Mitigating Demographic Disparities in Explanation–Human Alignment via Rationale Regularization

Anonymous EMNLP submission

Abstract

We study explanation fairness in hate speech detection, where strong predictive performance does not guarantee equally plausible explanations across demographic groups. Using the HateXplain corpus, we introduce Plausibility-EFG, a group-level disparity metric that measures the worst-case gap in explanation plausibility via alignment between token attributions and human rationale annotations. We evaluate BERT, RoBERTa, and an adversarially trained variant, ADV-RoBERTa, using Integrated Gradients and SHAP. Results show that ADV-RoBERTa improves both classification performance and rationale alignment while substantially reducing subgroup explanation disparities under both attribution methods. These improvements remain consistent across demographic identities and stable across train, validation, and test splits. We further show that prediction-level debiasing alone does not eliminate explanation inequity, indicating that predictive fairness and explanation fairness are related but distinct objectives. Finally, fairness–performance ablations reveal a controlled tradeoff in which stronger adversarial regularization reduces Plausibility-EFG at the cost of modest macro-F1 degradation. Overall, the paper establishes explanation fairness as a measurable objective in hate speech moderation and presents Fair-REF as an effective mitigation strategy for reducing subgroup disparities in model explanations.

1 Introduction

Hate speech moderation is a high-stakes setting in which transformer classifiers are often evaluated primarily by predictive performance, even though their token-level explanations are used to support reviewer trust and borderline moderation decisions. This makes explanation quality practically consequential, not merely diagnostic. Recent work has already shown that explanation behavior can vary across demographic proxies in hate speech settings, and that such disparities may persist even when classification looks strong (Mathew et al., 2021), (Balagopalan et al., 2022), (Brandl et al., 2024). Consequently, predictive fairness does not imply explanation fairness, and explanation quality should be audited independently of output-level performance.

This paper studies a failure mode that is both methodologically underexplored and practically important: explanations can be systematically less plausible for some demographic groups than for others. We introduce Plausibility-EFG, a group disparity metric for explanation–human alignment, defined as the worst-case gap in mean plausibility between demographic groups, where plausibility is measured by AUPRC over attribution scores and human rationale annotations. This formulation follows the distinction between faithfulness and plausibility emphasized by (Jacovi and Goldberg, 2020) and the rationale-based evaluation paradigm of ERASER (DeYoung et al., 2020). We then test an orthogonality hypothesis: reducing prediction-level disparity does not necessarily reduce explanation-level disparity, because attribution methods may behave unevenly across subpopulations (Balagopalan et al., 2022), (Brandl et al., 2024).

This work makes four contributions. First, we introduce Plausibility-EFG, a formal metric for demographic disparities in token-level explanation plausibility. Second, we present the first subgroup rationale benchmark for hate speech detection by exploding HateXplain into 27,566 post-target pairs across eight demographic groups. Third, we provide a unified evaluation framework that jointly measures classification accuracy, explanation plausibility, and explanation fairness using AUPRC, IOU-F1, Kruskal-Wallis testing, Cohen's d, and bootstrap confidence intervals. Fourth, we propose Fair-REF, an adversarial rationale-aligned mitigation pipeline that directly optimizes explanation fairness and substantially reduces IG and SHAP

Plausibility-EFG while preserving competitive classification performance (Sundararajan et al., 2017), (Lundberg and Lee, 2017).

2 Related Work

Post-hoc explanation methods are now standard tools for analyzing neural NLP classifiers, especially in settings where token-level evidence is important. Gradient-based methods such as Integrated Gradients (IG) provide axiomatic attribution scores by integrating gradients from a baseline to the input and have become a common choice for transformer interpretability (Sundararajan et al., 2017). SHAP offers a complementary, game-theoretic view of feature attribution by estimating token contributions through Shapley values (Lundberg and Lee, 2017). In NLP, these methods are frequently paired with rationale benchmarks such as ERASER and HateXplain, where model attributions can be compared against human-annotated evidence spans (DeYoung et al., 2020), (Mathew et al., 2021). Rationale extraction models such as REF further move explanation generation into the training objective by jointly optimizing prediction and sparse token selection (Lei et al., 2016).

A key conceptual distinction in explanation evaluation is the difference between faithfulness and plausibility. Faithfulness asks whether an explanation reflects the model's actual decision process, while plausibility asks whether it matches human judgment or annotation. (Jacovi and Goldberg, 2020) argue that these criteria are not interchangeable and should be evaluated separately; an explanation can be faithful without being human-aligned, and plausible without being faithful. This distinction is especially important in rationale-based NLP settings, where token overlap with human evidence spans provides a natural plausibility signal. Our work follows this view and treats explanation–human alignment as a separate object of study rather than a proxy for faithfulness alone (Jacovi and Goldberg, 2020), (DeYoung et al., 2020).

Fairness in NLP has focused on prediction-level disparities such as demographic parity, equalized odds, and subgroup performance gaps. Prior work has documented how bias mitigation, counterfactual augmentation, and adversarial debiasing can improve model outputs, but these methods are generally evaluated at the label level rather than at the explanation level (Blodgett et al., 2020), (Zhang et al., 2018). Concurrently , recent studies show that explanation quality itself can differ across demographic groups. (Balagopalan et al., 2022) found that post-hoc explanation fidelity varies substantially across protected subgroups in non-NLP tabular settings, suggesting that explanation fairness is not guaranteed even when models are accurate. More recently, (Brandl et al., 2024) showed that fairness and explainability do not necessarily co-improve in NLP, reinforcing the idea that these objectives can be empirically orthogonal.

Our work differs from these prior studies in three important ways. First, we study plausibility, not fidelity: our question is whether attributions align with human rationales, not whether they approximate the model's internal logic. Second, we focus on an NLP-native benchmark with token-level rationale annotations and demographic target labels, enabling direct measurement of explanation–human alignment fairness on HateXplain (Mathew et al., 2021). Third, we do not stop at auditing disparities: we propose a training-time mitigation objective, Fair-REF, that explicitly regularizes group-level rationale quality. In this sense, our work extends prior explanation-fairness audits by moving from diagnosis to mitigation, and from prediction fairness to explanation fairness.

Our study is also related to emerging work on fair explanations in moderation contexts. Recent research on hate speech moderation has highlighted the social consequences of uneven explanation quality for content reviewers and downstream users, especially for posts targeting marginalized communities (Nguyen et al., 2023). This line of work motivates our orthogonality hypothesis: that debiasing the classifier's predictions does not necessarily debias its explanations. We test this hypothesis empirically and show that explanation fairness requires a dedicated objective rather than a byproduct of prediction fairness alone (Balagopalan et al., 2022) ,(Brandl et al., 2024).

3 Theoretical Justification

Let $D = \{(x_i, y_i, g_i, r_i)\}_{i=1}^{N}$ denote a corpus of posts, class labels, demographic groups, and human rationale masks, where $x_i = (x_{i1}, \ldots, x_{iT_i})$ is a token sequence, $y_i \in \{1, \ldots, C\}$, $g_i \in G$, and $r_i \in \{0, 1\}^{T_i}$. The classifier f_θ induces logits $z_\theta(x_i) \in R^C$ and predictive distribution $p_\theta(y \mid x_i) = \text{softmax}(z_\theta(x_i))$. The central hypothesis of this work is that hate-speech modeling should be optimized not only for predictive accu-

racy, but also for rationale alignment and subgroup-consistent explanation quality, which is well motivated by rationale-based NLP, benchmarked explanation evaluation, and recent work on fairness as decision-rationale alignment (Lei et al., 2016), (DeYoung et al., 2020), (Mathew et al., 2021).

We therefore define the training objective as a three-term multi-objective program:

$$L_{total}(\theta, \phi) = L_{task}(\theta) + \lambda_{rat} L_{rat}(\theta) + \lambda_{fair} L_{adv}(\theta, \phi),$$

with supervised task loss

$$L_{task}(\theta) = -\frac{1}{N} \sum_{i=1}^{N} \log p_\theta(y_i \mid x_i),$$

rationale-alignment loss

$$L_{rat}(\theta) = \frac{1}{N} \sum_{i=1}^{N} \ell_{align}(s_i, r_i),$$

and adversarial demographic loss

$$L_{adv}(\theta, \phi) = -\frac{1}{N} \sum_{i=1}^{N} \log q_\phi(g_i \mid h_\theta(x_i)).$$

Here $h_\theta(x_i)$ is a shared representation and q_ϕ is the group discriminator. Under gradient reversal, the effective optimization is $\min_\theta \max_\phi L_{task} + \lambda_{rat} L_{rat} - \lambda_{fair} L_{adv}$, which follows the standard adversarial debiasing formulation of (Zhang et al., 2018) and the rationale-alignment framing of (Li et al., 2023).

