultural, and sociodemographic characteristics that may shape moral and hate-speech judgments(Trager et al., 2022). These measures included sexual orientation, household income, primary and secondary language(s), religious affiliation, and political ideology, assessed both via self-identification and using the Social and Economic Conservatism Scale (SECS; Everett, 2013). Annotators also completed validated instruments measuring moral values and related individual differences, including the Moral Foundations Questionnaire-2 (Atari et al., 2023), Big Five personality traits (Soto and John, 2017), Social Dominance Orientation (Ho et al., 2015), Collectivism/Individualism (Singelis et al., 1995), Cultural Tightness–Looseness (Gelfand et al., 2011), and Dark Triad traits (Jonason and Webster, 2010). Basic descriptive analyses indicated that our annotators, overall, leaned more liberal politically, came from higher-income households compared to national averages, and had higher levels of formal education. Following the recommendations of Prabhakaran et al. (2021) and Davani et al. (2024b), we include these annotator measures to increase transparency and utility for downstream researchers. We encourage future work to investigate how these annotator characteristics may influence labeling decisions, particularly given the subjectivity of moral and hate speech judgments (see Davani et al., 2024b; Salles et al., 2025b).

A.1.7 Annotation Evaluation

Inter-Annotator Agreement. Inter-annotator agreement was assessed using Cohen's Kappa (McHugh, 2012), including its weighted variant with quadratic weights (Marchal et al., 2022; Mittal et al., 2021), in order to measure the level of agreement between two independent annotators considering our multi-class moral labels dataset. Agreement was computed separately for each annotation dimension, considering only instances in which both annotators provided an explicit label; instances with null or missing annotations were excluded from the analysis. Since the labels follow an ordinal scale with three categories ($k = 3$), weighted Cohen's Kappa was employed to account for the degree of disagreement between categories. Quadratic weights were defined as $w_{ij} = 1 - \frac{(i-j)^2}{(k-1)^2}$, assigning lower penalties to disagreements between adjacent labels and higher penalties to disagreements between distant labels. The weighted Kappa was computed as $\kappa_w = 1 - \frac{\sum_{i,j} w_{ij} O_{ij}}{\sum_{i,j} w_{ij} E_{ij}}$, where O_{ij} denotes the observed proportion of label pairs (i, j) and E_{ij} denotes the expected proportion of agreement by

chance. All agreement scores were computed using the `cohen_kappa_score` implementation from `scikit-learn` with `weights="quadratic"`.

Class	Quadratic Weighted Kappa
Class A	0.811
Class B	0.671
Class C	0.612

Table 9: Inter-annotator agreement measured using Cohen's weighted Kappa with quadratic weights, computed separately for each annotation class. Instances with null annotations were excluded.

The values of quadratic weighted Cohen's Kappa indicate moderate to substantial inter-annotator agreement across the three evaluated classes. In particular, Class A exhibits substantial to almost perfect agreement ($\kappa = 0.811$), suggesting a high degree of shared understanding between annotators for this category. Classes B and C show moderate to substantial agreement ($\kappa = 0.671$ and $\kappa = 0.612$, respectively), reflecting the greater subjectivity associated with these categories. The use of quadratic weights is especially appropriate in this context, as it penalizes disagreements between semantically distant categories more strongly while assigning lower penalties to disagreements between adjacent labels. The observed agreement patterns suggest that most disagreements occur between neighboring categories rather than extreme opposites, which is expected in annotation tasks involving nuanced moral and discourse-related judgments. Overall, these agreement scores are consistent with prior work on subjective semantic annotation tasks and support the reliability of the annotation protocol and the suitability of the dataset for downstream modeling and analysis.

A.1.8 Descriptive Statistics

We provide additional descriptive statistics to further clarify the structure and annotation properties of the HateBRMoralXplain dataset. The dataset consists of 7,000 posts, with 3,500 labeled as hate speech and 3,500 as non-hate speech. Among the 3,500 hate posts, only one post did not correspond to any moral category. All remaining hate posts contain at least one moral annotation with associated rationales. Annotators were allowed to assign between one and three moral labels per post.

Considering only hate speech posts, annotators did not agree on any moral label in 306 out of 3,500 posts (8.74%). For posts with at least one agreed-upon label, the distribution of moral categories is shown in Table 10. As the task allows for multi-label annotations, percentages sum to more than 100%.

Moral Category	Count	Percentage (%)
Harm	1287	40.29
Cheating	1200	37.57
Degradation	1094	34.25
Subversion	291	9.11
Loyalty	96	3.01
Betrayal	72	2.25
Purity	40	1.25
Care	23	0.72
Fairness	18	0.56
Authority	7	0.22

Table 10: Distribution of moral categories for hate speech posts with at least one agreed label. Percentages are calculated relative to posts with at least one agreed label and may sum to more than 100% due to multi-label annotations.

