The Fortieth AAAI Conference on Artificial Intelligence (AAAI-26)

Aligning Attention with Human Rationales for Self-Explaining Hate Speech Detection

Brage Eilertsen[1], **Røskva Bjørgfinsdóttir**[1], **Francielle Vargas**[2], **Ali Ramezani-Kebrya**[1,3,4]
[1]University of Oslo
[2]University of São Paulo
[3]Integreat - Norwegian Centre for knowledge-driven machine learning
[4]TRUST – The Norwegian Centre for Trustworthy AI
brageei@uio.no, roskvatb@uio.no, ali@uio.no

Abstract

The opaque nature of deep learning models presents significant challenges for the ethical deployment of hate speech detection systems. To address this limitation, we introduce Supervised Rational Attention (SRA), a framework that explicitly aligns model attention with human rationales, improving both interpretability and fairness in hate speech classification. SRA integrates a supervised attention mechanism into transformer-based classifiers, optimizing a joint objective that combines standard classification loss with an alignment loss term that minimizes the discrepancy between attention weights and human-annotated rationales. We evaluated SRA on hate speech benchmarks in English (HateXplain) and Portuguese (HateBRXplain) with rationale annotations. Empirically, SRA achieves 2.4× better explainability compared to current baselines, and produces token-level explanations that are more faithful and human-aligned. In terms of fairness, SRA achieves competitive fairness across all measures, with second-best performance in detecting toxic posts targeting identity groups, while maintaining comparable results on other metrics. These findings demonstrate that incorporating human rationales into attention mechanisms can enhance interpretability and faithfulness without compromising fairness.

Introduction

The proliferation of social media platforms has necessitated the development of automated hate speech detection systems. These systems operate at large scale, with platforms such as Meta processing over seven million hate speech appeals monthly (Gorwa, Binns, and Katzenbach 2020; Meta Oversight Board 2025). These automated classifiers pose the risk of reinforcing societal biases and marginalizing vulnerable groups unless carefully designed and evaluated for fairness and representational harms (Davidson, Bhattacharya, and Weber 2019; Davani et al. 2023; Ungless et al. 2025; Attanasio et al. 2022; Vargas et al. 2023). In this work, we define offensive language as confrontational, rude, or aggressive content (Davidson et al. 2017; Zampieri et al. 2019), while hate speech specifically targets individuals based on their social identities (Fortuna and Nunes 2018). Sentiments can be expressed explicitly or implicitly (Poletto et al. 2021; Vargas et al. 2024). Existing automated classifiers exhibit systematic biases such as oversensitivity to identity terms, reinforcement of unfair associations of ethnic stereotypes, favouring European American names over African American names and associating negative sentiments with people with disabilities (Caliskan, Bryson, and Narayanan 2017; Hutchinson et al. 2020; Dixon et al. 2018).

While deep learning (DL) approaches achieve high performance in hate speech detection, they present two critical limitations. First, their black-box nature prevents understanding of whether decisions rely on problematic shortcuts or genuine indicators of harm (Gongane, Munot, and Anuse 2024). Second, systematic biases in hate speech detection can marginalize the communities they aim to protect (May et al. 2019). Developing transparent and responsible AI systems has essential important due to both ethical considerations and regulatory requirements (Brunk, Mattern, and Riehle 2019; European Parliament and Council 2016). Given the potential consequences of misidentifying or overlooking harmful content, developing interpretable hate speech detection systems represents a key challenge in Natural Language Processing (NLP) research (Mathew et al. 2021; Calabrese et al. 2024; Salles, Vargas, and Benevenuto 2025). Explainable AI methods that clarify the rationales (supporting evidence) behind predictions are the key components for the ethical deployment of such systems (Balkir et al. 2022).

Prior work has explored aligning model attention with human rationales, text snippets that guide and justify labeling decisions (DeYoung et al. 2020; Jain et al. 2020). Rationale-based learning improves interpretability across NLP tasks including Natural Language Inference, Question Answering, Information Retrieval, and Sentiment Analysis (DeYoung et al. 2020; Bastings, Aziz, and Titov 2019; Lei, Barzilay, and Jaakkola 2016; Zhang, Marshall, and Wallace 2016; Jiang et al. 2021; Lehman et al. 2019; Jørgensen et al. 2022). In these applications, attention-based models (Vaswani et al. 2017) have been explored since they offer a certain degree of interpretability. However, these explanations are often superficial and lack insights into why a model might consider specific features or tokens as relevant. For example, Strout, Zhang, and Mooney (2019) found that explanations generated using supervised attention are judged superior compared to explanations generated using normal unsuper-

---

vised attention. Bao et al. (2018) also showed that mapping human-annotated rationales into a continuous space significantly improves over the baselines and reduces error by over 15% on average across benchmark datasets. These studies did not, however, address fairness implications or evaluate token-level explanations for sensitive applications.

We propose Supervised Rational Attention (SRA), a method that enhances the explainability and fairness of hate speech detection systems. We argue that by using rationales that are explicitly labeled by domain experts, the system will generate more transparent and meaningful explanations. This can build trust with end-users and enable better decision-making in sensitive applications, such as content moderation.

Our contributions are summarized as follows:

- We introduce SRA, a framework to align model attention with human-annotated rationales.
- We evaluate SRA on English (HateXplain) and Portuguese (HateBRXplain) benchmarks (Mathew et al. 2021; Salles, Vargas, and Benevenuto 2025) and find that SRA improves explainability metrics (IoU F1 and Token F1) and achieves competitive fairness with second-best (GMB-BNSP) performance among all methods, with minimal impact on predictive performance.
- We provide publicly available code, datasets and models including rationale mask construction and attention supervision, to facilitate future research on trustworthy NLP and DL.

Related Work

Explainable hate speech detection methods are commonly categorized as either *self-explaining* or *post-hoc explaining* (Guidotti et al. 2018). Self-explaining methods integrate interpretability directly into the model architecture, while post-hoc methods generate explanations after training using the model's input-output behavior (Balkir et al. 2022). Our work focuses on self-explaining methods that leverage human rationales during training.

