

No Better Than Random: Evaluating LIME-Guided Saliency Penalisation for Fine-Grained Sexism Classification

Olarenwaju Adisa and Chidinma Ndubuisi

Department of Computing, College of Sciences
Afe Babalola University, Ado-Ekiti, Ekiti State, Nigeria
adisao@abuad.edu.ng
ndubuisichidinma@student.abuad.edu.ng

Abstract

Fine-grained sexism classifiers are known to rely on surface gender terms, and explainability methods are increasingly proposed not merely to diagnose this reliance but to correct it during training. We test that proposition directly. Using LIME, we identify the tokens a DeBERTa-v3-large classifier relies on most when it misclassifies EDOS Task C development posts, and penalise those tokens’ hidden-state magnitudes during continued fine-tuning. Across three random seeds the method does not improve macro-F1: the best penalised setting reaches 0.477 ± 0.003 against a 0.468 ± 0.004 baseline—within seed variance—and continued training with the penalty disabled scores highest of all (0.483). A controlled ablation isolates why. Suppressing only the gendered subset of flagged tokens, only the function-word subset, and a random frequency-matched set all leave test macro-F1 within seed-level variation of the baseline (0.466, 0.467, and 0.461); critically, LIME-guided suppression performs *no better than random suppression*, indicating the attribution signal carries no usable corrective information. The flagged tokens are themselves predominantly gendered terms and slurs—65–80% of LIME’s top tokens across seeds—that is, the genuine evidence for sexism rather than spurious cues. We further find that attribution methods disagree sharply on which tokens are most influential (LIME 65%, SHAP 30%, Integrated Gradients 35% gendered), cautioning against single-attributor conclusions. We conclude that attribution magnitude is an unreliable criterion for token suppression in abuse detection, and that explanation-guided correction cannot succeed where the most salient features are the task’s real signal.

1 Introduction

Automated detection of online sexism has moved from coarse binary judgements toward fine-grained taxonomies that distinguish, for example, threats of

harm from gendered stereotyping or condescending advice. The Explainable Detection of Online Sexism (EDOS) benchmark formalised this with a three-level hierarchy whose most demanding tier, Task C, assigns one of eleven specific categories to each sexist post (Kirk et al., 2023). Task C remains difficult—reported macro-F1 spans roughly 0.38–0.54—both because several categories are rare and because participating systems are documented to lean on surface cues, particularly gendered keywords, rather than on a genuine reading of intent (Hassan et al., 2023; Roy and Shrivastava, 2023; Hemati et al., 2023).

A natural idea follows. If a post-hoc explainer can reveal which tokens drive a model’s errors, that signal might be fed back into training to steer the model away from those tokens—turning explanation from a diagnostic into a corrective. This “explanation-as-supervision” framing underlies a line of work on constraining model explanations during training (Ross et al., 2017), and is especially attractive for abuse detection, where reliance on identity terms is a known failure mode (Dixon et al., 2018; Sap et al., 2019). Yet no EDOS system has used attribution-method output as a training signal, and the premise has not been tested where it matters most: when the most salient tokens are not spurious correlates but the task’s genuine evidence.

This paper tests the proposition directly and reports a carefully controlled negative answer. We take a strong DeBERTa-v3-large classifier, use LIME to identify the tokens it most relies on when it errs on the development set, and continue training with a penalty that suppresses those tokens’ hidden-state magnitudes. We evaluate the penalty across four strengths and three seeds, against a naive masking control, a token-type ablation, and—decisively—a random frequency-matched control. Our contributions are:

- We introduce **LIME-guided saliency**

penalisation, which operationalises a non-differentiable post-hoc attributor as a representation-level, gradient-bearing training signal at misleading-token positions, with no architectural change or additional labelled data.

- Across three seeds with a penalty-weight sweep and a naive-masking control, we show the penalty does not improve macro-F1 beyond seed variance, and that the best configuration disables the penalty.
- Through a controlled token-type ablation with a random frequency-matched control, we show LIME-guided suppression performs **no better than random suppression**—the attribution signal carries no usable corrective information. This is our central result.
- We characterise the cause: the flagged tokens are predominantly gendered terms and slurs (65–80% of LIME’s top tokens), i.e. the genuine evidence for sexism; and we show three attribution methods disagree on composition, cautioning against single-attributor conclusions.

2 Related Work

2.1 Online Sexism and the EDOS Benchmark

Detection of online sexism has grown from binary abusive-language classification (Davidson et al., 2017; Founta et al., 2018) to fine-grained taxonomies. EDOS (SemEval-2023 Task 10) provides 20,000 English comments from Reddit and Gab under a three-level hierarchy; Task C is an eleven-way classification (Kirk et al., 2023). Strong systems use transformer encoders with task-adaptive pre-training and ensembling: the highest reported Task C macro-F1 is about 0.54 via a DeBERTa ensemble (Hassan et al., 2023), while competitive single-model systems fall between 0.38 and 0.48 (Peng and Kim, 2023; Mahmoudi, 2023). A recurring theme is confusion among adjacent categories and reliance on surface gender terms (Hemati et al., 2023; Roy and Shrivastava, 2023; Rallabandi et al., 2023). The Chride system (Peng and Kim, 2023) showed that a DeBERTa-v3-large encoder trained with a Task-C-only loss outperforms joint multi-task training, which degrades Task C through gradient interference; we adopt this configuration as our baseline.