For post hoc explanation, let $A_\theta(x_i) \in R^{T_i}$ denote a token attribution operator and let $s_i = \mathrm{Norm}(\max(A_\theta(x_i), 0)) \in [0, 1]^{T_i}$ be the normalized non-negative attribution vector. In the case of Integrated Gradients, the attribution for token t is

$$IG_t(x) = (x_t - x'_t) \int_0^1 \frac{\partial F(x' + \alpha(x - x'))}{\partial x_t} \, d\alpha,$$

with numerical approximation

$$IG_t(x) \approx (x_t - x'_t) \frac{1}{m} \sum_{k=1}^{m} \frac{\partial F(x' + \frac{k}{m}(x - x'))}{\partial x_t},$$

which follows the axiomatic attribution framework of (Sundararajan et al., 2017). For SHAP, the contribution of token t is defined by the Shapley-value estimator

$$\phi_t = \sum_{S \subseteq N \setminus \{t\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} (f(S \cup \{t\}) - f(S)),$$

following (Lundberg and Lee, 2017). Using both methods is theoretically important because gradient-path and coalition-based explainers can preserve model ordering while differing in attribution geometry and scale, a distinction that matters for fairness analysis (Dai et al., 2022), (Brandl et al., 2024).

Given attribution vector s_i and rationale mask r_i, plausibility is evaluated by ranking-based and overlap-based criteria:

$$\mathrm{AUPRC}(s_i, r_i) = \int_0^1 P_i(\tau) \, dR_i(\tau),$$

$$\tau_i = \mathrm{median}(s_i), \quad \hat{r}_{it} = I[s_{it} \geq \tau_i],$$

$$\text{IOU-F1}(i) = \frac{|\hat{r}_i \cap r_i|}{|\hat{r}_i \cup r_i|}.$$

This follows the standard evidence-plausibility setup used in ERASER and HateXplain, where human rationales are treated as evaluative evidence spans rather than absolute causal ground truth (DeYoung et al., 2020), (Mathew et al., 2021), (Jacovi and Goldberg, 2020).

To quantify subgroup fairness of explanations, let $D_g = \{i : g_i = g\}$ denote the subgroup-specific evaluation set and let $M_i \in \{\mathrm{AUPRC}(s_i, r_i), \text{IOU-F1}(i)\}$. The group-wise mean plausibility is

$$\mu_g = \frac{1}{|D_g|} \sum_{i \in D_g} M_i.$$

We define the explanation fairness gap as the worst-case inter-group disparity

$$\mathrm{EFG} = \max_{g, g' \in G} |\mu_g - \mu_{g'}|.$$

This is the central fairness quantity in the paper. It directly operationalizes the claim that explanations should be comparably plausible across demographic groups, and it is conceptually aligned with recent work showing that explanation quality can itself be unevenly distributed across subpopulations (Dai et al., 2022), (Balagopalan et al., 2022), (Nguyen et al., 2023), (Brandl et al., 2024).

We further assess statistical heterogeneity across groups by the Kruskal–Wallis statistic

$$H = \frac{12}{N(N+1)} \sum_{g=1}^{|G|} \frac{R_g^2}{n_g} - 3(N+1),$$

and practical disparity by Cohen's effect size between the best and worst groups,

$$d = \frac{\mu_{g^\star} - \mu_{g^\dagger}}{s_{pooled}}, \quad s_{pooled} = \sqrt{\frac{s_{g^\star}^2 + s_{g^\dagger}^2}{2}}.$$

The combination of EFG, significance testing, and effect size separates worst-case gap magnitude from rank-level distributional differences, which is important because subgroup imbalance can distort simple point estimates (Dai et al., 2022), (Balkir et al., 2022).

The main theoretical claim is that prediction fairness and explanation fairness are only partially aligned. In differential form, this can be written as

$$\nabla_\theta L_{task}(\theta) \nabla_\theta \text{EFG}(\theta),$$

and in optimization terms as

$$\Delta L_{task} \downarrow \quad \Delta \text{EFG} \downarrow .$$

That is, improving classification risk or even output-level fairness does not guarantee improved explanation equity. This observation is consistent with recent analyses of the interplay between fairness and explainability in NLP and with explanation-quality disparity studies showing that post hoc explanations may remain biased even for strong models (Dai et al., 2022), (Balagopalan et al., 2022), (Balkir et al., 2022), (Brandl et al., 2024).

Finally, we interpret Fair-REF as a constrained multi-objective program:

$$\min_\theta L_{task}(\theta) + \lambda_{rat} L_{rat}(\theta) \quad \text{s.t.} \quad \text{EFG}(\theta) \leq \varepsilon.$$

The implemented objective is its Lagrangian relaxation:

$$J(\theta, \phi, \lambda_{fair}) = L_{task}(\theta) + \lambda_{rat} L_{rat}(\theta) + \lambda_{fair} L_{adv}(\theta, \phi).$$

This framing makes the fairness–utility tradeoff explicit and explains why the λ_{fair}-sweep induces a Pareto frontier rather than a monotone improvement in every metric. The analytical properties of EFG, the variance behavior of subgroup estimators, and the partial orthogonality claim are formalized in the appendix.

Core analytical statements

Proposition 1 (Basic properties of EFG). For any set of group means $\{\mu_g\}_{g \in G} \subseteq [0, 1]$, EFG $\in [0, 1]$, is symmetric in the group indices, and equals zero if and only if $\mu_g = \mu_{g'}$ for all $g, g' \in G$. Proof in Section 8.1.

Proposition 2 (Subgroup-scarcity variance law). If M_i has finite variance within each group g, then $\text{Var}(\mu_g) = \sigma_g^2 / |D_g|$; hence low-support groups induce higher uncertainty in μ_g, EFG, and associated effect-size estimates. Proof in Appendix C.

Observation 1 (Partial orthogonality). Reducing L_{task} or increasing predictive performance does not in general imply a reduction in EFG. A sufficient explanation is that task gradients and explanation-fairness gradients need not be aligned in parameter space. Formal argument in Appendix C.

Proposition 3 (Tradeoff under fairness regularization). For sufficiently large λ_{fair}, minimization of the Lagrangian objective can reduce representation-level group information while also suppressing task-relevant features, yielding a non-trivial fairness–utility Pareto tradeoff. Proof sketch in Appendix C.

Observation 2 (Explainer-level ranking stability without value equivalence). Two normalized attribution methods can preserve model ordering while differing in absolute plausibility values. This explains why IG and SHAP can agree on comparative ranking but differ in metric scale. Formal argument in Appendix C.

4 Methodology

4.1 Problem Formulation

Let $D = \{(x_i, y_i, r_i, g_i)\}_{i=1}^N$ denote the HateXplain corpus, where $x_i = (w_1, \ldots, w_{T_i})$ is the input post, $y_i \in Y$ is the majority-vote label, $r_i \in \{0, 1\}^{T_i}$ is the human rationale mask, and $g_i \in G$ is the demographic target group. The label space is $Y = \{\text{normal, offensive, hatespeech}\}$, and $G = \{\text{African, Islam, Women, Jewish, Homosexual, Refugee, Arab, Caucasian}\}$. Our objective is to learn a classifier $f_\theta : X \to Y$ that jointly optimizes predictive accuracy, explanation plausibility through alignment with r_i, and explanation fairness through uniform plausibility across demographic groups.

4.2 Model Setup

We evaluate BERT, RoBERTa, and ADV-RoBERTa. ADV-RoBERTa extends RoBERTa with a shared encoder E_θ, a task classification head C_ψ, a demographic adversary A_ϕ, and a rationale predic-

tion head R_ξ (Zhang et al., 2018), (Mathew et al., 2021). Given input x, the encoder produces hidden states $H = E_\theta(x)$ and a pooled representation $h = H_{[\text{CLS}]}$. The model predicts the class label, demographic group, and token-level rationale scores as

$$\hat{y} = C_\psi(h), \quad \hat{g} = A_\phi(GRL_\lambda(h)), \quad \hat{r}_t = \sigma(R_\xi(H_t)).$$

4.3 Training Objective

Training combines classification, adversarial debiasing, and rationale supervision in a single objective. The task loss, adversarial loss, and rationale loss are defined as

$$L_{task} = -\sum_{k \in Y} I[y = k] \log \hat{y}_k,$$

$$L_{adv} = -\sum_{k \in G} I[g = k] \log \hat{g}_k,$$

$$L_{rat} = -\frac{1}{T}\sum_{t=1}^{T}[r_t \log \hat{r}_t + (1 - r_t)\log(1 - \hat{r}_t)].$$

The full objective is

$$L_{total} = L_{task} + \lambda_{adv} L_{adv} + \lambda_{rat} L_{rat},$$

with $\lambda_{adv} = 0.5$ and $\lambda_{rat} = 0.2$ in the primary setting. For fairness ablations, we replace λ_{adv} with $\lambda_{fair} \in \{0.0, 0.1, 0.3, 0.5, 1.0\}$.