HateXplain Baseline. Mathew et al. (2021) proposed the foundational approach for explainable hate speech detection with human rationales. They transform human-annotated text spans into ground truth attention vectors by averaging across annotators and applying softmax with temperature. During training, attention-based models (BiRNN-Attention and BERT) minimize cross-entropy loss between predicted attention weights and ground truth attention vectors, encouraging alignment with human rationales. Models are evaluated on both classification metrics (accuracy, macro-F1, AUROC) and explainability metrics following the ERASER benchmark (DeYoung et al. 2020), including plausibility (IoU-F1, Token-F1, AUPRC) and faithfulness (comprehensiveness, sufficiency).

Attention Supervision Methods. Kim, Lee, and Sohn (2022) introduced Masked Rationale Prediction (MRP), which masks portions of rationale embeddings during an intermediate token-level classification task. The model learns to predict masked rationale labels before fine-tuning on hate speech classification. Clarke et al. (2023) developed Rule by Example, a contrastive learning framework that grounds hate-speech predictions in logical rules rather than token-level rationales.

Limitations of Prior Work. Existing attention supervision methods have improved explanation quality (Strout, Zhang, and Mooney 2019; Bao et al. 2018), with significant error reductions on benchmark datasets. However, these studies have not adequately addressed fairness implications or evaluated token-level explanations specifically for sensitive applications like hate speech detection, where both interpretability and equitable treatment across demographic groups are critical. Additional related work on post-hoc and self-explaining methods is provided in Supplementary.

Supervised Rational Attention (SRA)

The SRA framework aligns model attention with human rationales for hate speech detection. SRA enhances standard transformer-based text classifiers by explicitly aligning model attention with human-annotated rationales, leading to explanations that are plausible and faithful (DeYoung et al. 2020).

Problem Setting. Let $x = (w_1, \ldots, w_L)$ denote a tokenized input sequence, and $y \in \{0, 1, \ldots, C-1\}$ its class label, where C is the number of classes (e.g., $C = 3$ for multiclass hate speech classification). In the HateXplain benchmark (Mathew et al. 2021), labels are defined as normal $= 0$, offensive $= 1$, and hate speech $= 2$. For a subset of training examples, we are additionally given a *rationale mask* $r = (r_1, \ldots, r_L)$, where $r_i \in \{0, 1\}$ indicates whether token w_i is part of the human rationale for the label y.

Model Architecture. We use a pre-trained transformer encoder f_θ (e.g., BERT or BERTimbau) that computes contextual representations $\mathbf{h}_i$ for each input token w_i. A classification head predicts the class label:

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

where $\mathbf{h}_{\text{[CLS]}} \in \mathbb{R}^d$ is the representation of the [CLS] token, $\mathbf{W}_c \in \mathbb{R}^{1 \times d}$ and $\mathbf{b}_c \in \mathbb{R}$ are the learned weight matrix and bias term for class c, and C is the number of classes ($C = 3$ for HateXplain, $C = 2$ for HateBRXplain).

Attention Extraction. Let $A^{(l,h)} \in [0, 1]^{L \times L}$ denote the attention matrix at layer l and head h of encoder f_θ. To obtain model explanations, we first extract the attention from the [CLS] token to all input tokens:

$$\mathbf{a}^{(l,h)} = (a_1^{(l,h)}, \ldots, a_L^{(l,h)}) = A^{(l,h)}_{\text{[CLS]},:} \in \mathbb{R}^L$$

To select appropriate attention weights for supervision, we have conducted comprehensive ablation studies across layers and heads (see Supplementary for detailed results). Our experiments show that SRA achieves consistent performance across layers 6-11 and all attention heads, with minimal variation in both classification performance and explainability metrics. This robustness indicates that our method is not dependent on specific architectural choices. Based on these ablation studies, we use layer 8 and head 7 for our main experiments, as this configuration provides a good balance between classification and explainability performance. To reduce computational costs, we also evaluated a variant of SRA by taking an average of attention weights over all heads within a layer. Our experiments for this variant show consistent results with details provided in Supplementary.

Attention Alignment Loss (AAL). Given a sample with rationale mask r, we encourage the model to attend to tokens deemed important by human annotators by minimizing the mean squared error (MSE) between normalized attention $\mathbf{a}$ and r:

$$\ell_{\mathrm{AAL}}(\mathbf{a}, r) = \frac{1}{\sum_{i=1}^{L} m_i} \sum_{i=1}^{L} m_i \left(\frac{a_i}{\sum_{j=1}^{L} m_j a_j + \epsilon} - r_i \right)^2$$

where $m_i \in \{0, 1\}$ is the padding mask (1 for valid tokens, 0 for padding) and $\epsilon = 10^{-10}$ for numerical stability.

Overall Training Objective. We combine the standard cross-entropy loss for classification and the supervised attention alignment loss:

$$\ell_{\mathrm{total}} = \ell_{\mathrm{CE}}(\hat{y}, y) + \alpha \mathbf{1}[y > 0] \mathbf{1} \left[\sum_{i=1}^{L} r_i > 0 \right] \ell_{\mathrm{AAL}}(\mathbf{a}, r)$$

where α is a hyperparameter controlling the strength of attention supervision, balancing attention alignment and classification, $\mathbf{1}[y > 0]$ is an indicator function that equals 1 when the label is offensive or hate speech (i.e., $y \in \{1, 2\}$), and $\mathbf{1}[\sum_{i=1}^{L} r_i > 0]$ ensures rationales exist. The supervised attention loss is only applied for training examples labeled as offensive or hate speech with human-annotated rationales.

Rationale Extraction (Dataset-dependent).

- For datasets with *token-level rationales*, r is obtained by majority voting **across annotators who provided rationales for each example**. Specifically, a token is included in the rationale mask if it is selected by at least 50% of the available annotators (i.e., those who supplied rationales), corresponding to a threshold of 0.5.
- For datasets with *free-text rationale spans*, r is constructed by mapping character-level rationale spans to token indices using offset mappings from the tokenizer.