2.2 Attribution Methods in Abuse Detection

LIME fits a local linear surrogate to a model’s behaviour on perturbed inputs (Ribeiro et al., 2016); SHAP computes Shapley-value attributions with consistency guarantees (Lundberg and Lee, 2017); Integrated Gradients produces differentiable, path-integrated attributions (Sundararajan et al., 2017). In abuse detection, and EDOS specifically, these are applied almost exclusively post-hoc: Nakwiji et al. (2023) build sexism lexicons with SHAP but do not feed them back into training, and surveys confirm attribution methods are used for diagnosis rather than correction (Mosca et al., 2022). It is also well documented that perturbation-based and gradient-based attributions can disagree substantially on the same model—a point our cross-attributor analysis revisits. LIME in particular has known limitations: a recent survey of LIME and its variants categorises stability (explanations varying across repeated runs and perturbation draws) and fidelity (the surrogate failing to capture the underlying model) as two of its central open issues (Knab et al., 2025). We return to what this implies for the scope of our conclusion in §6.

2.3 Explanation-Guided Training

A separate line modifies training to constrain what models exploit. Ross et al. (2017) penalise input-gradient attributions against annotated relevance masks; Selvaraju et al. (2017) supervise attention via Grad-CAM; Pruthi et al. (2022) study when explanation utility transfers across models. These methods rely on differentiable signals directly accessible from the model, and assume that high-attribution but undesirable features are spurious and safe to suppress. What this literature rarely confronts—and what we examine—is the case where the highest-attribution features are the defining evidence for the label, which arises sharply in sexism detection, where slurs and gendered terms are simultaneously the most salient and the most diagnostic tokens.

Closest to our setting, Attanasio et al. (2022) propose Entropy-based Attention Regularization (EAR), which penalises tokens receiving low self-attention entropy in order to stop hate-speech classifiers overfitting to identity terms. EAR is deliberately *knowledge-free*: it avoids pre-specified term lists precisely because such lists require a-priori knowledge and can introduce further bias. Our method sits at the opposite pole, deriving its sup-

pression targets from a post-hoc attributor rather than from attention statistics or a curated list, and our negative result can be read as evidence about what that particular source of targets does and does not supply. Relatedly, Sen et al. (2022) show that promoting core features over spurious artefacts can itself induce unintended bias: models trained on construct-driven counterfactually augmented data show *higher* false positive rates on non-hateful and non-sexist uses of identity and gendered terms. Their finding reinforces a point we take up in §6, namely that whether a gendered term is evidential depends on the context in which it is used rather than on the token alone.

3 Method

3.1 System Overview

Our objective is measurable: to test whether an explanation-guided penalty improves Task C macro-F1 over a strong baseline with no architectural change, no additional labelled data, and no ensembling. The system has five stages, shown in Figure 1. Stage 1 prepares the EDOS Task C data; Stage 2 trains the baseline; Stage 3 runs LIME on the baseline’s development-set errors to extract a misleading-token set M ; Stage 4 continues training with a penalty on M ; Stage 5 evaluates baseline and penalised checkpoints, together with the controls and ablations that form the core of this study.

3.2 Baseline Classifier

Following Peng and Kim (2023), we fine-tune DeBERTa-v3-large (He et al., 2021) with a single linear head for the eleven-class task on the sexist-only subset of EDOS. Given an input $x = (x_1, \dots, x_T)$, the encoder produces last-layer hidden states h_t ; the pooled [CLS] representation h_0 passes through dropout ($p = 0.1$) and a linear classifier to the eleven logits, trained under cross-entropy. We fine-tune for up to 20 epochs with early stopping (patience 5) on development macro-F1, AdamW (Loshchilov and Hutter, 2019) at learning rate 1×10^{-5} , weight decay 0.01, an effective batch size of 16 (micro-batch 2, accumulation 8), and maximum length 128. Gradient checkpointing fits the model on a single 16 GiB GPU.

3.3 LIME Diagnostic Extraction

After baseline training we evaluate the best checkpoint on the development set and collect its misclassified posts. For up to 50 of these we run LIME

with 3,000 perturbations each, accumulating absolute attribution weights across both the true and predicted classes into a global token dictionary. The 20 tokens with the highest cumulative weight form the misleading-token set M for that seed.

Two properties of this construction should be stated plainly, since both bear on how our result is to be read. First, we accumulate the *absolute magnitude* of attribution weights; the sign of an attribution is not used, and weights for the true and predicted class are pooled into one dictionary. A token that contributes strongly and positively to the correct class can therefore enter M on magnitude alone, without having contributed to the error. Aggregating over the true class was intended to capture evidence the model failed to use, but the consequence is that M is not a set of error-causing tokens in any directional sense; it is a set of high-influence tokens on inputs the model got wrong. A direction-aware variant, restricted to attributions that push toward the incorrect class, is the most immediate follow-up to this study and we do not claim to have tested it. Second, we base the intervention on LIME rather than SHAP or Integrated Gradients because LIME is the attributor most widely used in the EDOS systems we build on and the one a practitioner would most plausibly reach for; we did not select it on the basis of the results it produced. As §5.1 shows, the three attributors disagree considerably about which tokens are most influential, and we did not run SHAP or Integrated Gradients through the penalty pipeline. The extraction hyperparameters (50 misclassified posts, 3,000 perturbations per post, top- $k = 20$) were fixed in advance and not tuned; we did not test the sensitivity of M to them, and note this in §7.