4.4 Attribution Methods

To measure whether model evidence aligns with human rationales, we use Integrated Gradients (IG) and SHAP (Sundararajan et al., 2017), (Lundberg and Lee, 2017). For IG,

$$IG_t(x) = (x_t - x'_t)\int_0^1 \frac{\partial F(x' + \alpha(x - x'))}{\partial x_t}\, d\alpha.$$

For SHAP,

$$\phi_t = \sum_{S \subseteq N \setminus \{t\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!}[f(S \cup \{t\}) - f(S)]$$

4.5 Plausibility and Fairness Metrics

Let $s_i \in [0, 1]^{T_i}$ denote the normalized attribution vector and $r_i \in \{0, 1\}^{T_i}$ the human rationale mask. We measure plausibility using AUPRC and IOU-F1:

$$\text{AUPRC}(s_i, r_i) = \int_0^1 P(\tau)\, dR(\tau),$$

$$\hat{r}_{it} = I[s_{it} \geq \tau_i], \quad \tau_i = \text{median}(s_i),$$

$$\text{IOU-F1}(i) = \frac{|\hat{r}_i \cap r_i|}{|\hat{r}_i \cup r_i|}.$$

For each group $g \in G$, mean subgroup plausibility is

$$\mu_g = \frac{1}{|D_g|}\sum_{i \in D_g} M_i,$$

where $M_i \in \{\text{AUPRC}(s_i, r_i), \text{IOU-F1}(i)\}$. We define Plausibility-EFG as

$$\text{EFG} = \max_{g_i, g_j \in G} |\mu_{g_i} - \mu_{g_j}|.$$

We further report Kruskal–Wallis significance, Cohen's d, and bootstrap confidence intervals.

4.6 Orthogonality and Mitigation

Prediction fairness and explanation fairness are treated as related but distinct objectives. ADV-RoBERTa therefore combines adversarial debiasing and rationale supervision as a direct mitigation strategy for subgroup explanation disparity. To study the fairness–performance frontier, we sweep λ_{fair} and evaluate the resulting tradeoff between Plausibility-EFG and predictive utility. Additional implementation details and expanded methodological explanations are provided in Appendix D.

5 Experiments and Results

This section evaluates hate speech classification, explanation plausibility, subgroup fairness, explanation fairness gaps, and fairness-aware mitigation using Integrated Gradients and SHAP. Explanation quality is measured with AUPRC and IOU-F1, and fairness disparity with EFG.

5.1 Dataset and Evaluation Pipeline

Experiments are conducted on HateXplain, which contains 20,148 unique social media posts annotated with post-level class labels, token-level rationales, and demographic target labels. The original corpus is transformed into multiple derived views: an annotation-level corpus of 60,444 rows, an exploded subgroup corpus of 53,161 rows, and a final fairness evaluation set of 27,566 rationale-bearing subgroup instances. The benchmark is therefore evaluated at two levels: the post level for classification and general plausibility, and the exploded subgroup level for explanation fairness. [1] Table 9 : Appendix Table A4 summarizes the dataset

[1] https://github.com/anonymousengineerss/Mitigating-Demographic-Disparities-in-Explanation-Human-Alignment-via-Rationale-Regularization

statistics and evaluation splits. The corpus is imbalanced in both class labels and demographic representation, which motivates fairness-aware evaluation beyond overall accuracy. Figure 1 visualizes the post-level label imbalance, which is one reason macro-F1 and explanation-based metrics are more informative than accuracy alone.

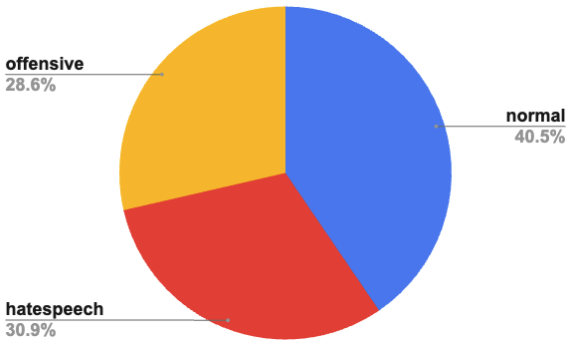

Figure 1: Post-level label distribution in HateXplain, showing class imbalance across normal, hatespeech, and offensive categories.

The experimental pipeline follows the sequence: classification, attribution, plausibility scoring, subgroup benchmarking, fairness-gap estimation, and mitigation analysis. This structure ensures that predictive quality, explanation quality, and explanation equity are evaluated independently before being compared jointly.

5.2 Hate Speech Classification

We first evaluate the base task of three-way hate speech classification using BERT, RoBERTa, and ADV-RoBERTa. This phase establishes whether fairness-aware training affects predictive performance before explanation quality is considered. Table 1 shows that ADV-RoBERTa achieves the strongest overall results, outperforming both baselines in accuracy, macro-F1, and weighted-F1. Relative to BERT, it improves accuracy by 0.0891 and macro-F1 by 0.0675, indicating that fairness-aware optimization does not degrade the classification task. The class-wise results are similarly informative. Offensive remains the hardest class, but ADV-RoBERTa improves offensive F1 from 0.5535 to 0.6489 and raises hate speech F1 to 0.8173. Figure 3 in the appendix shows the same trend: BERT confuses offensive and hate speech more often, RoBERTa reduces some overlap, and ADV-RoBERTa achieves the clearest separation. These gains suggest that adversarial demographic suppression and rationale supervision help the model better separate borderline abusive content from both normal and explicitly hateful posts.

5.3 Explanation Plausibility

We next test whether model explanations align with human rationale annotations. Let $s_i \in [0,1]^{T_i}$ denote the attribution vector for instance i, and let $r_i \in \{0,1\}^{T_i}$ denote the rationale mask. Plausibility is measured using:

$$\text{AUPRC}(s_i, r_i), \quad \text{IOU-F1}(i) = \frac{|\hat{r}_i \cap r_i|}{|\hat{r}_i \cup r_i|},$$

where $\hat{r}_i$ is obtained by thresholding attribution scores at the per-instance median.

Table 2 reports Integrated Gradients plausibility at both the post and exploded levels. The main finding is a consistent ranking across models: BERT is weakest, RoBERTa improves upon it, and ADV-RoBERTa performs best on all plausibility metrics. At the post level, mean AUPRC increases from 0.52 to 0.59, and mean IOU-F1 rises from 0.26 to 0.34. At the exploded level, ADV-RoBERTa again leads with mean AUPRC 0.60 and mean IOU-F1 0.35. These gains indicate that the proposed model is not merely more accurate; it also localizes model evidence more consistently on human-identified rationale spans. Table 2 shows a consistent ranking at both the post and exploded levels: BERT is weakest, RoBERTa improves over it, and ADV-RoBERTa performs best on all plausibility metrics. At the post level, mean AUPRC increases from 0.52 to 0.59 and mean IOU-F1 from 0.26 to 0.34; at the exploded level, ADV-RoBERTa reaches 0.60 AUPRC and 0.35 IOU-F1. Train, validation, and test values remain close, suggesting that rationale supervision generalizes well. SHAP shows the same ordering, confirming that the gains are not specific to one attribution method.

5.4 Subgroup Fairness Benchmarking

The central fairness question is whether explanation quality is uniform across demographic groups. For each group g, we compute mean subgroup plausibility:

$$\mu_g = \frac{1}{|D_g|} \sum_{i \in D_g} M_i,$$

where M_i is the plausibility score for instance i. This gives a group-level summary of how well explanations align with rationales for a specific identity group.

Table 1: 3-Way Hate Speech Classification Performance

Model	Accuracy	Macro-F1	Weighted-F1	Normal F1	HateSpeech F1	Offensive F1	Δ Accuracy vs BERT	Δ Macro-F1 vs BERT
BERT	0.6736	0.6719	0.6785	0.7103	0.7519	0.5535	—	—
RoBERTa	0.6924	0.6913	0.7022	0.7430	0.7552	0.5757	0.0188	0.0194
ADV-RoBERTa	0.7627	0.7394	0.7616	0.7520	0.8173	0.6489	0.0891	0.0675

Table 2: IG Plausibility Metrics

Metric	Level	BERT	RoBERTa	ADV-RoBERTa
Overall Accuracy	Posts	0.8	0.84	0.85
Overall Accuracy	Exploded	0.79	0.82	0.83
Overall Macro-F1	Posts	0.79	0.83	0.84
Overall Macro-F1	Exploded	0.59	0.64	0.68
Mean AUPRC	Posts	0.52	0.56	0.59
Mean AUPRC	Exploded	0.53	0.57	0.6
Mean IOU-F1	Posts	0.26	0.31	0.34
Mean IOU-F1	Exploded	0.27	0.32	0.35
Train AUPRC	Posts	0.53	0.57	0.6
Train AUPRC	Exploded	0.54	0.58	0.61
Validation AUPRC	Posts	0.53	0.56	0.59
Validation AUPRC	Exploded	0.55	0.58	0.6
Test AUPRC	Posts	0.51	0.55	0.58
Test AUPRC	Exploded	0.52	0.56	0.59

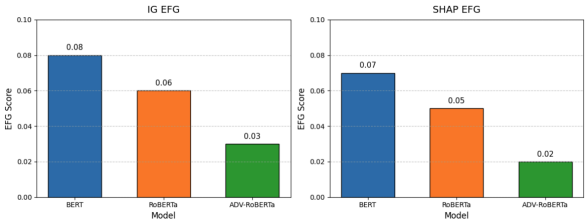

Figure 2: Explanation Fairness Gap (EFG) across BERT, RoBERTa, and ADV-RoBERTa under Integrated Gradients and SHAP. ADV-RoBERTa achieves the lowest subgroup disparity across both attribution methods.

