

Evaluating Automatically Inferred Linguistic Perspective as an Auxiliary Signal for Toxicity Detection

Meghana Gottapu

Department of Computer Science, Rutgers University, New Jersey, USA

Abstract

Identity-related language occurs in both self-referential discussion and speech about other groups, complicating toxicity detection. We ask whether an automatically inferred linguistic-perspective signal is useful as an auxiliary training objective. A regex and zero-shot NLI pipeline labels comments as SELF or OTHER; these are noisy linguistic proxies, not verified speaker identities. On Jigsaw Unintended Bias, we fine-tune a Toxic-BERT model with toxicity and proxy heads and compare it with a matched control whose perspective-loss weight is zero. The auxiliary objective gives a small descriptive mean ROC-AUC increase (+0.0038) but lower mean accuracy. On decontaminated CivilComments-WILDS, ROC-AUC is 0.9346 while worst-group accuracy is 0.5805. The main contribution is a leakage-controlled, matched evaluation showing that the full fine-tuning recipe improves toxic-class F1 and ranking substantially, whereas the automatic perspective proxy provides only limited additional benefit.

1 Introduction

Motivation: Toxicity models can associate identity terms with harmful labels even when a comment is non-toxic, producing unintended identity-term bias (Dixon et al., 2018). Abusive-language systems are also sensitive to how harm categories and annotations are defined (Davidson et al., 2017; Founta et al., 2018).

Problem statement: The paper addresses a narrower question: does a distinction between self-referential and other-directed linguistic framing provide a useful auxiliary signal during toxicity fine-tuning? By auxiliary signal, we mean an additional supervised training target: the model is optimized to predict toxicity and the automatic framing label, but the main evaluation remains binary toxicity detection. We operationalize the framing

distinction with two automatic proxies. SELF requires an explicit cue such as “as a Muslim” or “our community.” OTHER denotes either explicit out-group framing or no detected self-reference. The labels do not identify the author’s demographic membership; they are linguistic proxies that enable large-scale auxiliary supervision but do not resolve who authored a comment or whether an identity claim is sincere.

Our contributions are: (1) a provenance-tracked regex/NLI proxy-annotation pipeline; (2) a leakage-controlled protocol in which normalized duplicates stay in one split and all layer, calibration, and policy choices use validation data; and (3) a three-seed matched comparison of perspective-loss weights $\lambda = 0.3$ and $\lambda = 0$, followed by decontaminated CivilComments-WILDS evaluation.

Research objectives. We seek to (1) separate improvement from toxicity fine-tuning from any additional effect of the perspective objective using a matched $\lambda = 0$ condition at a shared raw threshold of 0.50; (2) evaluate a validation-selected routed policy without using test labels; and (3) measure external aggregate and Worst-Group Accuracy (WGA) behavior on CivilComments-WILDS. We therefore treat the work as a controlled evaluation of noisy proxy supervision, not as evidence that true speaker perspective has been measured or that a deployment-ready moderation policy has been produced.

2 Related Work

Prior abusive-language datasets show that toxicity is not a single homogeneous phenomenon: hate speech, offensiveness, targets, and rationales can lead to different annotation and modeling behavior (Davidson et al., 2017; Founta et al., 2018; Mathew et al., 2021). Work on subtle toxicity similarly argues for more fine-grained distinctions than a binary harmful/not-harmful label (Gilda et al., 2022).

Jigsaw comments → regex proxy pass → BART-MNLI for unresolved cases → confidence filtering → duplicate-safe split → Toxic-BERT fine-tuning with toxicity and proxy heads → matched $\lambda = 0$ control → validation-only calibration/routing → Jigsaw and decontaminated CivilComments-WILDS evaluation.

Figure 1: Pipeline overview and component roles. The automatically inferred linguistic-perspective signal is an auxiliary training target, not a replacement for toxicity labels and not a verified speaker-identity annotation.

Our SELF/OTHER labels are best understood in this family of fine-grained descriptive signals, but they describe linguistic framing rather than a new gold taxonomy of harm types.