3.4 LIME-Guided Saliency Penalty

Initialising from the baseline checkpoint, we continue training for five epochs at half the learning rate (5×10^{-6}) with an added penalty. Let B_M denote the set of sequences in the batch containing at least one token from the misleading-token set M , P_b the positions of such tokens within sequence b , d the hidden size, and $h_{t,k}^{(b)}$ the k -th component of the last-layer hidden state at position t . The training objective is then:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} + \frac{\lambda}{|B_M|} \sum_{b \in B_M} \frac{1}{|P_b|} \sum_{t \in P_b} \sum_{k=1}^d |h_{t,k}^{(b)}| \quad (1)$$

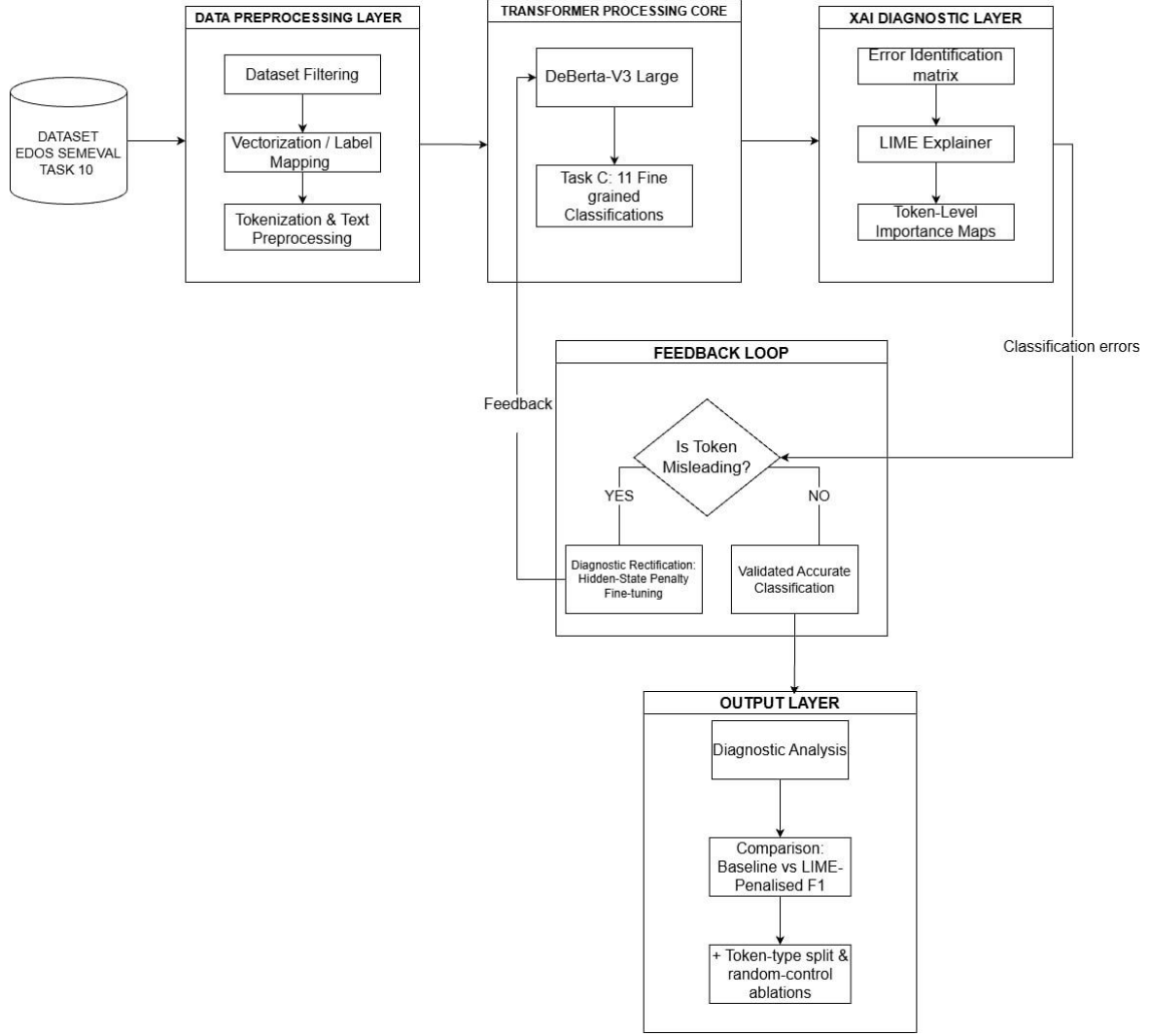


Figure 1: The five-stage LIME-guided saliency penalisation pipeline. The feedback loop (teal) carries the misleading-token set M from the LIME diagnostic into a continued fine-tuning stage that penalises those tokens’ hidden-state magnitudes; Stage 5 adds the token-type and random-control ablations.

No `torch.no_grad()` wrapper is applied, so gradients flow through the penalty and update the encoder. We sweep $\lambda \in \{0.0, 0.1, 0.3, 0.5\}$; $\lambda = 0.0$ is continued training with the penalty disabled, isolating the effect of extra training from the penalty itself. We penalise the mean absolute activation rather than the summed L1 norm; since λ is swept, the two differ only by a constant factor of d . We penalise hidden-state magnitude rather than gradient-based saliency because it is cheap (no second-order gradients) and operates on the final integrated representation; a gradient-based variant exceeds our memory budget.

