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

by Brage Eilertsen and Røskva Bjørgfinsdóttir

Introduction

Hate speech detection systems are widely used in online media platforms, with Meta reporting over 7 million hate speech appeals monthly.

They rely on classification algorithms that currently operate as ‘black boxes’. Because we do not know exactly how decisions are reached, we risk reinforcing societal biases and marginalizing vulnerable groups.

We need explainability in AI for transparency, accountability and fairness.

Method

We trained BERT-models on HateXplain and HateBRXplain datasets with Supervised Rational Attention (SRA), a novel framework to explicitly align model attention with human-annotated rationales.

AAL minimizes MSE between normalized attention and human rationales r:

Attention alignment loss

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

We combine the standard cross-entropy loss for classification and the supervised AAL:

Overall training objective

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

α is a hyperparameter controlling the strength of attention supervision, balancing attention alignment and classification.

Results

2.4x better explainability (IoU F1: 0.539 vs best baseline 0.222). 9.4% bias reduction in negative subgroup predictions. Maintains 98% classification performance. SRA learns to focus on the same harmful language patterns that human annotators identify as evidence of hate speech.

Table 1: HateXplain Results

Model	Classification		Explainability			Bias (AUC)		
[Explanation method]	F1	AUROC	IoU F1	Token F1	AUPRC	GMB-Sub	GMB-BPSN	GMB-BNSP
CNN-GRU [LIME]	0.606	0.793	0.167	0.385	0.648	0.654	0.623	0.659
BiRNN [LIME]	0.575	0.767	0.162	0.361	0.605	0.66	0.64	0.671
BiRNN-Attn [Attn]	0.614	0.795	0.167	0.369	0.643	0.653	0.662	0.668
BiRNN-Attn [LIME]	0.614	0.795	0.162	0.386	0.65	0.653	0.662	0.668
BiRNN-HateXplain [Attn]	0.629	0.805	0.222	0.506	**0.841**	0.691	0.691	0.674
BiRNN-HateXplain [LIME]	0.629	0.805	0.174	0.407	0.685	0.691	0.691	0.674
BERT [Attn]	0.674	0.843	0.13	0.497	0.778	0.762	0.709	0.757
BERT [LIME]	0.674	0.843	0.118	0.468	0.747	0.762	0.709	0.757
BERT-HateXplain [Attn]	**0.687**	0.851	0.12	0.411	0.626	**0.807**	**0.745**	0.763
BERT-HateXplain [LIME]	**0.687**	0.851	0.112	0.452	0.722	**0.807**	**0.745**	0.763
SRA Ours (α=10)	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)

IoU F1 measures overlap between model attention and human-identified rationales (1.0 = perfect alignment). **Token F1** treats each token as a separate prediction.

GMB-Sub compares offensive and normal posts mentioning identities. Higher score = better at distinguishing. **GMB-BPSN**: Normal posts with identities vs offensive posts without identities. Higher score = avoids false positives. **GMB-BNSP**: Offensive posts with identities vs normal posts without identities. Higher score = avoids false negatives.

Refugee example from HateXplain:

	Baseline BERT (α = 0)	**SRA (α = 10)**
Model rationale:	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	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*	*allowing, refugees, rabid, foxes, nation*
Probabilities:	Normal: 2.59%, Offensive: 90.17%, Hate speech: 7.24%	Normal: 1.70%, Offensive: 59.56%, Hate Hate speech: 38.74%

Correct label is hate speech because of metaphorical dehumanization. Colour intensity represents attention weights from layer 8, head 7. Baseline attention is on neutral framing terms, while SRA catches the metaphor. SRA’s strength lies in learning to attend to bias indicators beyond explicit hate terms.

Real world implications

SRA can make content moderation more transparent by providing real-time explanations that align with human reasoning. This can help platforms meet regulatory requirements like GDPR’s ‘right to explanation’. The approach shows promise in reducing some aspects of bias against identity groups, but more research is needed in developing methods that improve fairness metrics. When we are implementing hate speech detection systems we need to act carefully, because there is a fine line between protecting vulnerable communities and censorship. By making these systems more transparent we can hopefully strike a good balance.

SRA could also enhance interpretability in other high-stakes domains like healthcare diagnostics, criminal justice risk assessment, and hiring algorithms. One limiting factor is the cost of annotating rationales manually, but it could be well worth it in areas where both accuracy and accountability matter.

link-to-paper-or-code.com?

USP Universidade de São Paulo

UiO