Combination	Count
Cheating + Degradation	215
Cheating + Harm	206
Harm + Degradation	172

Table 11: Most frequent co-occurring moral label combinations (occurring more than 100 times).

We further examine co-occurrence patterns among moral labels. The most frequent multi-label combinations are presented in Table 11, with the full distribution of combinations provided in the supplementary materials.

Finally, we analyze the average length of rationales associated with each moral label. As shown in Table 12, rationale length varies across categories, with some labels (e.g., Fairness, Authority) associated with longer explanations on average.

A.2 Additional Annotated Examples and Cross-Cultural Considerations

To further illustrate the structure of the HateBR-MoralXplain dataset and the role of moral rationales in context, we provide additional annotated examples in the original Portuguese alongside their English translations. Each comment is annotated with up to three moral foundations and associated rationales highlighting the specific spans of text that justify each label.

These examples also highlight challenges in cross-cultural interpretation. As discussed in Section 6, moral reasoning and the linguistic expres-

Label	Mean Tokens
Fairness	10.79
Authority	10.43
Cheating	9.11
Harm	9.06
Care	8.68
Purity	8.20
Loyalty	8.15
Subversion	7.50
Betrayal	6.67
Degradation	5.80

Table 12: Average rationale length per moral label (measured in tokens separated by spaces).

sion of hate speech may not transfer cleanly across languages, as subtle pragmatic, cultural, and political cues can shift meaning. Providing examples in the original language allows readers to better understand how moral rationales are grounded in culturally specific expressions.

B Hate and Morality

A growing literature in moral psychology and computational social science shows that moral language is routinely used to moralize intergroup conflict, legitimize hostility, and structure political judgment, with distinct moral foundations systematically associated with ideological positions and group-based boundaries (Graham et al., 2013; Wang and Inbar, 2021; Hackenburg et al., 2023; Trager et al., 2025a). In the specific case of hateful and derogatory discourse, recent work argues that hate is often expressed through morally loaded justifications (e.g., purity, authority, loyalty) rather than only explicit slurs, and that moral-emotional dynamics such as moral outrage are central to how hostile content is produced and circulated online (Kennedy et al., 2023; Brady et al., 2017, 2023). At the same time, research on online political communication consistently finds that women in public life, including politicians, face disproportionate and gendered abuse, and that abusive and hateful content varies with political identities and partisan contexts (Rheault et al., 2019; Meriläinen, 2024; Petersen et al., 2025). These patterns are also shaped by culture and language: most morality-aware and explainable hate resources are English-centric, while moral and political language usage and its correlation vary across contexts, motivating the need for low-resource, culturally situated datasets (Salles et al., 2025a; Vargas et al., 2022b; Simonsen and Widmann, 2025). By incorporating metadata on the gender and political party of the targeted politician and focusing on Brazilian Portuguese as a low-resource language, HateBR-MoralXplain supports socio-demographic and linguistic analyses of moralized hate speech that remain underexplored in the current explainable hate speech literature.

C LLM Prompts

Hate Speech Definition: Hate Speech can result due to some of the followings:

- Having a term or expression with any pejorative connotation.
- Having a sequence of swear words.
- Having a sequence of at least two terms, or/and expressions with any pejorative connotation expressed explicitly or implicitly.

Context: The data was collected during the Bolsonaro government in 2019. We collected balanced data from left- and right-wing Brazilian politicians, ensuring balanced gender representation. The Jair Bolsonaro government began on 1 January 2019, after his election in late 2018 — Bolsonaro won on a wave of anti-establishment sentiment, capitalizing on widespread frustration with corruption scandals and economic stagnation. Early in his presidency, he pursued a conservative, pro-market agenda: notably, he enacted a major pension-reform in 2019 aiming to reduce social-security costs. His government was marked by sharp shifts in environmental and Indigenous-land policy — protections were scaled back, enforcement relaxed, and deforestation pressures increased, drawing both domestic and international criticism.

Hate only: Analyze the following text "{text}" for hate speech.
Provide your analysis in this exact format:
hate_label: [YES if the text contains hate speech, NO otherwise]
Provide ONLY the required output format with no additional text, explanations, or justifications.