All rationale masks are aligned with the tokenization of $x = (w_1, \ldots, w_L)$ and truncated or padded to the model's maximum input length.

Inference and Explainability. At inference, the same attention mechanism provides explanation for each prediction: the tokens with highest attention from [CLS] correspond to the model's "rationale" for its decision. This allows evaluation of faithfulness and plausibility metrics (DeYoung et al. 2020) with respect to human rationales.

Experimental Setup

Implementation Overview. We implement our framework in PyTorch with Hugging Face Transformers. Full code, including rationale mask construction, will be released upon publication.

Models. We use **BERT-base-uncased** (Devlin et al. 2019) for English and **BERTimbau-base** (Souza, Nogueira, and Lotufo 2020) for Portuguese, extracting attention from layer 8 and head 7 based on preliminary validation.

Datasets. Models are trained on **HateXplain** (Mathew et al. 2021) (English, 20,148 samples with three classes; strong inter-annotator reliability) and **HateBRXplain** (Vargas et al. 2024) (Portuguese, 7,000 samples with two classes). We created 80/10/10 stratified train/validation/test splits with no data overlap between splits, using fixed random seeds for reproducibility, and converted rationales to binary masks via majority vote across annotators.

Training. Hyperparameters are tuned on validation macro F1. We use AdamW (English: 2×10^{-5}, batch 16, max len 128; Portuguese: 1×10^{-5}, batch 8, max len 512), 5 epochs, and select the checkpoint with best validation F1. Unless otherwise specified, we set $\alpha = 10.0$ in the overall objective. All results are averaged over 5 random seeds.

Hardware. Our experiments have been conducted on a cluster node with one A100 GPU.

Results and Discussion

Our results show that explicit rationale supervision can improve explainability (2.4× better IoU F1) while maintaining classification performance and competitive fairness across demographic groups.

We compare SRA with the following baselines from (Mathew et al. 2021) on HateXplain (English): (1) **CNN-GRU** (Zhang, Robinson, and Tepper 2018), which combines CNN filters with GRU layers; (2) **BiRNN** (Schuster and Paliwal 1997), a bidirectional RNN model; (3) **BiRNN-Attention** (Liu and Lane 2016), which adds an attention mechanism to BiRNN; and (4) **BERT** (Devlin et al. 2019), the transformer-based model. For models with attention mechanisms (BiRNN-Attention and BERT), we evaluate both unsupervised attention variants (denoted with suffix "-Attn") and variants trained with attention supervision (denoted with suffix "-HateXplain"). Additionally, we compare against BERT-MRP (Kim, Lee, and Sohn 2022), a masked rationale prediction approach, and post-hoc explanation methods [LIME] (Ribeiro, Singh, and Guestrin 2016) on the baseline models. The bracketed labels [Attn] and [LIME] indicate the explanation method used: attention-based or LIME post-hoc explanations, respectively. Table 1 uses these labels for clarity.

We evaluate SRA on both English and the Portuguese benchmarks, showing consistent improvements in attention alignment, explainability metrics, and bias reduction across languages.

Results on HateXplain (English) in Table 1

Classification Performance. SRA achieves a macro F1 score of 0.682 and accuracy of 0.696, marginally below the best-performing BERT-MRP baseline (F1: 0.699, Acc: 0.704). SRA achieves AUROC of 0.855 that is competitive with rationale-based baselines, including BERT-MRP

Model [Explanation Method]	**Classification**			**Explainability**			**Bias (AUC)**			**Faithfulness**	
	Acc.↑	Macro F1↑	AUROC↑	IoU F1↑	Token F1↑	AUPRC↑	GMB-Sub.↑	GMB-BPSN↑	GMB-BNSP↑	Comp.↑	Suff.↓
CNN-GRU [LIME]	0.627	0.606	0.793	0.167	0.385	0.648	0.654	0.623	0.659	0.316	**-0.082**
BiRNN [LIME]	0.595	0.575	0.767	0.162	0.361	0.605	0.660	0.640	0.671	0.421	-0.051
BiRNN-Attn [Attn]	0.621	0.614	0.795	0.167	0.369	0.643	0.653	0.662	0.668	0.278	-0.001
BiRNN-Attn [LIME]	0.621	0.614	0.795	0.162	0.386	0.650	0.653	0.662	0.668	0.308	-0.075
BiRNN-HateXplain [Attn]	0.629	0.629	0.805	0.222	0.506	**0.841**	0.691	0.691	0.674	0.281	0.039
BiRNN-HateXplain [LIME]	0.629	0.629	0.805	0.174	0.407	0.685	0.691	0.691	0.674	0.343	-0.075
BERT [Attn]	0.690	0.674	0.843	0.130	0.497	0.778	0.762	0.709	0.757	0.447	0.057
BERT [LIME]	0.690	0.674	0.843	0.118	0.468	0.747	0.762	0.709	0.757	0.436	0.008
BERT-HateXplain [Attn]	0.698	0.687	0.851	0.120	0.411	0.626	0.807	0.745	0.763	0.424	0.160
BERT-HateXplain [LIME]	0.698	0.687	0.851	0.112	0.452	0.722	0.807	0.745	0.763	**0.500**	0.004
BERT-MRP [Attn]	**0.704**	**0.699**	**0.862**	0.141	0.504	0.745	**0.815**	**0.748**	**0.854**	0.479	0.067
SRA, Ours ($\alpha = 10$)	0.696 (±0.007)	0.682 (±0.010)	0.855 (±0.002)	**0.539** (±0.005)	**0.651** (±0.002)	0.753 (±0.001)	0.714 (±0.002)	0.718 (±0.003)	0.835 (±0.012)	0.417 (±0.019)	-0.013 (±0.012)

Table 1: Performance comparison of hate speech detection models on the HateXplain test set. We evaluate classification performance (Accuracy, Macro F1, AUROC), explainability metrics (IoU F1, Token F1, AUPRC), fairness metrics (GMB-Subgroup, GMB-BPSN, GMB-BNSP AUCs), and faithfulness metrics (Comprehensiveness, Sufficiency). Higher values are better for all metrics except Sufficiency (lower is better). Models with [LIME] use post-hoc explanations, while [Attn] indicates attention-based explanations. Our proposed SRA method achieves the best IoU F1 and Token F1 explainability scores while maintaining competitive classification performance, fairness, and faithfulness. Best results are in **bold**, second-best are underlined. SRA results are averaged across 5 random seeds with standard deviations shown in parentheses.