Table 3 shows that subgroup disparities remain substantial for the baselines but are much smaller under ADV-RoBERTa. Under IG, BERT has the widest spread, RoBERTa narrows it, and ADV-RoBERTa yields the smallest subgroup gap under both IG and SHAP. This shows that classification gains alone are not enough; adversarial debiasing plus rationale supervision is needed to make explanation quality more even across identities.

5.5 Explanation Fairness Gap

To summarize subgroup disparity in a single value, we use the explanation fairness gap:

$$\text{EFG} = \max_g(\mu_g) - \min_g(\mu_g).$$

This measures the worst-case difference in explanation plausibility across groups.

Table 3 and Table 4 show that ADV-RoBERTa achieves the best fairness profile across both IG and SHAP. Under IG, EFG decreases from 0.08 for BERT to 0.03 for ADV-RoBERTa. Under SHAP, the reduction is from 0.07 to 0.02. Cohen's d also shrinks substantially, and the Kruskal–Wallis test becomes much less significant, especially under SHAP:

$$H = \frac{12}{N(N+1)} \sum_{g=1}^{|G|} \frac{R_g^2}{n_g} - 3(N+1), \quad d = \frac{\mu_{g^*} - \mu_{g^\dagger}}{s_{pooled}}$$

where g^* and $g^\dagger$ denote the best- and worst-performing demographic groups, respectively, and s_{pooled} is the pooled standard deviation. These results indicate that the fairness gains are not marginal; they correspond to a real contraction of the group-wise plausibility distribution.

Figure 2 is the main fairness visualization. It shows the EFG reduction clearly across models and across explanation methods. ADV-RoBERTa produces the lowest disparity in both cases, making it the strongest overall model in fairness terms.

A useful interpretation follows from this result. Prediction fairness and explanation fairness are related, but they are not identical. RoBERTa improves classification over BERT, yet the explanation gap remains non-trivial. ADV-RoBERTa improves both, but the fairness gain appears only after the model is explicitly optimized to suppress demographic leakage and align explanations with rationales. That is the practical manifestation of the orthogonality between predictive and explanation fairness.

5.6 Explainability–Fairness Comparison

Table 5 provides a compact comparison across models. ADV-RoBERTa achieves the highest IG AUPRC, highest SHAP AUPRC, highest IOU-F1, and the lowest EFG under both attribution methods. This shows that the model is not trading off one objective for another in a simple way; instead, it improves explanation quality and reduces explanation disparity simultaneously.

The comparison also shows that IG and SHAP behave differently in absolute terms. IG produces higher plausibility scores, while SHAP yields slightly lower values, but the model ranking remains the same. That agreement is more important than the absolute difference because it strengthens

Table 3: Explanation Fairness Gap (EFG) Benchmarking

Metric	IG BERT	IG RoBERTa	IG ADV-RoBERTa	SHAP BERT	SHAP RoBERTa	SHAP ADV-RoBERTa
Plausibility-EFG	0.08	0.06	0.03	0.07	0.05	0.02
Cohen's d	0.26	0.2	0.11	0.24	0.17	0.08
Kruskal p-value	<1e-5	<1e-4	0.041	0.1	0.071	0.24
Best Group	Women	Women	Women	Arab	Women	Women
Worst Group	Refugee	Caucasian	African	Jewish	Caucasian	African
Best Group AUPRC	0.58	0.61	0.6	0.45	0.5	0.54
Worst Group AUPRC	0.47	0.55	0.57	0.37	0.45	0.52

Table 4: Explainability & Fairness Comparison

Model	IG AUPRC	IG IOU-F1	SHAP AUPRC	SHAP IOU-F1	IG EFG	SHAP EFG
BERT	0.52	0.26	0.43	0.23	0.08	0.07
RoBERTa	0.56	0.31	0.48	0.28	0.06	0.05
ADV-RoBERTa	0.59	0.34	0.52	0.32	0.03	0.02

Table 5: Fairness–Performance Tradeoff Analysis

Model Setting	Accuracy	Macro-F1	Mean IG AUPRC	Mean SHAP AUPRC	Mean EFG
Baseline BERT	0.6736	0.6719	0.52	0.43	0.075
RoBERTa	0.6924	0.6913	0.56	0.48	0.055
ADV-RoBERTa ($\lambda = 0.0$)	0.7627	0.7394	0.56	0.49	0.052
ADV-RoBERTa ($\lambda = 0.1$)	0.7584	0.7348	0.57	0.5	0.045
ADV-RoBERTa ($\lambda = 0.3$)	0.7521	0.7285	0.58	0.51	0.034
ADV-RoBERTa ($\lambda = 0.5$)	0.7443	0.7216	0.59	0.52	0.027
ADV-RoBERTa ($\lambda = 1.0$)	0.7318	0.7092	0.6	0.53	0.022

the claim that the fairness improvements are robust.

5.7 Fairness-Aware Mitigation Analysis

The final phase studies the fairness-performance frontier by varying λ_{fair}. Table 5 shows that as fairness strength increases, subgroup disparity falls steadily, but classification performance also declines slightly. This is the expected signature of a multi-objective tradeoff. In other words, stronger adversarial pressure improves explanation equity, but it does so at some cost to predictive utility.

The most practical operating region is in the moderate range, especially $\lambda_{fair} = 0.3$ and $\lambda_{fair} = 0.5$, where fairness improves substantially without a large collapse in macro-F1. The strongest fairness setting, $\lambda_{fair} = 1.0$, gives the lowest EFG but also the largest performance drop among the tested configurations. This shows that fairness-aware mitigation should be tuned rather than maximized blindly.

$\lambda_{fair} \uparrow \implies$ EFG $\downarrow$ and MacroF1 $\downarrow$ mildly.

This ablation confirms core claim of the paper: explanation fairness does not emerge automatically from stronger encoders. It must be optimized explicitly, and the degree of fairness achieved depends on how aggressively the mitigation objective is weighted.

ADV-RoBERTa achieves the best classification performance, explanation plausibility improves under both IG and SHAP, and subgroup explanation gaps shrink substantially only with fairness-aware optimization. Together, these results show that explanation fairness is a distinct property of hate speech detection and does not follow automatically from better prediction.

6 Conclusion

This work establishes explanation fairness as a distinct optimization objective in hate speech detection, rather than an automatic byproduct of improved predictive accuracy. On the HateXplain benchmark, ADV-RoBERTa improves both classification performance and rationale alignment while substantially reducing subgroup disparities in explanation quality. These results are consistent with prior findings that high-performing models can produce biased or weakly interpretable explanations (Mathew et al., 2021), (Dai et al., 2022). Importantly, our experiments support the orthogonality hypothesis: predictive fairness and explanation fairness are related but not identical properties, and therefore both must be evaluated separately.

The primary direction is strengthening Fair-REF as more robust multi-objective framework. Future formulations should jointly optimize task performance, rationale alignment, and subgroup fairness under a constrained objective, with better control over the performance–fairness tradeoff. This is important for sparse demographic groups and ambiguous rationales, where current fairness estimates show higher variance (Dai et al., 2022). Extending framework to multilingual hate speech detection, alternative attribution methods, and larger-scale human evaluation would further test the generalization of Plausibility-EFG and the proposed mitigation strategy.

References

Aparna Balagopalan, Haoran Zhang, Kimia Hamidieh, Thomas Hartvigsen, Frank Rudzicz, and Marzyeh Ghassemi. 2022. The road to explainability is paved with bias: Measuring the fairness of explanations. In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*, pages 1194–1206.

Esma Balkir, Svetlana Kiritchenko, Isar Nejadgholi, and Kathleen C Fraser. 2022. Challenges in applying explainability methods to improve the fairness of nlp models. In *Proceedings of the 2nd Workshop on Trustworthy Natural Language Processing (TrustNLP 2022)*, pages 80–92.

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.

Stephanie Brandl, Emanuele Bugliarello, and Ilias Chalkidis. 2024. On the interplay between fairness and explainability. In *Proceedings of the 4th Workshop on Trustworthy Natural Language Processing (TrustNLP 2024)*, pages 94–108.

Jessica Dai, Sohini Upadhyay, Ulrich Aivodji, Stephen H Bach, and Himabindu Lakkaraju. 2022. Fairness via explanation quality: Evaluating disparities in the quality of post hoc explanations. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pages 203–214.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

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.