3.5 Controls and Ablations

Three controls make the evaluation diagnostic rather than merely confirmatory, and each tests a

distinct hypothesis.

(i) **Naive masking** replaces misleading-token positions with [MASK] at probability 0.5 during continued fine-tuning. This is an input-level analogue of the penalty, and it tests whether any benefit comes from suppressing the tokens *at all* rather than from the particular representation-level mechanism we propose. If masking matched or beat the penalty, the hidden-state formulation would be doing no distinctive work.

(ii) **A token-type split** partitions M by linguistic category rather than by attribution score, and applies the penalty to each subset separately. The partition is into function words and dataset artefacts on one side, and gendered terms and slurs on the other. By “artefacts” we mean high-frequency tokens specific to the platform or community from which the

data was drawn rather than general vocabulary, for example *block* and *stacies*; the full assignment for every seed is given in Appendix A. This control tests the mechanistic hypothesis that the penalty fails because it suppresses genuine evidence: if so, suppressing the gendered subset should hurt while suppressing the function-word subset should not.

(iii) **A random frequency-matched control** suppresses a randomly drawn token set of the same size as M , matched for corpus frequency. This is the strictest test in the paper. It asks whether LIME’s specific selection carries any information at all: if informed suppression cannot beat suppression of an arbitrary set of the same size and frequency profile, the attribution signal is not contributing.

Every condition is run across three seeds.

4 Experimental Setup

We use the EDOS aggregated release restricted to its sexist-labelled posts, following the Chride Task C configuration, giving 3,398 training, 486 development, and 970 test posts across the eleven categories (Kirk et al., 2023). The restriction to sexist-labelled posts is imposed by the task definition rather than chosen by us: EDOS Task C is the finest tier of a three-level hierarchy and asks, given that a post has already been identified as sexist, which of eleven categories it belongs to (Kirk et al., 2023). Non-sexist posts have no Task C label. One consequence is that the classifier never sees non-sexist uses of gendered vocabulary during Task C training, so what LIME flags as influential is conditioned on a label space in which such vocabulary is almost always evidential; we return to this in §7. We use EDOS rather than another sexism or misogyny corpus because it is the only widely used English benchmark providing this fine-grained eleven-way taxonomy together with an established set of comparable published systems. The distribution is severely imbalanced—the rarest class, 3.4 (condescending advice), has only a handful of test instances—so per-class F1 in the tail is volatile and, because Task C is scored by macro-F1, that volatility propagates into the headline metric. The primary metric is macro-averaged F1 over the eleven classes; development macro-F1 is used only for checkpoint selection. We report results for three seeds (42, 7, 123) as mean and sample standard deviation, and treat a difference as meaningful only when it is large relative to that

standard deviation. Implementation uses PyTorch, Hugging Face Transformers (Wolf et al., 2020), and scikit-learn (Pedregosa et al., 2011). Code and results are available at <https://github.com/adisa22/woah2026-lime-saliency>.

5 Results

5.1 What the Diagnostic Flags

LIME applied to the baseline’s development-set errors yields a misleading-token set dominated by gendered vocabulary. Across the three seeds, gendered terms and slurs—*woman*, *women*, *she*, *her*, *bitch*, *whore(s)*, *slut(ty)*, and related forms—account for 65%, 65%, and 80% of the top-20 tokens; the remainder are high-frequency function words (*a*, *the*, *that*, *like*, *you*) and a few dataset artefacts. The tokens LIME identifies as most responsible for errors are therefore, in the main, the very terms that define sexist content. This single observation frames everything that follows. The complete per-seed token lists, and the category assignment used for the ablation in §5.3, are given in Appendix A so that these percentages can be audited.

Because LIME is known to be unstable across repeated runs (Knab et al., 2025), we also asked how much M itself varies with the seed. It varies less than that literature might suggest. Pairwise, the three seeds share 13, 13 and 15 of their 20 tokens (Jaccard 0.48, 0.48, 0.60), and twelve tokens—*a*, *bitch*, *female*, *her*, *like*, *men*, *she*, *slutty*, *whore*, *whores*, *woman*, *women*—appear in all three, ten of them gendered terms or slurs. The union across seeds is 31 distinct tokens, so variation is concentrated in the tail rather than the core. This matters for interpreting Tables 2 and 3: the instability of the $\lambda = 0.5$ and naive-masking conditions cannot be attributed to wholesale turnover in M , since the set is largely stable and consistently dominated by the same gendered core.

Because conclusions of this kind often rest on one attributor, we repeated the diagnostic with SHAP and Integrated Gradients on the same errors (seed 42). The three methods agree on a core of gendered tokens—*woman*, *women*, *she*, *her*, *bitch*, *whore* appear in the top set of all three—but disagree sharply on overall composition (Table 5.1). LIME ranks gendered terms most heavily (65%), whereas SHAP and the gradient-based Integrated Gradients disperse more mass onto function words (30% and 35% gendered). We therefore base our

Attributor	Gend./slur	Type
LIME	13/20 (65%)	perturbation
SHAP	6/20 (30%)	perturbation
Integ. Grad.	7/20 (35%)	gradient

Table 1: Composition of the top-20 most influential tokens on the baseline’s errors (seed 42), by attribution method (Ribeiro et al., 2016; Lundberg and Lee, 2017; Sundararajan et al., 2017). A core of gendered terms is common to all three, but the methods disagree on overall composition. Single seed; SHAP uses a regex word masker, so token granularity differs from LIME’s. Indicative of the disagreement, not a stable estimate.

intervention on LIME, consistent with the EDOS systems we build on, but we caution that the precise ranking of “influential” tokens is attributor-dependent and should not be over-interpreted. We stress that the figures in Table 5.1 come from a single seed and that SHAP attributions were computed with a regex word masker, which sets a different token granularity from LIME’s; they indicate the direction and rough size of the disagreement, and should not be read as stable estimates of attributor composition.