(0.862) and BERT-HateXplain (0.851). This small performance difference (less than 2% in F1) suggests that incorporating human rationale supervision can significantly improve interpretability with little impact on classification performance.

Explainability Improvements. SRA shows improvements in explainability metrics compared to the baselines. **IoU F1**, which measures the Intersection-over-Union F1 score between model attention and human rationales (with matches defined by an overlap exceeding 0.5 (DeYoung et al. 2020)), improves to 0.539—representing a 4.5× improvement over BERT-HateXplain's supervised attention and a 2.4× gain over the best baseline (BiRNN-HateXplain). **Token F1**, which evaluates rationale alignment as token-level binary classification based on precision and recall for individual token matches (DeYoung et al. 2020), reaches 0.651, outperforming the best baseline by 29%. For SRA, token-level precision and recall are 0.937 and 0.579, respectively, indicating that when the model highlights tokens, they are likely to align with human rationales. The **AUPRC** metric evaluates soft token scoring by sweeping thresholds over attention weights and confirms SRA's strong ranking ability for rationale tokens.

These improvements indicate that SRA learns to focus on the same textual evidence that human annotators identify as indicators of hate speech.

Fairness Metrics. We evaluate fairness using three Generalized Mean Bias (GMB) metrics (Borkan et al. 2019). SRA achieves a **GMB-Subgroup AUC** of 0.714, which measures the model's ability to distinguish toxic from normal posts that mention identity groups. While this is lower than BERT-MRP's 0.815 and BERT-HateXplain's 0.807, it remains comparable. The **GMB-BPSN AUC** of 0.718 (compared to 0.748 for BERT-MRP and 0.745 for BERT-HateXplain) evaluates false-positive rates by measuring performance on normal posts mentioning identities versus toxic posts without identity terms. SRA achieves a **GMB-BNSP AUC** of 0.835, the second-best result after BERT-MRP's 0.854, showing improved performance at avoiding false negatives when toxic posts mention identity groups compared to BERT-HateXplain's 0.763. These results suggest a trade-off among fairness metrics. SRA shows slightly reduced performance on subgroup classification and false positive measures but significantly improves false negative detection for posts targeting identity groups compared to non-MRP baselines. This indicates that supervised rationale alignment helps the model better identify genuinely harmful content targeting protected groups.

Faithfulness Analysis. Faithfulness metrics evaluate whether model explanations correspond to features actually used for predictions (DeYoung et al. 2020). SRA achieves a **comprehensiveness** score of 0.417, which measures the performance drop when removing high-attention (rationale) tokens. Higher values indicate that rationales were influential in predictions. While SRA is outperformed by baselines using post-hoc explanation, e.g., BERT-HateXplain [LIME], it is competitive or outperforms other attention-based baselines. This suggests that the model uses both highlighted rationales and wider context. The **sufficiency** score compares the model's prediction confidence on the full input text versus its confidence when using only the rationale tokens. The negative values, which are common among attention-based methods including SRA, indicate that the model becomes more confident for prediction of the class given only rationales since it is less distracted by the context.

Ablation and Robustness. To evaluate the robustness of our method to the choice of rationale alignment hyperparameter, we conduct an ablation study by varying α from 0 (baseline without supervision) to 100. As illustrated in Figure 1, increasing α improves explainability metrics: IoU F1 improves from 0.019 (baseline) to 0.572 ($\alpha = 100$),

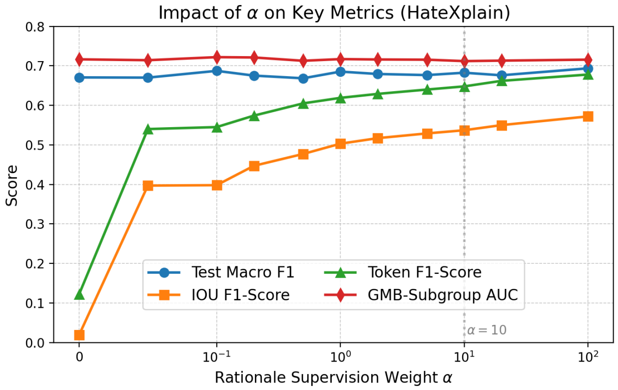

Figure 1: Impact of the rationale alignment hyperparameter α on test performance, explainability, and fairness metrics for HateXplain dataset. Increasing α yields improvements in rationale alignment (IoU F1, Token F1) while maintaining stable classification performance (Test Macro F1) and fairness (GMB-Subgroup AUC). The vertical line indicates our chosen operating point ($\alpha = 10$).

while Token F1 increases from 0.122 to 0.678. These gains come with limited impact on classification performance. Test Macro F1 remains relatively stable across all α values. Similarly, fairness remains relatively stable, with GMB-Subgroup AUC ranging from 0.7112-0.7220. We use $\alpha = 10$ as our primary configuration, achieving good explainability (IoU F1: 0.537, Token F1: 0.648) while maintaining comparable performance (Test F1: 0.683) and fairness (GMB-Sub: 0.7120). The range of effective α values suggests that SRA can substantially improve interpretability with minimal impact on performance and biases.