Hate with definition: Analyze the following text "{text}" for hate speech.
{definition}
Provide your analysis in this exact format:
hate_label: [YES if the text contains hate speech, NO otherwise]

Comment (PT)	Comment (EN)	Moral	Rationale	Moral	Rationale
O povo ODEIA vcs, bando de marginais	People HATE you, bunch of criminals	FN	bando de marginais	HN	O povo ODEIA vcs
Boa Manu, pelos ataques dos Bozonarianos, vc está no caminho certo, da para notar que esses lixos humanos estão literalmente "se cagando"... miliciano, e sua família de bandidos!!!	Good Manu... we prefer Venezuelans and Argentinians over those crazy fascists and their criminal family!!!	PN	lixos humanos; se cagando	AN	maluco fascista
		FN	miliciano; família de bandidos		
Colonialista de merda	F***ing colonialist	AN	Colonialista de merda		

Table 13: Examples of annotated Instagram comments with moral labels and rationales in Portuguese (PT) and English (EN). Each comment may contain multiple moral labels with associated rationale spans.

Comment (PT)	Comment (EN)	Post Summary	Themes
O povo ODEIA vcs, bando de marginais	People HATE you, bunch of criminals	Announcement of a political event in defense of national and popular sovereignty.	National sovereignty; Political mobilization
Colonialista de merda	Shitty colonialist	Post showing a politician attending a WWII commemoration event.	Political content; Election-related content

Table 14: Examples with contextual metadata illustrating how moral reasoning is embedded in political discourse.

Provide ONLY the required output format with no additional text, explanations, or justifications.

Hate with context: Analyze the following text "{text}" for hate speech.
{context}
Provide your analysis in this exact format:
hate_label: [YES if the text contains hate speech, NO otherwise]
Provide ONLY the required output format with no additional text, explanations, or justifications.

Hate and moral: Analyze the following text "{text}" for hate speech and identify its underlying moral value dimensions:
The Moral Foundations Theory framework represents core ethical ad psychological concerns that come in paired positive vs negative expressions:

- care vs harm: Involves concern for the well-being of others, with virtues expressed through care, protection, or nurturance, and vices involving harm, cruelty, or indifference to suffering.
- fairness vs cheating: morals related to justice, rights, and reciprocity, with fairness indicating equity, rule-following, and cheating denoting exploitation, dishonesty, or manipulation.
- loyalty vs betrayal: morals related to group-based morality, where loyalty refers to solidarity, allegiance, and in-group defense, while betrayal signals disloyalty or abandonment of one's group.
- authority vs subversion: morals related to respect for tradition, and legitimate hierarchies, with authority indicating respect or deference to leadership or norms, and subversion indicating rebellion, disrespect, or disobedience.
- sanctity vs degradation: morals related to purity, contamination, with Purity is associated with cleanliness, modesty, or moral elevation, while degradation includes defilement, obscenity,or perceived corruption.

Provide your analysis in this exact format:
hate_label: [YES if the text contains hate speech, NO otherwise]
moral_value: [the single most prominent moral foundations from: care, harm, fairness, cheating, authority, subversion, sanctity, degradation, loyalty, betrayal. If no clear moral foundation applies,

write “None”]
explanation: [provide a brief evidence based justification, specifically highlighting the words or phrases that triggered your moral value classification. If none, write “None”]
Provide ONLY the required output format with no additional text, explanations, or justifications.

Hate and moral with context: Analyze the following text “{text}” for hate speech and identify its underlying moral value dimensions:
The Moral Foundations Theory framework represents core ethical ad psychological concerns that come in paired positive vs negative expressions:

- care vs harm: Involves concern for the well-being of others, with virtues expressed through care, protection, or nurturance, and vices involving harm, cruelty, or indifference to suffering.
- fairness vs cheating: morals related to justice, rights, and reciprocity, with fairness indicating equity, rule-following, and cheating denoting exploitation, dishonesty, or manipulation.
- loyalty vs betrayal: morals related to group-based morality, where loyalty refers to solidarity, allegiance, and in-group defense, while betrayal signals disloyalty or abandonment of one’s group.
- authority vs subversion: morals related to respect for tradition, and legitimate hierarchies, with authority indicating respect or deference to leadership or norms, and subversion indicating rebellion, disrespect, or disobedience.
- sanctity vs degradation: morals related to purity, contamination, with Purity is associated with cleanliness, modesty, or moral elevation, while degradation includes defilement, obscenity,or perceived corruption.

{context}
Provide your analysis in this exact format:
hate_label: [YES if the text contains hate speech, NO otherwise]
moral_value: [the single most prominent moral foundations from: care, harm, fairness, cheating, authority, subversion, sanctity, degradation, loyalty, betrayal. If no clear moral foundation applies, write “None”]
explanation: [provide a brief evidence based justification, specifically highlighting the words or phrases that triggered your moral value classification. If none, write “None”]
Provide ONLY the required output format with no additional text, explanations, or justifications.