5.2 The Penalty Does Not Improve Classification

Table 2 reports test macro-F1 for every configuration across the three seeds. At its best setting ($\lambda = 0.3$) the penalty reaches 0.477 ± 0.003 against a baseline of 0.468 ± 0.004 —a gain of $+0.009$ that barely exceeds run-to-run variation—and stronger penalties degrade performance. The highest mean of all is obtained at $\lambda = 0.0$, continued training with the penalty switched off, at 0.483: whatever movement appears comes from additional fine-tuning, not from the explanation-guided penalty. On the development set the same holds, with $\lambda = 0.3$ (0.535 ± 0.008) marginally below the baseline (0.548 ± 0.026). A naive-masking control is no better and the least stable of all (0.463 ± 0.028); its single-seed advantage over the penalty does not survive replication, reversing in sign across seeds ($+0.039$, -0.011 , $+0.014$).

Two aspects of this table deserve comment. First, our baseline of 0.468 sits in the middle of the published Task C range of roughly 0.38–0.54 noted in §2, above several single-model SemEval systems and below the best reported ensemble at about 0.54 (Hassan et al., 2023). It is a competitive single-model baseline rather than a state-of-the-art one, which is the appropriate reference point for an in-

tervention study of this kind. Second, the $\lambda = 0.0$ condition gains 0.015 over the original baseline, and that improvement comes from continued fine-tuning alone. We did not design the study to explain it and can only note the plausible causes: five further epochs at half the learning rate act as a late-stage annealing schedule, and the checkpoint is re-selected on development macro-F1 after those epochs, which affords an additional selection opportunity. A gain of this size is unsurprising for extended fine-tuning of a large encoder on a small dataset. Its relevance here is methodological: without the $\lambda = 0.0$ arm, this 0.015 would have been available to attribute to the penalty.

5.3 No Better Than Random: The Controlled Ablation

If the penalty helped for the reason its proponents would hope—removing reliance on spurious cues—then suppressing genuinely spurious tokens should help, suppressing genuine-evidence tokens should hurt, and LIME’s informed selection should beat a random one. Table 3 tests all three expectations and none holds. Suppressing only the function-word subset, only the gendered subset, and a random frequency-matched set all leave test macro-F1 within seed-level variation of the baseline (0.467, 0.466, 0.461; baseline 0.468). The gendered-minus-function contrast is $+0.001 \pm 0.013$, centred on zero and reversing across seeds. Most tellingly, LIME-guided suppression performs no better than random suppression: the attribution signal carries no information that a random token set of the same size does not. Whatever LIME identifies, penalising it is not the operative mechanism of any change in performance.

A note on how we read these differences. With three seeds we do not conduct formal hypothesis tests, because inferential statistics at $n = 3$ would offer little power and could convey false precision. We instead compare each difference against the seed-level standard deviation observed for the conditions involved, and treat a difference as meaningful only when it is large relative to that spread. Statements in this paper that two conditions are indistinguishable are therefore claims about practical significance relative to observed run-to-run variability, not the outcome of a significance test.

5.4 Where the Small Movements Go

Although the macro effect is null, the penalty does shift per-class performance (Table 4): modest

Configuration	Seed 42	Seed 7	Seed 123	Mean $\pm$ SD
Baseline (no intervention)	0.467	0.472	0.465	0.468 $\pm$ 0.004
Continued FT, $\lambda = 0.0$ (penalty off)	0.487	0.479	0.484	0.483 $\pm$ 0.004
LIME penalty, $\lambda = 0.1$	0.466	0.465	0.457	0.463 $\pm$ 0.005
LIME penalty, $\lambda = 0.3$	0.474	0.480	0.477	0.477 $\pm$ 0.003
LIME penalty, $\lambda = 0.5$	0.436	0.472	0.466	0.458 $\pm$ 0.019
Naive masking ($p = 0.5$)	0.435	0.491	0.464	0.463 $\pm$ 0.028

Table 2: Test macro-F1 on EDOS Task C across three seeds. $\lambda = 0.0$ is continued fine-tuning with no penalty. The highest mean is obtained with the penalty disabled; no penalised setting separates from the baseline beyond seed-level variation.