Results on HateBRXplain (Portuguese)

For HateBRXplain, we compare against the following baselines from (Salles, Vargas, and Benevenuto 2025): (1) mBERT (Devlin et al. 2019), the multilingual BERT model; (2) BERTimbau (Souza, Nogueira, and Lotufo 2020), a Portuguese-specific BERT model; (3) DistilBERTTim- bau (Junior 2024), a distilled version of BERTimbau; and (4) PTT5 (Carmo et al. 2020), a Portuguese T5 model. Since these baselines were evaluated using post-hoc explanation methods, we compare against both LIME (Ribeiro, Singh, and Guestrin 2016) and SHAP (Lundberg and Lee 2017) explanations for each model. The bracketed labels [LIME] and [SHAP] indicate the post-hoc explanation method used for each baseline model.

Comparison with Post-hoc Methods. We compare SRA against LIME and SHAP explanations across multiple Portuguese language models in Table 2. While post-hoc methods like PTTS [SHAP] achieve comparable IoU F1 scores, SRA with BERTimbau ($\alpha = 10$) shows high token-level precision (0.935) and improved overall token F1 (0.745). SRA provides these explanations intrinsically during prediction, unlike post-hoc methods that require additional computation and may not reflect the model's actual decision process. This intrinsic explainability makes SRA more appealing for real-time applications where computational efficiency is important.

Classification Performance. On HateBRXplain, SRA shows robust performance across varying rationale alignment hyperparameter values. Our ablation study reveals that Test Macro F1 remains stable across all values of α, from baseline 0.903 to a peak of 0.921 at $\alpha = 0.5$, with only minor variations (±2%) even at $\alpha = 100$. This stability, mirroring the pattern observed in English, suggests that incorporating rationale supervision does not compromise the model's hate speech detection capabilities. At $\alpha = 10$, SRA achieves a Test F1 of 0.910 (±0.008) with an accuracy of 0.907 (±0.007) and AUROC of 0.966 (±0.004), demonstrating consistent performance.

Explainability Improvements. Similar to our English results, incorporating attention alignment loss yields monotonic improvements in explainability metrics. IoU F1 increases from 0.387 at $\alpha = 0.05$ to 0.751 at $\alpha = 100$, while the baseline (no rationale alignment) achieves 0. Token F1 shows parallel gains, improving from 0.574 to 0.771. Token Precision peaks at $\alpha = 1.0$, maintaining values above 0.91 across all supervised settings. At our selected operating point ($\alpha = 10$), SRA achieves IoU F1 of 0.716 (±0.025), Token F1 of 0.745 (±0.010), with Token Precision of 0.935 (±0.005) and Token Recall of 0.668 (±0.014). These improvements suggest that SRA learns to focus on textual evidence that human annotators identify as critical for hate speech classification.

Fairness and Cross-lingual Validation. The ablation analysis shows that fairness metrics remain relatively stable across different values of α. GMB-Subgroup AUC varies slightly from 0.7165 (baseline) to 0.7157 ($\alpha = 100$), indicating minimal trade-off between explainability and subgroup fairness. At $\alpha = 10$, GMB-Subgroup AUC is 0.7120 (±0.003), GMB-BPSN AUC is 0.7211 (±0.002), and GMB-BNSP AUC improves to 0.8186 (±0.004), representing slight improvement over the baseline. These results suggest that supervised rationale alignment preserves fairness while improving explainability, with the strongest gains observed in reducing false negatives on identity-targeted toxic content (GMB-BNSP). Similar patterns across HateXplain (English) and HateBRXplain (Portuguese) suggest SRA generalizes across languages. Despite differences in dataset characteristics, HateXplain with 20,148 multi-platform samples versus HateBRXplain with 7,000 Instagram-focused samples, both exhibit similar behaviors with stable classification performance across α values, substantial explainability improvements, and comparable fairness trade-offs. These results suggest SRA is applicable to interpretable hate speech detection in different languages.

Qualitative Analysis

While quantitative metrics show SRA's overall performance, qualitative examination of attention patterns provides insights into how rationale supervision influences model behavior in practice. Understanding these patterns is essential

Model [XAI method]	Explainability				Faithfulness	
	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, Ours ($\alpha = 10$)	0.716	**0.935**	0.668	**0.745**	0.454	**-0.036**
	(±0.025)	(±0.005)	(±0.014)	(±0.010)	(±0.114)	(±0.016)

Table 2: Comparison of explainability methods on the HateBRXplain (Portuguese) test set. We evaluate post-hoc explanation methods (LIME, SHAP) against our intrinsic SRA approach across multiple Portuguese language models. Explainability metrics (IoU F1, Token Precision/Recall/F1) measure alignment with human rationales, while faithfulness metrics (Comprehensiveness, Sufficiency) assess whether explanations reflect actual model reasoning. Higher values indicate better performance for all metrics except Sufficiency (lower is better). SRA provides explanations intrinsically during prediction, while LIME and SHAP require additional post-hoc computation. Our SRA method achieves superior token precision (0.935) and competitive overall performance while providing real-time explanations. Best results are in **bold**. SRA results are averaged across 5 random seeds with standard deviations shown in parentheses.

for hate speech detection systems, where explainability requirements demand that models not only achieve high performance, but also focus on linguistically meaningful features that align with human reasoning. We examine specific examples to illustrate how SRA addresses key challenges in self-explaining hate speech detection.

Attention Alignment Patterns. Our systematic analysis reveals consistent improvements in attention-rationale alignment across rationale alignment hyperparameters, with Pearson correlation coefficients increasing from -0.084 at baseline with $\alpha = 0$ to 0.649 at $\alpha = 100$ for English, and similar patterns in Portuguese reaching 0.757. This suggests SRA guides model attention toward human-identified rationales.

SRA demonstrates qualitative differences from unsupervised approaches in attention distribution, as we see in this example from the English dataset: “Go back to where you came from — but I don’t hate all immigrants,”. Baseline BERT distributes attention across structural elements (highest attention on “you” = 0.286, “go” = 0.160, “back” = 0.120), while SRA focuses on sentiment-critical terms (“hate” = 0.394, “all” = 0.445) that human annotators identified as decisive for classification. This shift indicates SRA’s capacity to learn linguistically meaningful patterns beyond surface-level features.