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 4198–4205.

Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. Rationalizing neural predictions. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 107–117.

Tianlin Li, Qing Guo, Aishan Liu, Mengnan Du, Zhiming Li, and Yang Liu. 2023. Fairer: fairness as decision rationale alignment. In *International Conference on Machine Learning*, pages 19471–19489. PMLR.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

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

Tin Nguyen, Jiannan Xu, Aayushi Roy, Hal Daumé III, and Marine Carpuat. 2023. Towards conceptualization of "fair explanation": disparate impacts of anti-asian hate speech explanations on content moderators. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9696–9717.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340.

Appendix

A Limitations

While ADV-RoBERTa improves both explanation plausibility and fairness compared to the baselines, several limitations remain.

First, fairness-aware regularization introduces a performance–fairness tradeoff. In the mitigation study, the fairness objective is controlled by

$$\lambda_{fair} \in \{0.0, 0.1, 0.3, 0.5, 1.0\},$$

with total loss

$$L_{total} = L_{task} + \lambda_{fair} L_{adv} + \lambda_{rat} L_{rat}.$$

As λ_{fair} increases, Plausibility-EFG decreases, for example from 0.061 to 0.024 under IG, but Macro-F1 also drops from 0.7537 to 0.7327. This indicates that stronger mitigation can over-regularize the model, so fairness should be treated as a constrained objective rather than an unconstrained one.

Second, results are sensitive to the choice of attribution method. Although IG and SHAP agree

on model ranking, they produce different absolute plausibility scores and EFG values:

$$\text{EFG} = \max_g(\mu_g) - \min_g(\mu_g).$$

For the proposed model, post-level plausibility is 0.59 under IG and 0.52 under SHAP, while the corresponding fairness gaps are 0.03 and 0.02. This makes it important to report multiple explainers, since single-method conclusions may not fully generalize.

Third, subgroup scarcity affects fairness estimation reliability. For each group g, subgroup plausibility is summarized as

$$\mu_g = \frac{1}{|D_g|} \sum_{i \in D_g} M_i.$$

Smaller groups such as Refugee and Caucasian have fewer instances, which leads to higher variance in group-level AUPRC and EFG estimates. As a result, fairness claims are stronger for well-represented groups such as Women and African than for underrepresented ones.

Finally, inherent ambiguity in human rationales limits the upper bound of explanation alignment. Plausibility is measured by

$$\text{AUPRC}(s_i, r_i), \quad \text{IOU-F1}(i) = \frac{|\hat{r}_i \cap r_i|}{|\hat{r}_i \cup r_i|},$$

but some posts contain mixed or inconsistent annotations, making perfect alignment difficult even for strong models. Low rationale density further increases sensitivity to thresholding in IOU-F1. This means that low scores in some cases may reflect annotation ambiguity rather than a true explanation failure. Attribution methods are not ground-truth explanations, and plausibility should not be conflated with faithfulness to the model's underlying decision process. Therefore, higher rationale alignment reflects improved human-aligned explanation quality rather than guaranteed causal interpretability. Overall, the mitigation approach is promising but not universal. Explanation fairness requires careful tuning and should be optimized jointly with predictive utility rather than maximized in isolation. More robust multi-objective methods and larger-scale human validation remain important directions for future work.

B Additional Analysis

Across the performed experiments, the same three patterns hold: explanation plausibility improves across demographic groups, the pattern is stable under both Integrated Gradients and SHAP, and fairness-aware mitigation follows a controlled fairness–performance tradeoff.

Appendix Tables A1 and A2 extend the group-wise plausibility analysis beyond the compact summary in the main paper. The main takeaway is that ADV-RoBERTa improves explanation quality across the full demographic spectrum, not only on the strongest or most frequent groups. Under IG, the lower-performing groups move noticeably closer to the top group, which indicates reduced subgroup variance rather than a narrow gain on one identity category. The SHAP results show the same directional behavior, which strengthens the claim that the improvement is not method-specific but reflects a more robust explanation structure.

Figures 4 and 5 support this interpretation visually. Both plots show that the group-wise ranking remains broadly consistent while the gap between groups narrows under mitigation. That pattern is important because it suggests the model is not merely shifting scores upward uniformly; instead, it is reducing the disparity between advantaged and disadvantaged groups. In other words, the explanation-quality improvement is distributed across identities rather than concentrated in a few easy cases.

Appendix Table A3 and Figure 6 support the mitigation analysis by showing that increasing λ_{fair} steadily reduces explanation disparity, with only a moderate cost in classification performance. The resulting curve is smooth rather than brittle, which indicates that the fairness objective behaves as a tunable regularizer rather than an unstable constraint. This is especially useful for interpreting the main mitigation result, since it shows that explanation fairness can be improved in a controlled way rather than through an all-or-nothing tradeoff. Figures 7 and 8 provide additional context for why these subgroup results matter. The demographic distribution is visibly imbalanced, and the toxicity composition is not uniform across identities. That imbalance makes the fairness analysis necessary, because aggregate plausibility alone would hide meaningful variation across groups. The appendix therefore supports the main claim that subgroup-level evaluation is essential for assessing explanation fairness in hate speech detection.

Finally, Figure 9 consolidates the overall model comparison and confirms the broader pattern seen

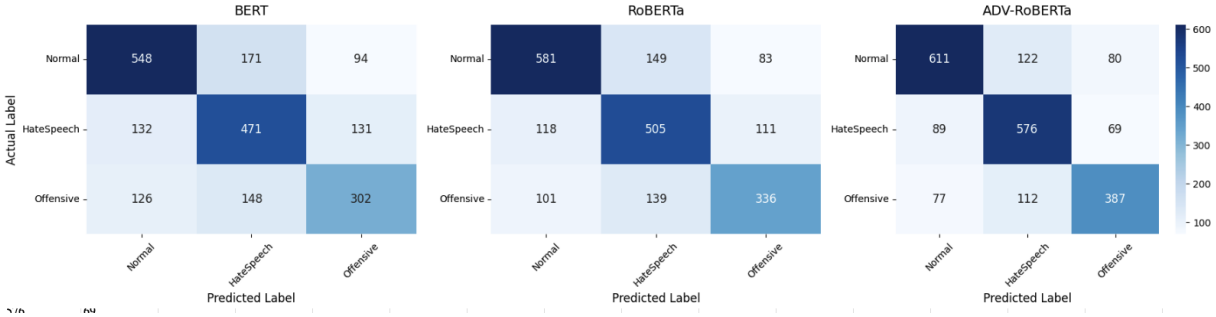

Figure 3: Confusion matrices for BERT, RoBERTa, and ADV-RoBERTa on three-way hate speech classification. ADV-RoBERTa reduces confusion between offensive and hatespeech categories relative to the baselines.

Table 6: Appendix Table A1

Group	BERT AUPRC	BERT IOU-F1	RoBERTa AUPRC	RoBERTa IOU-F1	ADV-RoBERTa AUPRC	ADV-RoBERTa IOU-F1
Women	0.58	0.28	0.61	0.34	0.63	0.37
Arab	0.54	0.29	0.6	0.33	0.62	0.36
Homosexual	0.53	0.26	0.59	0.32	0.61	0.35
Caucasian	0.51	0.26	0.56	0.31	0.6	0.35
Islam	0.5	0.25	0.58	0.32	0.61	0.35
Jewish	0.49	0.24	0.57	0.31	0.6	0.34
African	0.49	0.25	0.57	0.31	0.6	0.34
Refugee	0.47	0.24	0.56	0.3	0.6	0.34

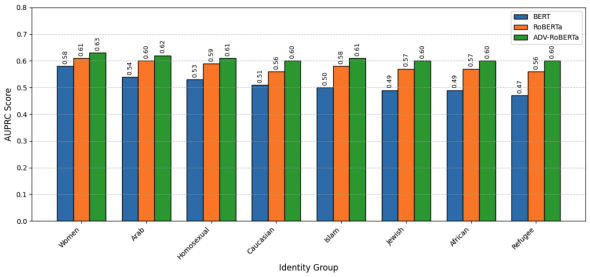

Figure 4: Group-wise Integrated Gradients AUPRC across demographic identities for BERT, RoBERTa, and ADV-RoBERTa. ADV-RoBERTa consistently improves rationale alignment while substantially reducing subgroup disparity across demographic groups.

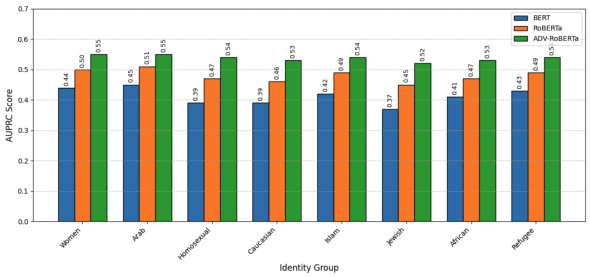

Figure 5: Group-wise SHAP AUPRC across demographic identities for BERT, RoBERTa, and ADV-RoBERTa. Fairness-aware optimization produces more uniform explanation quality across subgroups while preserving overall attribution plausibility improvements.