Suppressed token set ($\lambda = 0.3$)	Seed 42	Seed 7	Seed 123	Mean $\pm$ SD
Baseline (reference)	0.467	0.472	0.465	0.468 $\pm$ 0.004
Function words only	0.471	0.466	0.465	0.467 $\pm$ 0.003
Gendered terms / slurs only	0.461	0.479	0.457	0.466 $\pm$ 0.012
Random, frequency-matched	0.448	0.477	0.458	0.461 $\pm$ 0.015

Table 3: Token-type ablation ($\lambda = 0.3$), test macro-F1 across three seeds. Targeted suppression of gendered terms, of function words, and a random frequency-matched control all fall within seed-level variation of the baseline and of each other; LIME-guided suppression does not beat random suppression. No formal significance test is applied (see §5.3).

gains on several under-represented or less lexical categories—condescending advice (3.4, +0.059), individual mistreatment (4.1, +0.032), backhanded compliments (3.3, +0.026)—are offset by losses on the most explicitly gendered, lexically signalled categories—gendered slurs (3.1, −0.033), threats (1.1, −0.032), dehumanising attacks (2.3, −0.030). The penalty redistributes performance across classes rather than adding it, consistent with a mechanism that withdraws the model’s strongest lexical cue from exactly the categories that cue defines. The pattern is directional rather than random. Every category that loses ground under the penalty is one whose positive evidence is carried by explicit gendered vocabulary: gendered slurs (3.1), threats (1.1), dehumanising attacks (2.3), and systemic discrimination (4.2), the last only marginally. Every category that gains is one where the sexism is carried by phrasing, stance or implicature rather than by a single lexical marker: condescending advice (3.4), backhanded compliments (3.3), gender stereotypes (3.2), individual mistreatment (4.1). Suppressing the gendered vocabulary appears to shift the decision boundary away from lexical shortcuts and toward contextual cues, which helps where the evidence was never lexical and hurts where it was. The macro average is unchanged because the two effects are of comparable size. We note that movements on minority classes with very few test instances (3.4 has fewer than 15) are individually unstable and should not be read as reliable

Task C class	Base.	$\lambda=0.3$	$\Delta F1$
1.1 threats of harm	0.470	0.438	−0.032
1.2 incitement of harm	0.699	0.723	+0.024
2.1 descriptive attacks	0.548	0.565	+0.017
2.2 aggressive/emotive	0.565	0.582	+0.017
2.3 dehumanising attacks	0.419	0.389	−0.030
3.1 gendered slurs	0.673	0.640	−0.033
3.2 gender stereotypes	0.508	0.531	+0.023
3.3 backhanded compliments	0.239	0.265	+0.026
3.4 condescending advice	0.112	0.171	+0.059
4.1 individual mistreat.	0.341	0.373	+0.032
4.2 systemic discrim.	0.577	0.572	−0.005
Macro average	0.468	0.477	+0.009

Table 4: Per-class test F1 (mean over three seeds), baseline vs. the $\lambda = 0.3$ penalty. Gains concentrate in under-represented, less lexically marked categories; losses concentrate in categories most directly signalled by gendered terms and slurs.

behaviour change.

6 Discussion

Three results combine into a single conclusion. The penalty does not improve macro-F1 (Table 2); it performs no better than random token suppression (Table 3); and the tokens it suppresses are predominantly the gendered terms that constitute the task’s genuine evidence (§5.1). The natural reading is that explanation-guided suppression assumes a separation—between influential-but-spurious features and influential-but-legitimate ones—that does not exist in this task. When the most salient to-

kens are the real signal, an intervention that targets salience targets the signal, and the best it can do is redistribute errors (Table 4) rather than reduce them.

This explanation carries a caveat we should state ourselves rather than leave to the reader. Because M is built from absolute attribution magnitude pooled over the true and predicted classes (§3.3), tokens that are strongly and legitimately predictive of the correct class are eligible for inclusion on magnitude alone. To the degree that this occurs, the construction of M selects genuine evidence for suppression by design, and the explanation becomes partly circular: the intervention fails in part because of how its targets were chosen. We do not think this dissolves the finding, for two reasons. The random control fails equally, which cannot be explained by how M was built, and the per-class pattern in Table 4 shows losses concentrated precisely in the categories the suppressed vocabulary defines. But the circularity is real, and it means our result speaks most directly to magnitude-based, direction-agnostic selection rather than to explanation-guided training in general. A direction-aware construction is the experiment that would separate these.

A further qualification concerns our use of “spurious” and “genuine”. In §3.5 we partition M by linguistic category, which treats spuriousness as a property of the token. Our own discussion undercuts that: what makes a gendered term evidential is that in this label space it reliably marks sexist content, and that is a fact about the relationship between token and label in context, not about the word. A reclaimed slur, a quotation, or a counter-speech post can contain the same token with the opposite import, and Sen et al. (2022) show empirically that models which over-commit to such core features produce more false positives on precisely those non-sexist uses. Our partition should therefore be read as a practical operationalisation adequate for the ablation, not as a claim that these categories are intrinsically spurious or intrinsically evidential.

The random-control result sharpens this. It is not merely that LIME chose poorly; it is that no magnitude-based selection from this token vocabulary carries usable corrective information, because the vocabulary as a whole is evidential. The methodological lesson is that “spuriousness” must be defined linguistically and task-aware, not by attribution magnitude: a token is a legitimate suppression target when it is predictive but illegitimate—an

identity term used as a shortcut to infer unobserved hostility—not merely because it is influential. Distinguishing the two requires external knowledge of what counts as valid evidence, which a threshold over attribution scores does not supply.