Cross-linguistic consistency emerges in Portuguese examples. For “Só podia ser mulher dirigindo desse jeito,”, (“It could only be a woman driving like that”), SRA concentrates attention on stereotype-indicating phrases (“jeito” = 0.277, “desse” = 0.215, “mulher” = 0.113), while baseline models show more dispersed attention patterns, focusing on special tokens ([CLS] = 0.633). These improvements across languages indicate that SRA well aligns model attention with human rationales rather than learning dataset-specific artifacts.

Handling Identity Terms. SRA shows improved precision in aligning attention with human rationales for identity-related terminology, with token precision improving from 0.265 for unsupervised attention to 0.938 with $\alpha = 10$. This means that when SRA highlights tokens as important, they are much more likely to match the tokens that human annotators identified as rationales for their labeling decisions. BERT’s WordPiece tokenization segments certain terms into subword units, allowing examination of attention patterns across these components. We analyzed attention distributions for examples containing reclaimed identity terms used in non-offensive contexts. In one case containing “That’s my nigga right there!” (labeled as non-offensive in the dataset), BERT’s tokenizer produces subword pieces [“ni”, “##gga”]. SRA allocated attention weights of 0.665 and 0.328 to these respective tokens while correctly maintaining non-offensive classification, whereas the baseline model showed more dispersed attention across function words and sentence structure. This example illustrates how rationale supervision influences attention allocation at the subword level for identity terms that undergo WordPiece segmentation. Such patterns suggest that explicit attention supervision may help models focus on relevant linguistic features when distinguishing between harmful targeting and in-group usage contexts.

Implicit Hate Speech. SRA shows improved performance in detecting subtle offensive content where harmful intent is expressed indirectly. In the Portuguese example “Esses estrangeiros vêm aqui só para atrapalhar, deveriam voltar para o país deles” (“These foreigners come here just to cause trouble, they should go back to their country”), the model identifies xenophobic sentiment by focusing on key argumentative elements (“Esses” = 0.126, “deveriam” = 0.112, “só” = 0.111) that construct exclusionary narratives rather than explicit slurs. As illustrated in Table 3, SRA’s attention patterns reveal sensitivity to rhetorical structures common

	Baseline BERT ($\alpha = 0$)	**SRA** ($\alpha = 10$)
Refugee Example Label: Hate speech	allowing refugees into your nation is like allowing rabid foxes into your chicken coop it does not make you caring it makes you an asshole Human rationale: *allowing, refugees, rabid, foxes, nation* *Probabilities:* Normal: 2.59%, Offensive: 90.17%, Hate speech: 7.24%	allowing refugees into your nation is like allowing rabid foxes into your chicken coop it does not make you caring it makes you an asshole Human rationale: *allowing, refugees, rabid, foxes, nation* *Probabilities:* Normal: 1.70%, Offensive: 59.56%, Hate speech: 38.74%

Table 3: Attention heatmap visualization comparing baseline BERT ($\alpha = 0$) and SRA ($\alpha = 10$) on a test example involving implicit bias through metaphorical dehumanization. Color intensity represents attention weights from layer 8, head 7 (darker = higher attention). Baseline attention scatters across neutral framing terms, while SRA focuses on the problematic metaphor elements that human annotators identified as rationales for hate speech classification. This demonstrates SRA's effectiveness in learning to attend to subtle bias indicators beyond explicit hate terms.

in implicit offensive/hate speech. While the baseline model scatters attention across neutral framing terms, SRA focuses on problematic elements that human annotators identified as rationales. This suggests SRA learns to recognize argumentative patterns beyond lexical hate indicators, attending to subtle bias indicators such as categorical judgment terms, modal expressions of obligation, and restrictive qualifiers.

Performance Implications. Analysis of model predictions revealed potential labeling inconsistencies in the HateXplain dataset (see Supplementary for details). These discrepancies potentially impact fair model evaluation, as standard metrics may underestimate performance when models correctly identify harmful content that contradicts ground truth labels. To quantify this effect, we evaluated SRA on a filtered test set excluding those 335 identified problematic cases. We noticed that Test F1 of SRA improved from 0.682 to 0.796 (+16.7%), accuracy from 0.696 to 0.814 (+17.0%), and IoU F1 from 0.539 to 0.561. Fairness metrics also improved, with subgroup bias reduction showing a 9.9% relative gain compared to baseline. While these filtered results are not directly comparable to baseline methods evaluated on the full dataset, they suggest that actual model performance may be higher than what standard benchmarks indicate, especially for fairness-critical applications. The improved performance on the filtered dataset indicates that SRA's attention supervision helps the model learn patterns that sometimes conflict with the original ground truth labels. However, we acknowledge the inherent subjectivity in hate speech annotation and the possibility that some discrepancies reflect legitimate disagreements rather than clear errors.

Discussion

Key Findings and Implications. The observed improvements in attention alignment metrics, from negative correlations at baseline to significantly positive correlations under a range of alignment hyperparameters, indicate that explicit rationale supervision steers models toward more human-interpretable decisions. Our findings show that models trained with SRA focus on meaningful content words, understanding identity terms in context, and recognizing subtle hate speech patterns beyond explicit slurs, while maintaining cross-linguistic applicability between English and Portuguese datasets. These improvements suggest that human reasoning can be integrated into neural architectures for sensitive classification tasks.

A limitation of our approach is that improvements in explainability come with trade-offs in fairness metrics. While SRA achieves second-best GMB-BNSP, it shows lower GMB-Subgroup AUC compared to BERT-based baselines. This suggests that rationale supervision may improve some aspects of fairness while affecting others, requiring careful evaluation in deployment scenarios. Additionally, SRA requires rationale-annotated training data, which is more expensive to obtain than standard classification labels, potentially limiting scalability to new domains or languages.