The closest methodological neighbor is [Vaidya et al. \(2020\)](#), who study multi-task learning for reducing identity bias in toxic-comment detection on the same broad dataset family. That line of work motivates auxiliary supervision, but our question is narrower: whether an automatically inferred linguistic-framing proxy adds value beyond matched toxicity fine-tuning. We therefore compare $\lambda = 0.3$ with a seed-, split-, sampler-, optimizer-, and threshold-matched $\lambda = 0$ control, rather than comparing only against an untrained checkpoint. Other abusive-language debiasing and contextualization methods address related identity-bias failures through different interventions, including gender-bias mitigation ([Park et al., 2018](#)) and post-hoc contextual explanations ([Kennedy et al., 2020](#)). Our contribution is not a new debiasing architecture; it is a controlled test of whether this particular automatic proxy is a useful auxiliary objective.

Perspectivist work shows that annotator beliefs and identities affect toxicity judgments ([Sap et al., 2022](#)). Our use of “perspective” is deliberately more limited: the labels describe textual framing cues and do not verify author identity, annotator identity, or the sincerity of identity claims. This distinction matters for interpretation. If the auxiliary head yields only a small ROC-AUC change, the result should not be read as evidence that true speaker perspective is unhelpful; it may instead indicate that the automatic proxy is noisy or mismatched to the toxicity objective.

Our toxicity encoder is BERT ([Devlin et al., 2019](#)), initialized from the public Toxic-BERT/Detoxify checkpoint ([Hanu and Unitary team, 2020](#)). BART-MNLI ([Lewis et al., 2020](#)) is used only to annotate unresolved proxy-label

cases, not as the toxicity classifier. [Guo et al. \(2017\)](#) show that predictive accuracy does not ensure calibrated confidence; our implemented post-processing method is isotonic regression. We evaluate distribution shift with CivilComments in WILDS ([Koh et al., 2021](#)). Because prior work shows that identity-related toxicity benchmarks are sensitive to annotation choices, dataset composition, and operating thresholds, we interpret the observed 0.0038 mean ROC-AUC gain as a small ranking difference to be weighed against seed variability, lower fixed-threshold accuracy, proxy-label validity, and WGA behavior rather than as a practically meaningful breakthrough by itself.

Our central baseline is therefore internal and matched: the same initialized encoder, split, sampler, optimizer, and checkpoint rule are run with and without the proxy loss. Stronger recent encoders and instruction-tuned LLMs remain important for future competitive benchmarking, but they would not by themselves isolate the auxiliary proxy’s contribution.

3 Methodology

3.1 Data and proxy annotation

We use Jigsaw Unintended Bias in Toxicity Classification ([Jigsaw and Conversation AI, 2019](#)). Comments with at least five toxicity annotators are retained, toxicity is binarized at 0.5, and a seed-42 toxicity-stratified 30% sample yields 185,822 comments. As Figure 1 shows, proxy annotation is a cascade. The first stage uses deterministic expressions for comments where the writer explicitly signals a relation to an identity group, such as self-reference or clear out-group framing. This stage targets explicit cases rather than resolving ambiguous comments. The second stage sends unresolved comments to a pinned `facebook/bart-large-mnli` checkpoint with hypotheses for own-group and outside-group/no-reference framing. Each NLI input uses the first 400 comment characters and appends the names of identity columns with scores of at least 0.5 as group hints. Scores above 0.80 are high confidence, above 0.60 are medium confidence, and all others are low confidence. The pipeline produces 89,860 SELF and 95,962 OTHER labels: 25,372 high-confidence, 92,293 medium-confidence, and 68,157 low-confidence cases. Removing the last group leaves 117,665 examples. No human per-

Table 1: Human validation audit of automatic proxy labels, excluding one human NONE row. Rows are human labels; columns are automatic labels.

	Auto OTHER	Auto SELF
Human OTHER	279	321
Human SELF	98	87

spective labels are used for training, model selection, or threshold tuning, and the toxicity-adjacent `identity_attack` field does not choose proxy labels. The annotation cache records the raw-data checksum, sample settings, regular expressions, NLI revision, policy version, and output checksum. A literal BART end-of-sequence marker is escaped only in the NLI input before tokenization; the stored comment remains unchanged. Low-confidence annotations remain in the provenance output for inspection but are excluded before splitting and training. Medium-confidence proxy losses are down-weighted rather than treated as equivalent to rule-matched high-confidence examples.

Proxy annotation is completed before splitting, but it does not use toxicity labels, model outputs, or split membership. The human labels are never used for training, model selection, calibration, or threshold choice; they are used only afterward for construct-validity checking.