Hate and moral with definition: Analyze the following text “{text}” for hate speech and identify its underlying moral value dimensions:
The Moral Foundations Theory framework represents core ethical ad psychological concerns that come in paired positive vs negative expressions:

- care vs harm: Involves concern for the well-being of others, with virtues expressed through care, protection, or nurturance, and vices involving harm, cruelty, or indifference to suffering.
- fairness vs cheating: morals related to justice, rights, and reciprocity, with fairness indicating equity, rule-following, and cheating denoting exploitation, dishonesty, or manipulation.
- loyalty vs betrayal: morals related to group-based morality, where loyalty refers to solidarity, allegiance, and in-group defense, while betrayal signals disloyalty or abandonment of one’s group.
- authority vs subversion: morals related to respect for tradition, and legitimate hierarchies, with authority indicating respect or deference to leadership or norms, and subversion indicating rebellion, disrespect, or disobedience.
- sanctity vs degradation: morals related to purity, contamination, with Purity is associated with cleanliness, modesty, or moral elevation, while degradation includes defilement, obscenity,or perceived corruption.

{definition}
Provide your analysis in this exact format:
hate_label: [YES if the text contains hate speech, NO otherwise]
moral_value: [the single most prominent moral foundations from: care, harm, fairness, cheating, authority, subversion, sanctity, degradation, loyalty, betrayal. If no clear moral foundation applies, write “None”]
explanation: [provide a brief evidence based justification, specifically highlighting the words or phrases that triggered your moral value classification. If none, write “None”]

Provide ONLY the required output format with no additional text, explanations, or justifications.

Moral only: Identify the underlying moral value dimensions in the following text "text". The Moral Foundations Theory framework represents core ethical ad psychological concerns that come in paired positive vs negative expressions:

- care vs harm: Involves concern for the well-being of others, with virtues expressed through care, protection, or nurturance, and vices involving harm, cruelty, or indifference to suffering.
- fairness vs cheating: morals related to justice, rights, and reciprocity, with fairness indicating equity, rule-following, and cheating denoting exploitation, dishonesty, or manipulation.
- loyalty vs betrayal: morals related to group-based morality, where loyalty refers to solidarity, allegiance, and in-group defense, while betrayal signals disloyalty or abandonment of one's group.
- authority vs subversion: morals related to respect for tradition, and legitimate hierarchies, with authority indicating respect or deference to leadership or norms, and subversion indicating rebellion, disrespect, or disobedience.
- sanctity vs degradation: morals related to purity, contamination, with Purity is associated with cleanliness, modesty, or moral elevation, while degradation includes defilement, obscenity,or perceived corruption.

Provide your analysis in this exact format:
moral_value: [the single most prominent moral foundations from: care, harm, fairness, cheating, authority, subversion, sanctity, degradation, loyalty, betrayal. If no clear moral foundation applies, write "None"]
explanation: [provide a brief evidence based justification, specifically highlighting the words or phrases that triggered your moral value classification. If none, write "None"]

Provide ONLY the required output format with no additional text, explanations, or justifications.

Moral with definition: Identify the underlying moral value dimensions in the following text "text". The Moral Foundations Theory framework represents core ethical ad psychological concerns that come in paired positive vs negative expressions:

- care vs harm: Involves concern for the well-being of others, with virtues expressed through care, protection, or nurturance, and vices involving harm, cruelty, or indifference to suffering.
- fairness vs cheating: morals related to justice, rights, and reciprocity, with fairness indicating equity, rule-following, and cheating denoting exploitation, dishonesty, or manipulation.
- loyalty vs betrayal: morals related to group-based morality, where loyalty refers to solidarity, allegiance, and in-group defense, while betrayal signals disloyalty or abandonment of one's group.
- authority vs subversion: morals related to respect for tradition, and legitimate hierarchies, with authority indicating respect or deference to leadership or norms, and subversion indicating rebellion, disrespect, or disobedience.
- sanctity vs degradation: morals related to purity, contamination, with Purity is associated with cleanliness, modesty, or moral elevation, while degradation includes defilement, obscenity,or perceived corruption.

{definition}
Provide your analysis in this exact format:
moral_value: [the single most prominent moral foundations from: care, harm, fairness, cheating, authority, subversion, sanctity, degradation, loyalty, betrayal. If no clear moral foundation applies, write "None"]
explanation: [provide a brief evidence based justification, specifically highlighting the words or phrases that triggered your moral value classification. If none, write "None"]

Provide ONLY the required output format with no additional text, explanations, or justifications.