Our cross-attributor comparison adds a cautionary note for the broader explanation-guided-training programme. LIME, SHAP, and Integrated Gradients agreed on a core of gendered tokens but disagreed substantially on overall composition (65% vs. 30% vs. 35% gendered). A pipeline that suppresses “the most influential tokens” would therefore suppress materially different sets depending on the attributor chosen. This dependence is easy to overlook when a method is validated against a single explainer, and it also bounds what we can conclude. Our data are consistent with two readings. The first is a task-level claim: attribution magnitude is a poor criterion for suppression in this setting because the high-magnitude tokens are the evidence. The second is a method-level claim: LIME’s attributions are simply too noisy here to be distinguished from a random selection, and a more faithful attributor might behave differently. We ran SHAP and Integrated Gradients as a diagnostic only and did not put either through the penalty pipeline, so we cannot adjudicate between these readings, and we do not claim to. What our results do establish is narrower and still useful: LIME-guided suppression, as it would plausibly be implemented, offers no advantage over random suppression on this task, and the reason is at least partly that its targets are the task’s genuine evidence.

The scope of the conclusion is bounded in a second way. Task C is a setting in which much of the genuine evidence is lexically explicit; slurs and gendered terms are surface markers of the categories they define. Whether the same failure mode arises where the relevant signal is contextual, relational or distributed across a post—coordinated harassment, implicit threats, coded language, dog-whistle rhetoric—remains an open question, and one we would not want a reader to treat as settled by this study. It is plausible that in those settings high-attribution tokens are less often the evidence itself, and explanation-guided suppression correspondingly less self-defeating.

7 Limitations

Our study covers one dataset (EDOS), one language (English), one encoder, and one penalty for-

mulation (L1 hidden-state magnitude); a gradient-based penalty closer to what LIME approximates exceeds our memory budget and is left to future work. Three seeds bound the precision of our variance estimates; we therefore avoid quantitative variance claims, conduct no formal hypothesis tests, and report effects only relative to seed-level standard deviation. Per-class movements on minority categories with very small test support are individually unstable.

Four further limitations bear directly on the interpretation of our central result. First, M is constructed from absolute attribution magnitude with attribution sign discarded and the true and predicted classes pooled, so our finding applies to magnitude-based, direction-agnostic target selection; a direction-aware variant is untested. Second, the extraction hyperparameters (50 misclassified development posts, 3,000 perturbations per post, top- $k = 20$) were fixed in advance and never varied, so we cannot say how sensitive the composition of M , or the results that follow from it, would be to reasonable alternatives. A full sensitivity analysis was beyond the scope of this study. Third, we ran only LIME through the penalty pipeline; SHAP and Integrated Gradients were used diagnostically, so we cannot separate a task-level conclusion about attribution magnitude from a method-level conclusion about LIME specifically. Fourth, Task C is defined only over sexist-labelled posts, so the classifier never encounters non-sexist uses of gendered vocabulary; what LIME identifies as influential is conditioned on that label space, and the evidential status we ascribe to these tokens holds within it rather than in general. Finally, our cross-attributor analysis is at a single seed and SHAP attributions were computed with a regex word masker; the disagreement we report is a caution, not a precise measurement of attributor reliability.

8 Conclusion and Future Work

We introduced LIME-guided saliency penalisation for fine-grained sexism classification and evaluated it as a controlled, multi-seed study. The method does not improve macro-F1 over a strong baseline, the best configuration disables the penalty, and—our central finding—LIME-guided suppression performs no better than suppressing random tokens. The cause is that the flagged tokens are predominantly the gendered terms and slurs that carry the task’s genuine signal, so suppression redistributes

rather than reduces error. We conclude that attribution magnitude, at least as operationalised by LIME and applied without regard to attribution direction, is an unreliable criterion for token suppression in lexically explicit abuse detection. We are deliberate about that scope: we cannot separate this from the possibility that a more faithful attributor would behave differently, and we have not tested settings where the evidence for abuse is contextual rather than lexical. Constructive directions follow: define suppression targets linguistically (function words, names, platform artefacts) rather than by magnitude; pursue interventions that reshape how an evidential token is used in context rather than penalising its representation outright; and, given the attributor disagreement we observe, validate any explanation-guided method against multiple explainers before drawing conclusions.

Ethical Considerations

This study uses an existing, publicly released dataset and involves no new data collection or human subjects. It necessarily quotes offensive gendered slurs because they are the object of analysis and central to the finding; they appear only where required for clarity. Our aim is to make abuse-detection systems more faithful to genuine evidence, and we caution explicitly against interventions that, in seeking to reduce bias, instead remove the legitimate signal these systems depend on. Any artefact derived from this work is a research demonstrator and is not intended for deployment as a moderation decision system without further validation.

References

- Giuseppe Attanasio, Debora Nozza, Dirk Hovy, and Elena Baralis. 2022. [Entropy-based attention regularization frees unintended bias mitigation from lists](#). In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1105–1119, Dublin, Ireland. Association for Computational Linguistics.
- Thomas Davidson, Dana Warmley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*.
- Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (AIES)*.