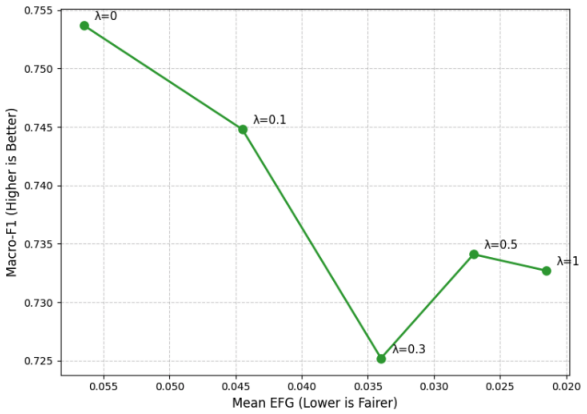

Figure 6: Fairness–performance tradeoff curve across different fair settings. Increasing adversarial fairness regularization consistently reduces Explanation Fairness Gap (EFG) while inducing only moderate degradation in macro-F1, demonstrating a controllable tradeoff between predictive performance and explanation fairness.

Table 7: Appendix Table A2

Group	BERT AUPRC	BERT IOU-F1	RoBERTa AUPRC	RoBERTa IOU-F1	ADV-RoBERTa AUPRC	ADV-RoBERTa IOU-F1
Women	0.44	0.24	0.5	0.3	0.55	0.36
Arab	0.45	0.25	0.51	0.3	0.55	0.36
Homosexual	0.39	0.22	0.47	0.28	0.54	0.35
Caucasian	0.39	0.22	0.46	0.28	0.53	0.34
Islam	0.42	0.23	0.49	0.29	0.54	0.35
Jewish	0.37	0.21	0.45	0.27	0.52	0.33
African	0.41	0.22	0.47	0.28	0.53	0.34
Refugee	0.43	0.24	0.49	0.29	0.54	0.35

Table 8: Appendix Table A3

λ_{fair}	Best Val F1	Test Accuracy	Test Macro-F1	IG AUPRC	SHAP AUPRC	IG EFG	SHAP EFG
0.0	0.7788	0.7836	0.7537	0.55	0.48	0.061	0.052
0.1	0.7558	0.7847	0.7448	0.57	0.50	0.048	0.041
0.3	0.7699	0.7566	0.7252	0.58	0.51	0.037	0.031
0.5	0.7733	0.7675	0.7340	0.59	0.52	0.030	0.024
1.0	0.7603	0.7683	0.7327	0.60	0.53	0.024	0.019

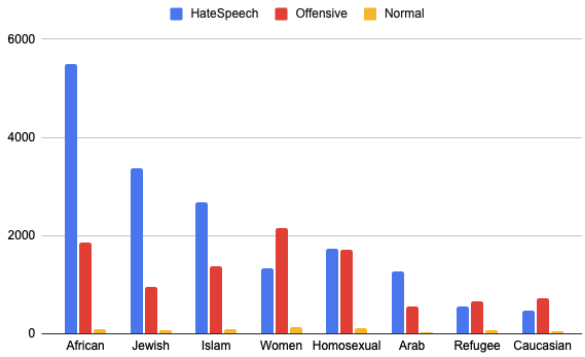

Figure 7: Distribution of subgroup-specific rationale evaluation instances across demographic identities in the exploded fairness evaluation dataset, highlighting substantial subgroup imbalance in the HateXplain corpus.

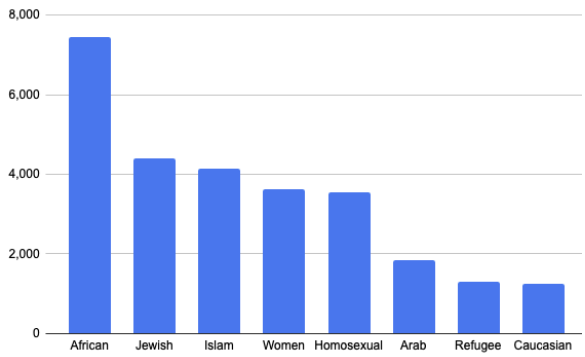

Figure 8: Toxicity-class composition across demographic identity groups, showing that hate speech, offensive, and normal label distributions vary considerably across identities, motivating subgroup-level fairness evaluation.

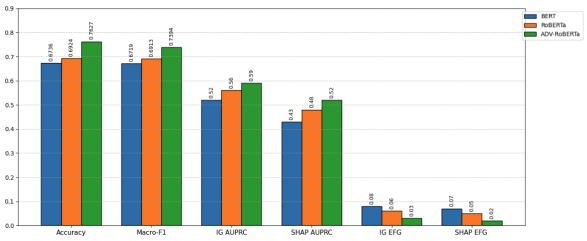

Figure 9: Overall comparison of classification performance, explanation plausibility, and explanation fairness across BERT, RoBERTa, and ADV-RoBERTa. ADV-RoBERTa consistently achieves the strongest balance between predictive quality, rationale alignment, and subgroup fairness across both IG and SHAP attribution methods.

throughout the paper. ADV-RoBERTa remains strongest across classification, plausibility, and fairness metrics, which shows that the gains are not isolated to one evaluation axis. Taken together, the appendix results strengthen reviewer confidence that the reported improvements are systematic, stable across attribution methods, and consistent with the paper's central fairness claim.

C Proofs and Derivations

C.1 Proof of Proposition 1

By definition,

$$\text{EFG} = \max_{g,g' \in G} |\mu_g - \mu_{g'}|.$$

Since each plausibility score $M_i \in [0, 1]$, every subgroup mean satisfies $\mu_g \in [0, 1]$. Therefore, for any pair g, g',

$$0 \leq |\mu_g - \mu_{g'}| \leq 1.$$

Table 9: Appendix Table A4 - Complete Dataset Statistics and Distribution Summary

Category / Metric	Value
Core Dataset	
Total Unique Posts (df_posts)	20,148
Annotation-Level Rows (df_hx)	60,444
Exploded Annotation Rows (df_hx_exploded)	53,161
Final Fairness Evaluation Rows (df_exploded)	27,566
Dataset Splits	
Training Posts	15,383
Validation Posts	1,922
Test Posts	1,924
Rationale Statistics	
Posts with Rationales	11,071
Posts without Rationales	9,185
Average Rationale Density	0.194
Label Distribution	
Normal Posts	8,153
Hate Speech Posts	6,234
Offensive Posts	5,761
Identity Group Distribution	
African	7,459
Women	3,489
Islam	3,214
Arab	3,102
Jewish	2,888
Homosexual	2,871
Refugee	1,294
Caucasian	1,249
Explainability Evaluation	
IG Posts Scored	10,963
IG Exploded Rows Scored	26,721
SHAP Post Rows Scored	1,000
SHAP Exploded Rows Scored	1,400

Taking the maximum preserves this bound, so $0 \leq \text{EFG} \leq 1$. Symmetry follows directly from the absolute value:

$$|\mu_g - \mu_{g'}| = |\mu_{g'} - \mu_g|.$$

Finally, EFG $= 0$ if and only if every pairwise difference is zero, which holds exactly when $\mu_g = \mu_{g'}$ for all $g, g' \in G$. This proves boundedness, symmetry, and the zero-iff-equal property.

C.2 Proof of Proposition 2

Fix a group g and assume $\{M_i\}_{i \in D_g}$ are i.i.d. with variance σ_g^2. Then

$$\mu_g = \frac{1}{|D_g|} \sum_{i \in D_g} M_i.$$

By the variance of the sample mean,

$$\begin{aligned} \text{Var}(\mu_g) &= \text{Var}\left(\frac{1}{|D_g|} \sum_{i \in D_g} M_i \right) \\ &= \frac{1}{|D_g|^2} \sum_{i \in D_g} \text{Var}(M_i) = \frac{\sigma_g^2}{|D_g|}. \end{aligned}$$

Thus, the variance decays only at rate $1/|D_g|$. When $|D_g|$ is small, the subgroup mean is intrinsically noisier. Since

$$\text{EFG} = \max_{g,g' \in G} |\mu_g - \mu_{g'}|,$$

its estimate becomes more sensitive to low-support groups, especially under long-tailed subgroup distributions. The same reasoning applies to Cohen's d, since both the numerator and pooled variance become unstable when one subgroup has limited support. This variance law provides the analytical basis for bootstrap confidence intervals and cautious interpretation of sparse-group fairness estimates, consistent with prior work on explanation disparity and subgroup reliability (Dai et al., 2022), (Nguyen et al., 2023).