Moral with context: Identify the underlying moral value dimensions in the following text "text". The Moral Foundations Theory framework represents core ethical ad psychological concerns that come in paired positive vs negative expressions:

- care vs harm: Involves concern for the well-being of others, with virtues expressed through care, protection, or nurturance, and vices involving harm, cruelty, or indifference to suffering.

- fairness vs cheating: morals related to justice, rights, and reciprocity, with fairness indicating equity, rule-following, and cheating denoting exploitation, dishonesty, or manipulation.
- loyalty vs betrayal: morals related to group-based morality, where loyalty refers to solidarity, allegiance, and in-group defense, while betrayal signals disloyalty or abandonment of one's group.
- authority vs subversion: morals related to respect for tradition, and legitimate hierarchies, with authority indicating respect or deference to leadership or norms, and subversion indicating rebellion, disrespect, or disobedience.
- sanctity vs degradation: morals related to purity, contamination, with Purity is associated with cleanliness, modesty, or moral elevation, while degradation includes defilement, obscenity,or perceived corruption.

{context}
Provide your analysis in this exact format: moral_value: [the single most prominent moral foundations from: care, harm, fairness, cheating, authority, subversion, sanctity, degradation, loyalty, betrayal. If no clear moral foundation applies, write "None"]
explanation: [provide a brief evidence based justification, specifically highlighting the words or phrases that triggered your moral value classification. If none, write "None"]

Provide ONLY the required output format with no additional text, explanations, or justifications.

Ablation: Analyze the following text "{text}" for hate speech and identify its moral value:
hate_label : [YES or NO]
moral_value: [care, harm, fairness, cheating, authority, subversion, sanctity, degradation, loyalty, betrayal, None]
explanation: [brief justification]

D Theory Trace Card

Theory Trace Card (Karimi-Malekabadi et al., 2026) for HateBR-MoralXplain Corpus

1. Theory

- **Framework:** Hate speech (Fortuna et al., 2019; Zampieri et al., 2019; Vargas et al., 2022a); Moral Foundations Theory (MFT),(Graham et al., 2013); Span-based rationales (Zaidan et al., 2007)
- **Core components:**
  - **Hate Speech:** Normative judgment of hostility or discrimination toward protected groups; theoretically includes multiple sub-dimensions (e.g., intent, target, implicitness, severity)
  - **Moral Foundations:** Five foundations—Care/Harm, Fairness/Cheating, Loyalty/Betrayal, Authority/Subversion, Purity/Degradation—each represented via a virtue/violation (vice) polarity.
  - **Rationales:** Minimal, annotator-selected spans of text that provide evidential justification for assigned hate and/or moral labels, reflecting both salience and sufficiency.

2. Components Exercised

- Recognition of hate speech under a collapsed binary operationalization.
- Classification of moral content into virtue/violation categories across five moral foundations plus non-morality.
- Identification of morally and normatively salient textual evidence via span-based rationales.

3. Task Operationalization

- **Task:** Given an Instagram comment responding to a Brazilian political post, models are required to (i) predict whether the comment constitutes hate speech, (ii) predict which moral foundations are present (if any), and (iii) identify textual spans that justify the assigned hate and/or moral labels.
- **Key specs:** Comments are short-form Brazilian Portuguese Instagram texts collected during specific political crises. Moral annotation allows assignment of 1–3 moral labels per comment, ordered by annotator-perceived salience, spanning 11 categories (five foundations × virtue/violation, plus Non-Morality). Multiple theoretically distinct dimensions of hate speech are collapsed into a single binary label.
- **Scoring criterion:** Model outputs are evaluated via agreement with expert human annotators. Classification performance is assessed using standard metrics (e.g., Macro F1), while rationale quality is evaluated using overlap-based plausibility and faithfulness metrics comparing predicted spans to expert-annotated rationales.

4. Inference and Limitations

- **Inference:** Performance on this benchmark is treated as evidence that a model can align hate speech and moral category predictions with expert normative judgments in short-form, politically contextualized Brazilian Portuguese social media comments, and can ground those predictions in human-interpretable textual evidence.
- **Limitations:** *Data limitations* include restriction to Brazilian Portuguese, Instagram comments responding to political actors, collection during Bolsonaro-era political crises, and a limited number of expert annotators. *Theoretical limitations* include reliance on Moral Foundations Theory rather than alternative moral frameworks (e.g., dyadic morality, utilitarian or virtue-ethics accounts), collapse of multiple hate speech sub-dimensions into a single label, absence of author intent modeling, and treatment of moral reasoning as classification rather than deliberation or tradeoff-based judgment.