Conclusion

This work suggests that explicit rationale supervision offers a potentially viable path toward self-explaining hate speech detection without compromising classification performance. The SRA framework demonstrates improvements in attention alignment across languages while maintaining competitive accuracy, contributing to a broader movement toward explainable AI in sensitive domains. However, broader implications extend beyond technical metrics to fundamental questions about automated systems moderating human communication. While improved interpretability facilitates responsible AI systems that meet desired transparency requirements, it cannot resolve the underlying tensions between protecting individuals from harm and preserving open dialogue in societies.

The systematic evaluation approach presented here provides a methodological foundation for future research at the intersection of interpretability, fairness, and performance in socially sensitive applications. As AI systems increasingly influence consequential decisions about human communication and beyond, the development of self-explaining models that better align with human rationales is an important step toward responsible AI.

Acknowledgments

This work was supported by the Research Council of Norway through FRIPRO Grant under project number 356103, its Centres of Excellence scheme, Integreat - Norwegian Centre for knowledge-driven machine learning under project number 332645, and UiO Life Sciences Summer Internship. The computations were performed on resources provided by Educloud Research infrastructure at UiO.

References

Attanasio, G.; Nozza, D.; Hovy, D.; and Baralis, E. 2022. Entropy-based Attention Regularization Frees Unintended Bias Mitigation from Lists. In Muresan, S.; Nakov, P.; and Villavicencio, A., eds., *Findings of the Association for Computational Linguistics: ACL 2022*, 1105–1119. Dublin, Ireland: Association for Computational Linguistics.

Balkir, E.; Kiritchenko, S.; Nejadgholi, I.; and Fraser, K. 2022. Challenges in Applying Explainability Methods to Improve the Fairness of NLP Models. In *Proceedings of the 2nd Workshop on Trustworthy Natural Language Processing*, 80–92. Seattle, USA.: Association for Computational Linguistics.

Bao, Y.; Chang, S.; Yu, M.; and Barzilay, R. 2018. Deriving Machine Attention from Human Rationales. In Riloff, E.; Chiang, D.; Hockenmaier, J.; and Tsujii, J., eds., *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 1903–1913. Brussels, Belgium: Association for Computational Linguistics.

Bastings, J.; Aziz, W.; and Titov, I. 2019. Interpretable Neural Predictions with Differentiable Binary Variables. In Korhonen, A.; Traum, D.; and Màrquez, L., eds., *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2963–2977. Florence, Italy: Association for Computational Linguistics.

Borkan, D.; Dixon, L.; Sorensen, J.; Thain, N.; and Vasserman, L. 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, 491–500. San Francisco, CA, USA: ACM.

Brunk, J.; Mattern, J.; and Riehle, D. M. 2019. Effect of Transparency and Trust on Acceptance of Automatic Online Comment Moderation Systems. In *2019 IEEE 21st Conference on Business Informatics (CBI)*, volume 01, 429–435.

Calabrese, A.; Neves, L.; Shah, N.; Bos, M.; Ross, B.; Lapata, M.; and Barbieri, F. 2024. Explainability and Hate Speech: Structured Explanations Make Social Media Moderators Faster. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 398–408. Bangkok, Thailand: Association for Computational Linguistics.

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

Clarke, C.; Hall, M.; Mittal, G.; Yu, Y.; Sajeev, S.; Mars, J.; and Chen, M. 2023. Rule By Example: Harnessing Logical Rules for Explainable Hate Speech Detection. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 364–376. Toronto, Canada: Association for Computational Linguistics.

Davani, A. M.; Atari, M.; Kennedy, B.; and Dehghani, M. 2023. Hate Speech Classifiers Learn Normative Social Stereotypes. *Transactions of the Association for Computational Linguistics*, 11: 300–319.

Davidson, T.; Bhattacharya, D.; and Weber, I. 2019. Racial Bias in Hate Speech and Abusive Language Detection Datasets. In *Proceedings of the 3rd Workshop on Abusive Language Online*, 25–35. Florence, Italy.

Davidson, T.; Warmsley, D.; Macy, M.; and Weber, I. 2017. Automated Hate Speech Detection and the Problem of Offensive Language. *Proceedings of the International AAAI Conference on Web and Social Media*, 11(1): 512–515.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

DeYoung, J.; Jain, S.; Rajani, N. F.; Lehman, E.; Xiong, C.; Socher, R.; and Wallace, B. C. 2020. ERASER: A Benchmark to Evaluate Rationalized NLP Models. In Jurafsky, D.; Chai, J.; Schluter, N.; and Tetreault, J., eds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4443–4458. Online: Association for Computational Linguistics.

Dixon, L.; Li, J.; Sorensen, J.; Thain, N.; and Vasserman, L. 2018. Measuring and Mitigating Unintended Bias in Text Classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, 67–73. New York, USA: Association for Computing Machinery. ISBN 9781450360128.

European Parliament and Council. 2016. General Data Protection Regulation. Official Journal of the European Union. Regulation (EU) 2016/679, Article 15(1)(h).

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

Gongane, V. U.; Munot, M. V.; and Anuse, A. D. 2024. A survey of explainable AI techniques for detection of fake news and hate speech on social media platforms. *Journal of Computational Social Science*, 7: 587–623.

Gorwa, R.; Binns, R.; and Katzenbach, C. 2020. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, 7(1).

Guidotti, R.; Monreale, A.; Ruggieri, S.; Turini, F.; Giannotti, F.; and Pedreschi, D. 2018. A Survey of Methods for Explaining Black Box Models. *ACM Computing Surveys*, 51(5).

Hutchinson, B.; Prabhakaran, V.; Denton, E.; Webster, K.; Zhong, Y.; and Denuyl, S. 2020. Social Biases in NLP Models as Barriers for Persons with Disabilities. In Jurafsky, D.; Chai, J.; Schluter, N.; and Tetreault, J., eds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5491–5501. Online: Association for Computational Linguistics.

Jain, S.; Wiegreffe, S.; Pinter, Y.; and Wallace, B. C. 2020. Learning to Faithfully Rationalize by Construction. In Jurafsky, D.; Chai, J.; Schluter, N.; and Tetreault, J., eds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4459–4473. Held Online: Association for Computational Linguistics.