To support construct-validity checking, we manually labeled a subset using a three-way SELF, OTHER, and NONE scheme from comment text alone. The validation audit contains 786 usable labels: 185 SELF, 600 OTHER, and 1 NONE. Excluding the single NONE row, agreement with the automatic binary proxy is 46.6% (macro-F1=0.4323, Cohen’s $\kappa = -0.0457$). Agreement is not monotonic in the automatic confidence tiers: 37.6% for high-confidence rows ($n = 109$), 48.7% for medium-confidence rows ($n = 372$), and 47.4% for low-confidence rows ($n = 304$). The binary confusion matrix is shown in Table 1. These results support treating the labels as noisy automatic linguistic proxies, not as human-validated perspective labels.

The audit is small and single-annotator, so it is not a final annotation study. Its main use is interpretive: any toxicity improvement should be attributed to a noisy automatically inferred framing signal, not to reliable speaker-perspective supervision.

For duplicate grouping only, text undergoes Unicode normalization, case folding, whitespace collapse, and trimming. StratifiedGroupKFold ap-

proximately preserves toxicity $\times$ proxy strata while keeping each normalized group in one split. Train, validation, and test sizes are 82,428, 17,604, and 17,633. The run records 831 duplicate rows beyond first occurrences, 74 duplicate groups with conflicting toxicity labels, and one with conflicting proxy labels. Normalized-text and ID overlap are zero for every split pair.

The split manifest records the source annotation checksum and row membership, so later steps reuse the same partition rather than recomputing a random split. Keeping contradictory duplicates inside a single split does not remove annotation noise, but it avoids turning that noise into train/test contamination.

3.2 Model implementation and training

We fully fine-tune Toxic-BERT with a pair of two-logit heads. The toxicity head reads final-layer [CLS]; the proxy head reads an earlier state selected by train-fitted, validation-evaluated logistic probes. Layer 5 gives the highest validation proxy F1 (0.7738) and is used in all runs. Probes use at most 5,000 training and 2,000 validation examples; the test set is not used for this choice. Layer 5 only narrowly exceeds the final layer (0.7702), so the selected configuration is an empirical choice specific to this setup rather than evidence of a general structural optimum. A plausible interpretation is that mid-layer representations retain lexical and syntactic framing cues relevant to explicit self-reference, while final layers are more adapted to toxicity ranking; this interpretation was not separately tested.

The objective is

$$\mathcal{L} = \mathcal{L}_{\text{focal}} + \lambda \mathcal{L}_{\text{proxy}}, \quad (1)$$

where toxicity focal loss uses $\gamma = 3$ and label smoothing 0.1. Proxy cross-entropy uses inverse-frequency class weights; high-confidence and medium-confidence examples receive weights 1.0 and 0.7. A weighted sampler balances the four toxicity $\times$ proxy cells.

Training uses batch size 32, maximum length 128, dropout 0.3, AdamW with learning rate 10^{-5} and weight decay 0.1, layer-wise decay 0.9, 10% warmup, a cosine schedule, gradient clipping at 1.0, and at most ten epochs. Early stopping uses validation F1 with patience five and restores the best checkpoint. Training uses bfloat16 autocasting on an NVIDIA A100; probability softmax is evaluated in float32. For seeds 42, 43, and 44, multi-task

($\lambda = 0.3$) and matched-control ($\lambda = 0$) runs share the architecture, data, initialization seed, optimizer, schedule, sampler, and checkpoint rule.

The $\lambda = 0$ condition retains the proxy head and sampler but removes the proxy-loss gradient. This isolates the incremental proxy loss, though both conditions still experience the artificial toxicity $\times$ proxy sampling distribution.

3.3 Calibration and policy

For the seed-42 multi-task model, 70% of validation fits isotonic regression; the disjoint 30% selects global and proxy-perspective-specific thresholds by F1. The executed sweep evaluates 0.05–0.95 in increments of 0.01 and selects 0.31 for SELF and 0.41 for OTHER. Routing uses the model-predicted proxy, never the automatic reference proxy. Test labels are not used for calibration or policy selection. The serialized artifact stores the encoder configuration, tokenizer, calibrator, thresholds, run configuration, and checksums; reloaded inference matches pre-export inference.