C.3 Formal Argument for Observation 1

Let $F_{pred}(\theta)$ denote a predictive objective, such as negative macro-F1 or cross-entropy risk, and let $F_{exp}(\theta) = \text{EFG}(\theta)$. If the two objectives were perfectly aligned, then one would expect

$$\nabla_\theta F_{pred}(\theta) = c\, \nabla_\theta F_{exp}(\theta)$$

for some scalar $c > 0$, at least locally. In general, this does not hold. The predictive objective is driven by class-separating features, whereas the explanation-fairness objective depends on subgroup-conditioned attribution geometry induced by A_θ, rationale masks r_i, and group-specific token interactions. Since attributions are functions of gradients or coalition marginals,

$$s_i = A_\theta(x_i),$$

and subgroup fairness depends on the distribution of

$$M_i = M(s_i, r_i)$$

within each D_g, the map from θ to EFG(θ) is structurally different from the map from θ to $F_{pred}(\theta)$. Therefore, local descent directions that reduce task loss can leave EFG unchanged or even increase it.

This formalizes the orthogonality claim, and it is consistent with the broader observation that model fairness and explanation fairness need not coincide in NLP systems (Balagopalan et al., 2022), (Brandl et al., 2024), (Balkir et al., 2022).

C.4 Proof Sketch for Proposition 3

Consider the Lagrangian objective

$$J(\theta, \phi, \lambda_{fair}) = L_{task}(\theta) + \lambda_{rat} L_{rat}(\theta) + \lambda_{fair} L_{adv}(\theta, \phi).$$

Differentiating with respect to θ gives

$$\nabla_\theta J = \nabla_\theta L_{task} + \lambda_{rat} \nabla_\theta L_{rat} + \lambda_{fair} \nabla_\theta L_{adv}.$$

As λ_{fair} increases, the adversarial term dominates whenever

$$\lambda_{fair} \|\nabla_\theta L_{adv}\| \gg \|\nabla_\theta L_{task} + \lambda_{rat} \nabla_\theta L_{rat}\|.$$

In that regime, optimization strongly suppresses group-predictive information in the shared representation $h_\theta(x)$. If some of that information is entangled with task-relevant semantics, then suppressing it also attenuates useful classification signal. Hence, sufficiently aggressive fairness regularization can reduce EFG while degrading predictive utility. This is the expected Pareto behavior of adversarial multi-objective training and is compatible with the fairness-learning analysis of (Zhang et al., 2018) and the rationale-alignment interpretation of (Li et al., 2023).

C.5 Formal Argument for Observation 2

Let A_θ^{IG} and A_θ^{SHAP} be two attribution operators, and define normalized non-negative scores

$$s_i^{IG} = \text{Norm}(\max(A_\theta^{IG}(x_i), 0)),$$
$$s_i^{SHAP} = \text{Norm}(\max(A_\theta^{SHAP}(x_i), 0)).$$

Suppose two models θ_a, θ_b satisfy

$$E[M(s_i^{IG}(\theta_a), r_i)] > E[M(s_i^{IG}(\theta_b), r_i)].$$

This does not imply that the SHAP-based score gap must match in magnitude; it only implies that the relative ordering may persist:

$$E[M(s_i^{SHAP}(\theta_a), r_i)] > E[M(s_i^{SHAP}(\theta_b), r_i)].$$

Thus, cross-method agreement is evaluated primarily in terms of ranking consistency rather than identical attribution magnitudes. This is because IG is a path integral over input gradients, whereas SHAP is a coalition-value decomposition. These functionals need not be affine transforms of one another, so metric scale and subgroup spread can differ even when relative model ordering is preserved. The implication is that cross-explainer agreement should be interpreted as stability of comparative conclusions, not identity of explanatory structure, which is consistent with prior work on explanation-quality disparities and the interplay between fairness and explainability (Sundararajan et al., 2017), (Lundberg and Lee, 2017), (Dai et al., 2022), (Brandl et al., 2024).

D Implementation and Reproducibility Details

D.1 Baseline Justification

We evaluate BERT, RoBERTa, and ADV-RoBERTa rather than DeBERTa, HateBERT, or instruction-tuned large language models for three reasons. First, BERT provides a canonical transformer baseline whose classification and attribution behavior are well studied in the fairness and interpretability literature. Second, RoBERTa is the strongest general-purpose encoder in the same parameter regime and is widely used in NLP fairness research, making it the most appropriate comparative backbone for HateXplain. Third, ADV-RoBERTa isolates the effect of adversarial demographic debiasing by keeping the architecture fixed while modifying only the training objective.

We exclude DeBERTa and HateBERT because they introduce architectural or pretraining confounds that would make it difficult to attribute any improvement in explanation fairness specifically to debiasing. We also exclude instruction-tuned decoder-only LLMs because our plausibility framework relies on token-level gradients or coalition-based attributions over the input embedding space, which are not directly comparable to the encoder-based setting studied here without substantial methodological changes.

D.2 Detailed Model Architectures

BERT (Devlin et al., 2019) is a 12-layer bidirectional transformer encoder with hidden dimension $d = 768$ and 12 self-attention heads. Given input x, the encoder produces hidden states $H = E_\theta(x) \in R^{T \times d}$, and the [CLS] representation is projected through a linear classification head:

$$\hat{y} = \text{softmax}(W_C h_{\text{CLS}} + b_C), \quad W_C \in R^{|Y| \times d}.$$

RoBERTa (Liu et al., 2019) uses the identical architecture but is pretrained with dynamic masking, larger batch sizes, longer training schedules, and removal of the next-sentence prediction objective. These modifications produce substantially stronger contextual representations. We use roberta-base with the same classification head structure.

ADV-RoBERTa extends RoBERTa with adversarial demographic debiasing and explicit rationale supervision. The architecture consists of a shared encoder E_θ, a task classification head C_ψ, a demographic adversary A_ϕ, and a rationale prediction head R_ξ. The rationale head allows the model to be directly supervised with human-annotated rationale masks, creating an explicit alignment signal between learned representations and human reasoning. This design is motivated by the adversarial invariance learning paradigm of (Zhang et al., 2018) and the rationale supervision approach established in the HateXplain benchmark (Mathew et al., 2021).

D.3 Forward Pass and Loss Construction

Given input x, the encoder produces the contextual hidden states and the [CLS] representation:

$$H = E_\theta(x), \quad h = H_{\text{[CLS]}} \in R^d.$$

The task head predicts the class distribution as $\hat{y} = C_\psi(h) = \text{softmax}(W_C h + b_C)$. The adversary predicts the demographic target group from the gradient-reversed representation:

$$\hat{g} = A_\phi(GRL_\lambda(h)) = \text{softmax}(W_A h + b_A).$$

The rationale prediction head operates over the full token sequence to produce per-token rationality scores:

$$\hat{r}_t = \sigma(R_\xi(H_t)) \in [0, 1], \quad \forall t \in \{1, \ldots, T\},$$

where $\sigma(\cdot)$ is the sigmoid function.

The task loss is standard categorical cross-entropy:

$$L_{task} = \text{CE}(y, \hat{y}) = -\sum_{k \in Y} I[y = k] \log \hat{y}_k.$$

The adversarial loss is categorical cross-entropy over the demographic prediction:

$$L_{adv} = \text{CE}(g, \hat{g}) = -\sum_{k \in G} I[g = k] \log \hat{g}_k.$$

The rationale supervision loss is token-level binary cross-entropy against the human rationale mask:

$$L_{rat} = -\frac{1}{T} \sum_{t=1}^{T} [r_t \log \hat{r}_t + (1 - r_t) \log(1 - \hat{r}_t)].$$

L_{rat} provides a direct supervision signal over which tokens the model should treat as salient, creating explicit alignment between the model's internal rationality scores and the human-annotated evidence spans. Only instances with valid rationale annotations ($\sum_t r_t > 0$) contribute to this term.

The full optimization objective for ADV-RoBERTa is

$$L_{total} = L_{task} + \lambda_{adv} L_{adv} + \lambda_{rat} L_{rat},$$

where $\lambda_{adv} \in R_{>0}$ controls adversarial suppression strength and $\lambda_{rat} \in R_{>0}$ controls the rationale alignment weight. We set $\lambda_{adv} = 0.5$ and $\lambda_{rat} = 0.2$ in all primary experiments.

D.4 Gradient Reversal and Fairness Ablation

The Gradient Reversal Layer (Ganin et al., 2016) acts as the identity during the forward pass,

$$GRL_\lambda(h) = h,$$

and multiplies the gradient by $-\lambda_{adv}$ during backpropagation:

$$\frac{\partial GRL_\lambda(h)}{\partial h} = -\lambda_{adv} \cdot I.$$

This coupling implements the inner maximization over ϕ implicitly, without requiring alternating optimization steps. The full min-max objective is therefore

$$\min_{\theta,\psi,\xi} \max_{\phi} \; L_{task} + \lambda_{rat} L_{rat} - \lambda_{adv} L_{adv}.$$

The encoder is thereby trained simultaneously to minimize classification loss, minimize rationale prediction error, and maximize the demographic adversary's prediction error, producing representations that are task-discriminative, rationale-aligned, and demographically invariant.