Jiang, Z.; Zhang, Y.; Yang, Z.; Zhao, J.; and Liu, K. 2021. Alignment Rationale for Natural Language Inference. In Zong, C.; Xia, F.; Li, W.; and Navigli, R., eds., *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 5372–5387. Online: Association for Computational Linguistics.

Jørgensen, R.; Caccavale, F.; Igel, C.; and Søgaard, A. 2022. Are Multilingual Sentiment Models Equally Right for the Right Reasons? In Bastings, J.; Belinkov, Y.; Elazar, Y.; Hupkes, D.; Saphra, N.; and Wiegreffe, S., eds., *Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, 131–141. Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics.

Kim, J.; Lee, B.; and Sohn, K.-A. 2022. Why Is It Hate Speech? Masked Rationale Prediction for Explainable Hate Speech Detection. In Calzolari, N.; Huang, C.-R.; Kim, H.; Pustejovsky, J.; Wanner, L.; Choi, K.-S.; Ryu, P.-M.; Chen, H.-H.; Donatelli, L.; Ji, H.; Kurohashi, S.; Paggio, P.; Xue, N.; Kim, S.; Hahm, Y.; He, Z.; Lee, T. K.; Santus, E.; Bond, F.; and Na, S.-H., eds., *Proceedings of the 29th International Conference on Computational Linguistics*, 6644–6655. Gyeongju, Republic of Korea: International Committee on Computational Linguistics.

Lehman, E.; DeYoung, J.; Barzilay, R.; and Wallace, B. C. 2019. Inferring Which Medical Treatments Work from Reports of Clinical Trials. In Burstein, J.; Doran, C.; and Solorio, T., eds., *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 3705–3717. Minneapolis, Minnesota: Association for Computational Linguistics.

Lei, T.; Barzilay, R.; and Jaakkola, T. 2016. Rationalizing Neural Predictions. In Su, J.; Duh, K.; and Carreras, X., eds., *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 107–117. Austin, Texas: Association for Computational Linguistics.

Liu, B.; and Lane, I. 2016. Attention-Based Recurrent Neural Network Models for Joint Intent Detection and Slot Filling. In *Interspeech 2016*.

Mathew, B.; Saha, P.; Yimam, S. M.; Biemann, C.; Goyal, P.; and Mukherjee, A. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 14867–14875.

May, C.; Wang, A.; Bordia, S.; Bowman, S. R.; and Rudinger, R. 2019. On Measuring Social Biases in Sentence Encoders. In Burstein, J.; Doran, C.; and Solorio, T., eds., *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 622–628. Minneapolis, Minnesota: Association for Computational Linguistics.

Meta Oversight Board. 2025. Content Moderation in a New Era for AI and Automation.

Poletto, F.; Basile, V.; Sanguinetti, M.; Bosco, C.; and Patti, V. 2021. Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, 55(3): 477–523.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *arXiv preprint arXiv:1602.04938*.

Salles, I.; Vargas, F.; and Benevenuto, F. 2025. HateBRXplain: A Benchmark Dataset with Human-Annotated Rationales for Explainable Hate Speech Detection in Brazilian Portuguese. In Rambow, O.; Wanner, L.; Apidianaki, M.; Al-Khalifa, H.; Eugenio, B. D.; and Schockaert, S., eds., *Proceedings of the 31st International Conference on Computational Linguistics*, 6659–6669. Abu Dhabi, UAE: Association for Computational Linguistics.

Schuster, M.; and Paliwal, K. K. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11): 2673–2681.

Souza, F.; Nogueira, R.; and Lotufo, R. 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)*.

Strout, J.; Zhang, Y.; and Mooney, R. 2019. Do Human Rationales Improve Machine Explanations? In Linzen, T.; Chrupała, G.; Belinkov, Y.; and Hupkes, D., eds., *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 56–62. Florence, Italy: Association for Computational Linguistics.

Ungless, E. L.; Vitsakis, N.; Talat, Z.; Garforth, J.; Ross, B.; Onken, A.; Kasirzadeh, A.; and Birch, A. 2025. The Only Way is Ethics: A Guide to Ethical Research with Large Language Models. In Rambow, O.; Wanner, L.; Apidianaki, M.; Al-Khalifa, H.; Eugenio, B. D.; and Schockaert, S., eds., *Proceedings of the 31st International Conference on Computational Linguistics*, 8992–9005. Abu Dhabi, UAE: Association for Computational Linguistics.

Vargas, F.; Carvalho, I.; Hürriyetoğlu, A.; Pardo, T.; and Benevenuto, F. 2023. Socially Responsible Hate Speech Detection: Can Classifiers Reflect Social Stereotypes? In Mitkov, R.; and Angelova, G., eds., *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, 1187–1196. Varna, Bulgaria: INCOMA Ltd., Shoumen, Bulgaria.

Vargas, F.; Carvalho, I.; Pardo, T. A. S.; and Benevenuto, F. 2024. Context-aware and expert data resources for Brazilian Portuguese hate speech detection. *Natural Language Processing*, 31(2): 435–456.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 6000–6010. Red Hook, NY, USA: Curran Associates Inc. ISBN 9781510860964.

Zampieri, M.; Malmasi, S.; Nakov, P.; Rosenthal, S.; Farra, N.; and Kumar, R. 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*, 1415–1420. Minnesota, United States.

Zhang, Y.; Marshall, I.; and Wallace, B. C. 2016. Rationale-Augmented Convolutional Neural Networks for Text Classification. In Su, J.; Duh, K.; and Carreras, X., eds., *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 795–804. Austin, Texas: Association for Computational Linguistics.

Zhang, Z.; Robinson, D.; and Tepper, J. 2018. Detecting Hate Speech on Twitter Using a Convolution-GRU Based Deep Neural Network. In *The Semantic Web: 15th International Conference, ESWC 2018*, 745–760. Springer.