Routing uses the model’s predicted proxy label rather than the automatic reference proxy, because reference proxy labels would not be available for arbitrary deployment inputs unless the full annotation cascade were run at inference time.

4 Experimental Setup

The non-fine-tuned Toxic-BERT baseline uses the pinned Toxic-BERT toxic logit at threshold 0.50. To isolate the auxiliary effect from policy routing, matched fine-tuned conditions are compared at the same raw 0.50 threshold. We report accuracy, toxic-class F1, ROC-AUC, Average Precision (AP), and 15-bin expected calibration error (ECE). Reported 95% percentile bootstrap intervals use 1,000 row-level resamples from a NumPy Generator initialized with seed 42. Exact McNemar tests apply only to paired correctness and are not interpreted as tests of F1, ROC-AUC, or AP. Automatic reference proxy labels are used only for descriptive stratification.

The baseline diagnostic obtains accuracy 0.8268, F1=0.5367, and ROC-AUC 0.8429. Its SELF false-positive rate is 0.0799, while recall on toxic OTHER examples is 0.4021. This diagnostic motivates fine-tuning but is not used as a test-dependent gate: all predeclared training conditions are run.

External evaluation uses the checksum-pinned CivilComments-WILDS file. Only its predefined

test split is retained, and every ID appearing in the primary Jigsaw sample is removed. From 133,782 predefined test rows, 11,974 overlapping IDs and one row unusable after required-field filtering are removed, leaving 121,807. Official Worst-Group Accuracy (WGA) is the minimum accuracy over positive membership in each of eight identities crossed with binary toxicity. We use CivilComments-WILDS because its predefined split and identity metadata support a standard external WGA evaluation; ToxiGen and HateXplain are not evaluated. This choice measures a published distribution shift close enough to the primary data family for ID decontamination to be meaningful, but it does not establish generalization to adversarial, rationale-annotated, or non-comment settings.

Reproducibility artifacts. The executed run saves the annotation file and manifest, split membership, per-seed raw prediction arrays, calibrated prediction CSV, final metrics, external WILDS metrics, error-analysis rows, tokenizer files, calibrator, configuration, and checksums. Regression checks pass for split contamination, prediction alignment, probability bounds, validation partitioning, artifact round-trip inference, and CivilComments ID removal. Large model weights are stored through Git LFS.

5 Results and Observations

Table 2 summarizes the main evidence. Fine-tuning improves F1 and ROC-AUC markedly over non-fine-tuned Toxic-BERT, but neither fine-tuned condition improves fixed-threshold accuracy. Relative to the matched $\lambda = 0$ control, the auxiliary objective gives a small descriptive mean ROC-AUC gain (+0.0038) and a much smaller mean F1 gain (+0.0013), while lowering mean accuracy (-0.0059). The three raw-threshold ROC-AUC deltas are all positive (+0.0026, +0.0038, +0.0050), but with only three seeds the exact sign test gives $p = 0.25$; this is directional evidence, not a statistically reliable or architecture-general result.

Calibration and routing improve the seed-42 operating point: isotonic regression reduces validation-policy ECE from 0.1402 to 0.0091, and test ECE is 0.0081. However, calibration does not imply a fair or cost-optimal threshold. The automatic-reference-proxy rows show a mixed trade-off: toxic SELF recall is higher than toxic OTHER recall, but non-toxic SELF false positives are also higher. On CivilComments-WILDS, aggre-

Table 2: Core results. Fixed-threshold rows report mean $\pm$ SD over three seeds except the non-fine-tuned checkpoint. The remaining panels use the seed-42 validation-selected multi-task policy.

Panel	System/group	n	Acc.	F1	ROC-AUC	AP/ECE	FPR/WGA
Fixed 0.50	Non-fine-tuned	–	.8268	.5367	.8429	–	–
Fixed 0.50	$\lambda = 0$	–	.8221 $\pm$.0033	.6635 $\pm$.0041	.8907 $\pm$.0009	–	–
Fixed 0.50	Multi-task	–	.8163 $\pm$.0117	.6648 $\pm$.0066	.8945 $\pm$.0012	–	–
Policy	Overall	17,633	.8426	.6727 [.6610,.6851]	.8936 [.8883,.8992]	AP .7477 / ECE .0081	–
Confusion	TN/FP/FN/TP	17,633	12004	1588	1188	2853	–
Proxy	SELF	8,950	–	.6824	–	Recall .7161	FPR .1594
Proxy	OTHER	8,683	–	.6547	–	Recall .6872	FPR .0799
Civil-WILDS	Decontaminated test	121,807	.9196	.6299	.9346 [.9323,.9367]	AP .6897 / ECE .0288	WGA .5805

gate ROC-AUC is strong, but WGA is only 0.5805. The worst group is Christian-identity-positive toxic comments ($n = 1,037$); the corresponding non-toxic Christian group has accuracy 0.9509, so the saved results do not support an over-flagging claim.