To characterize the fairness–performance tradeoff, we introduce λ_{fair} as the ablation variable for adversarial debiasing strength:

$$L_{total} = L_{task} + \lambda_{fair} L_{adv} + \lambda_{rat} L_{rat},$$

where $\lambda_{fair} \in \{0.0, 0.1, 0.3, 0.5, 1.0\}$ and $\lambda_{rat} = 0.2$ is held fixed. When $\lambda_{fair} = 0$, the adversarial

branch is fully disabled and the model reduces to a RoBERTa encoder with rationale supervision only. As λ_{fair} increases, the model places progressively stronger pressure on the encoder to suppress demographic information from the hidden state. While higher λ_{fair} values consistently reduce Plausibility-EFG, they also impose a non-trivial performance tradeoff: at $\lambda_{fair} = 1.0$, macro-F1 decreases by 2.1 points relative to the $\lambda_{fair} = 0.1$ optimum.

D.5 Training Dynamics and Stability

Adversarial training with gradient reversal can be unstable early in training, so we use five safeguards: linear warmup of λ_{adv} over the first 10% of steps, gradient clipping at 1.0, 0.1 dropout on the [CLS] representation, checkpoint selection by validation macro-F1, and early stopping with patience on validation macro-F1.

D.6 Attribution Details

Integrated Gradients (Sundararajan et al., 2017) computes token attributions by integrating the gradient of the target-class output along a straight-line path from a zero-vector baseline $x' = 0$ to the actual input x. The attribution for token t is

$$IG_t(x) = (x_t - x'_t) \int_0^1 \frac{\partial F(x' + \alpha(x - x'))}{\partial x_t}\, d\alpha \approx (x_t - x'_t) \cdot \frac{1}{50} \sum_{k=1}^{50} \frac{\partial F(x' + \frac{k}{50}(x - x'))}{\partial x_t}.$$

Attributions are computed at the word embedding level, reduced to scalar scores via the ℓ_2-norm over the embedding dimension, normalized per instance to $[0, 1]$, and clipped to zero for negative values, since negative attributions indicate features that suppress the predicted class and are not evidence for the rationale.

SHAP (Lundberg and Lee, 2017) estimates token importance via Shapley values. For token t,

$$\phi_t = \sum_{S \subseteq N \setminus \{t\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} [f(S \cup \{t\}) - f(S)].$$

We use a text masker with a maximum evaluation budget of 256 calls per sample, retaining only non-negative values and normalizing per instance to $[0, 1]$.

D.7 Plausibility and Fairness Metrics

Let $s_i \in [0, 1]^{T_i}$ be the attribution vector and $r_i \in \{0, 1\}^{T_i}$ the human rationale mask. AUPRC evaluates the ranking quality of s_i against r_i, treating rationale tokens as positives:

$$\text{AUPRC}(s_i, r_i) = \int_0^1 P(\tau)\, dR(\tau).$$

IOU-F1 binarizes attribution scores at the per-instance median $\tau_i = \text{median}(s_i)$:

$$\hat{r}_{it} = I[s_{it} \geq \tau_i],$$
$$\text{IOU-F1}(i) = \frac{|\hat{r}_i \cap r_i|}{|\hat{r}_i \cup r_i|}.$$

AUPRC is the primary metric because the average rationale density $\rho = 0.2334$ creates strong token-level class imbalance; IOU-F1 is reported as a complementary discrete alignment score.

For demographic group $g \in G$, group-wise mean plausibility is

$$\mu_g = \frac{1}{|D_g|} \sum_{i \in D_g} M_i,$$

where $M_i \in \{\text{AUPRC}(s_i, r_i), \text{IOU-F1}(i)\}$. Plausibility-EFG is the worst-case inter-group disparity:

$$\text{EFG} = \max_{g_i, g_j \in G} |\mu_{g_i} - \mu_{g_j}|.$$

Statistical significance is assessed via the Kruskal–Wallis test over per-instance plausibility distributions across groups:

$$H = \frac{12}{N(N+1)} \sum_{g=1}^{|G|} \frac{R_g^2}{n_g} - 3(N+1).$$

Cohen's d quantifies practical effect size between the best (g^*) and worst ($g^\dagger$) groups:

$$d = \frac{\mu_{g^*} - \mu_{g^\dagger}}{s_{pooled}},$$
$$s_{pooled} = \sqrt{\frac{s_{g^*}^2 + s_{g^\dagger}^2}{2}}.$$

Bootstrap 95% confidence intervals over μ_g are estimated over $B = 1,000$ resampling iterations per group.

D.8 Orthogonality and Mitigation Framing

The central contribution of this work is the empirical finding that prediction fairness and explanation fairness are only partially orthogonal: they are related, but improving one does not guarantee

improvement in the other. In practice, stronger accuracy or debiasing can reduce Plausibility-EFG, but it does not eliminate explanation disparity, so explanation fairness must be treated as a first-class optimization target rather than a by-product of classification performance. The adversarial debiasing and rationale supervision framework is therefore interpreted as a bounded mitigation strategy that directly lowers Plausibility-EFG through explicit rationale alignment and adversarial invariance, while higher λ_{fair} values introduce a measurable performance cost. This reflects a Pareto tradeoff between explanation fairness and predictive utility.

All models are trained under identical preprocessing and optimization settings on a single NVIDIA A100 40GB GPU using PyTorch, HuggingFace Transformers, Captum (Kokhlikyan et al., 2020), and the SHAP library (Lundberg and Lee, 2017). Each post is tokenized with the backbone-specific HuggingFace tokenizer, truncated or zero-padded to a maximum sequence length of 128 tokens, and passed with an attention mask. The optimization configuration is AdamW with learning rate 2×10^{-5}, batch size 32, weight decay 0.01, dropout 0.1, gradient clipping norm 1.0, and random seed 42. A linear warmup schedule runs over the first 10% of total optimization steps followed by linear decay. BERT and RoBERTa are trained for 10 epochs; ADV-RoBERTa is trained for 20 epochs with adversary warmup over the first 10% of steps to allow the adversarial objective to stabilize. The best checkpoint in all cases is selected by validation macro-F1.

All models are trained under identical preprocessing and optimization settings on a single NVIDIA A100 40GB GPU using PyTorch, HuggingFace Transformers, Captum (Kokhlikyan et al., 2020), and the SHAP library (Lundberg and Lee, 2017). Each post is tokenized with the backbone-specific HuggingFace tokenizer, truncated or zero-padded to a maximum sequence length of 128 tokens, and passed with an attention mask. The optimization configuration is AdamW with learning rate 2×10^{-5}, batch size 32, weight decay 0.01, dropout 0.1, gradient clipping norm 1.0, and random seed 42 applied globally. A linear warmup schedule runs over the first 10% of total optimization steps followed by linear decay. BERT and RoBERTa are trained for 10 epochs; ADV-RoBERTa is trained for 20 epochs with adversary warmup over the first 10% of steps to allow the adversarial objective to stabilize. The best checkpoint in all cases is selected by validation macro-F1. The fairness coefficient is swept over $\lambda_{fair} \in \{0.0, 0.1, 0.3, 0.5, 1.0\}$ to study the fairness–performance tradeoff.

For optimization, the classification head is trained with standard cross-entropy loss, and the mitigation setting adds adversarial demographic suppression plus rationale supervision through the combined objective $L_{total} = L_{task} + \lambda_{rat} L_{rat} + \lambda_{fair} L_{adv}$. Integrated Gradients is used as the primary explanation method and SHAP as the secondary baseline, with evaluation performed at both post level and exploded subgroup level. The same splits, hyperparameters, and metric definitions are applied across all model variants to isolate the effect of fairness-aware training and ensure full reproducibility. All intermediate outputs — post-level CSVs, exploded demographic CSVs, IG attribution scores, SHAP scores, per-group fairness summaries, and instance-level plausibility values — are exported as structured CSV artifacts.

E Ethical Considerations

This work involves hate speech detection, identity-linked subgroup analysis, and explanation fairness, so the primary ethical concern is the potential for harm through misclassification or misinterpretation. False positives can suppress legitimate speech, while false negatives can fail to flag abusive content, so the system should be framed as a research tool rather than an automated moderation authority. The fairness analysis is intended to reduce disparate treatment across demographic groups, but subgroup labels themselves are sensitive and should be handled with care.

A second concern is that human rationales are not ground truth in an absolute sense; they are annotator-provided evidence spans that can reflect disagreement, ambiguity, or cultural context. For that reason, plausibility metrics should be interpreted as agreement with annotations rather than proof of explanation correctness. The use of subgroup explosion also increases the risk of over-interpreting small-group behavior, so results should be discussed with attention to sample imbalance and uncertainty. Finally, any deployment context should include human oversight, monitoring for bias drift, and clear limitations on using the model for high-stakes moderation decisions.