- Antigoni-Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of Twitter abusive behavior. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*.
- Fadi Hassan, Abdessalam Boucekif, and Walid Aransa. 2023. FiRC at SemEval-2023 task 10: Fine-grained classification of online sexism content using DeBERTa. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*. Association for Computational Linguistics.
- Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTa: Decoding-enhanced BERT with disentangled attention. In *International Conference on Learning Representations (ICLR)*. DeBERTaV3: arXiv:2111.09543.
- Hamed Hematian Hemati, Sayed Hesam Alavian, Hamid Beigy, and Hossein Sameti. 2023. SUTNLP at SemEval-2023 task 10: RLAT-Transformer for explainable online sexism detection. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*. Association for Computational Linguistics.
- Hannah Rose Kirk, Wenjie Yin, Bertie Vidgen, and Paul Röttger. 2023. SemEval-2023 task 10: Explainable detection of online sexism. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2193–2210. Association for Computational Linguistics.
- Patrick Knab, Sascha Marton, Udo Schlegel, and Christian Bartelt. 2025. **Which LIME should I trust? concepts, challenges, and solutions**. In *Explainable Artificial Intelligence (xAI 2025)*, volume 2577 of *Communications in Computer and Information Science*, pages 28–52. Springer, Cham.
- Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*.
- Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Hadiseh Mahmoudi. 2023. IUST_NLP at SemEval-2023 task 10: Explainable detecting sexism with transformers and task-adaptive pretraining. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*. Association for Computational Linguistics.
- Edoardo Mosca, Ferenc Szegedi, Stella Tragianni, Daniel Gallagher, and Georg Groh. 2022. SHAP-based explanation methods: A review for NLP interpretability. In *Proceedings of the 29th International Conference on Computational Linguistics (COLING)*. Association for Computational Linguistics.
- Pakawat Nakwijit, Mahmoud Samir, and Matthew Purver. 2023. Lexicools at SemEval-2023 task 10: Sexism lexicon construction via XAI. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*. Association for Computational Linguistics.
- Fabian Pedregosa and 1 others. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.
- Letian Peng and Bosung Kim. 2023. Chride at SemEval-2023 task 10: Fine-tuned DeBERTa-V3 on detection of online sexism with hierarchical loss. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1670–1675. Association for Computational Linguistics.
- Danish Pruthi, Rachit Bansal, Bhuwan Dhingra, Livio Baldini Soares, Michael Collins, Zachary C. Lipton, Graham Neubig, and William W. Cohen. 2022. Evaluating explanations: How much do explanations from the teacher aid students? *Transactions of the Association for Computational Linguistics*, 10:359–375.
- Sai Kartheek Rallabandi, Sanchit Singhal, and Pratinav Seth. 2023. SSS at SemEval-2023 task 10: Explainable detection of online sexism using majority voted fine-tuned transformers. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*. Association for Computational Linguistics.
- Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. “Why should I trust you?”: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144.
- Andrew Slavin Ross, Michael C. Hughes, and Finale Doshi-Velez. 2017. Right for the right reasons: Training differentiable models by constraining their explanations. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI)*.
- Debashish Roy and Manish Shrivastava. 2023. Attention at SemEval-2023 task 10: Explainable detection of online sexism (EDOS). In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*. Association for Computational Linguistics.
- Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*.
- Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Indira Sen, Mattia Samory, Claudia Wagner, and Isabelle Augenstein. 2022. [Counterfactually augmented data and unintended bias: The case of sexism and hate speech detection](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4716–4726, Seattle, United States. Association for Computational Linguistics.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*.

Thomas Wolf and 1 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.

A Token Lists and Category Assignment

This appendix documents the vocabulary used to categorise tokens in the ablation of §5.3, and the resulting assignment for every seed, so that the composition percentages in §5.1 can be verified.

Category vocabularies. Membership was fixed before the ablation was run and was decided on linguistic grounds, not by which assignment produced a preferred result. The gendered-term list is: *woman, women, girl, girls, female, females, she, her, men, man, male, males, he, his, fem, lady, ladies*. The slur list is: *bitch, bitches, slut, slutty, whore, whores, cunt, cunts, hag, bastards, pussy, slag*. The function-word list comprises high-frequency closed-class English items: *a, an, the, that, this, and, or, but, if, not, of, to, in, for, on, with, as, at, be, is, it, so, we, you, i, they, like*. Tokens in *M* falling outside all three lists were treated as dataset artefacts and grouped with the function words for the ablation.

Per-seed assignment. For seed 42, the gendered/slur group was *bitch, female, girls, hag, her, men, she, slut, slutty, whore, whores, woman, women* (13/20) and the function/artefact group *a, block, fake, like, that, the, you* (7/20). For seed 7, the gendered/slur group was *bastards, bitch, bitches, female, girl, her, men, she, slutty, whore, whores, woman, women* (13/20) and the function/artefact group *a, fem, i, like, stacies, that, they* (7/20). For seed 123, the gendered/slur group was *bitch, bitches, cunt, cunts, female, girl, her, men, pussy, she, slut, slutty, whore, whores, woman, women* (16/20) and the function/artefact group *a, be, like, stacies* (4/20).

Borderline cases. Three assignments deserve comment, since they are the points at which a reader might reasonably disagree. The token *fem* appears in our gendered-term vocabulary but was assigned to the function/artefact group in the seed 7 run; reclassifying it would make that seed 14/20 (70%) rather than 13/20 (65%), leaving the reported range of 65–80% unchanged. We report the assignment as it was actually used in the experiments rather than adjust it after the fact. The token *stacies* is community-specific slang for women originating in manosphere discourse; we treated it as a platform artefact rather than a general gendered term, though it could defensibly be placed in either group. The token *bastards* is a slur but not gender-specific in the way the others are; we retained it in the slur group. None of these three decisions affects the direction of any result in Table 3, where the gendered, function and random conditions are all within seed-level variation of one another.

Stability of *M*. Across the three seeds the sets share 13, 13 and 15 of 20 tokens pairwise. Twelve tokens occur in all three: *a, bitch, female, her, like, men, she, slutty, whore, whores, woman, women*. The union is 31 distinct tokens.