6 Discussion

The strongest result is the F1 and ranking improvement from the complete fine-tuning recipe, not the auxiliary objective alone. The fixed-threshold comparison is the cleaner auxiliary-task test because the $\lambda = 0$ proxy head is unsupervised and unstable for routing. The small ROC-AUC increase is positive for all three seeds, but the sign test is underpowered and the experiment uses one Toxic-BERT architecture. Near-equal F1 and lower accuracy therefore argue against attributing the overall fine-tuning gain to perspective supervision.

The proxy-stratified results show a mixed operating-point trade-off: the model recovers more toxic SELF examples, but also falsely flags more non-toxic SELF examples. External ROC-AUC demonstrates ranking under shift, whereas the low WGA shows that aggregate performance is insufficient for deployment claims. The sampler is controlled in the $\lambda = 0.3$ versus $\lambda = 0$ comparison, but it may still affect both models’ absolute accuracy and external robustness relative to natural-distribution training.

The manual validation is the key interpretive constraint. Because automatic labels agree weakly with human text-only judgments and confidence does not predict agreement, the experiment cannot distinguish “perspective is not useful” from “this proxy does not capture perspective reliably enough.” The work is therefore best read as a cautionary, reproducible evaluation of noisy proxy supervision.

7 Limitations and Ethics

The human audit is single-annotator and small; OTHER conflates explicit out-group framing with

no detected self-reference; and the code file stores confidence but not a persistent regex-versus-NLI source indicator. The study uses one English primary dataset, one external distribution, one encoder, and three seeds. It does not ablate regex and NLI labels, confidence weighting, sampling, routing, threshold-grid resolution, or λ values beyond 0 and 0.3; it also does not evaluate ToxiGen (Hartvigsen et al., 2022), HateXplain, DeBERTa-v3, ModernBERT, or instruction-tuned LLM base-lines. These are future-work items, not claims made by the present paper.

Identity-term imbalances can produce unintended bias in toxicity classifiers (Dixon et al., 2018). Our evidence is limited to automatic-proxy FPR differences and CivilComments-WILDS group accuracies; neither identifies comment authors’ demographics. Autonomous moderation, human-review workflows, appeals, and deployment outcomes were not evaluated.

8 Conclusion

Under leakage-controlled evaluation, fine-tuning improves toxic-class F1 and ranking, while the automatic perspective proxy adds only a small, metric-dependent benefit. Calibration still coexists with proxy-group and CivilComments WGA disparities. The result supports protocol design, not verified speaker perspective or uniformly superior multi-task detection.

References

- 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*, 11(1):512-515.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171-4186.
- 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*, pages 67-73.
- 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*, 12(1):491-500.
- Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321-1330.
- Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. ToxiGen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 3309-3326.
- Laura Hanu and Unitary team. 2020. Detoxify. Software repository: <https://github.com/unitaryai/detoxify>.
- Jigsaw and Conversation AI. 2019. Jigsaw unintended bias in toxicity classification. Kaggle competition dataset.
- Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanus Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton Earnshaw, Imran Haque, Sara Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. 2021. WILDS: A benchmark of in-the-wild distribution shifts. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5637-5664.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871-7880.
- Shlok Gilda, Luiz Giovanini, Mirela Silva, and Daniela Oliveira. 2022. Predicting different types of subtle toxicity in unhealthy online conversations. *Procedia Computer Science*, 198:360-366.
- Brendan Kennedy, Xisen Jin, Aida Mostafazadeh Davani, Morteza Dehghani, and Xiang Ren. 2020. Contextualizing hate speech classifiers with post-hoc explanation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5435-5442.
- 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*, 35(17):14867-14875.
- Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing gender bias in abusive language detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2799-2804.
- Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5884-5906.
- Ameya Vaidya, Feng Mai, and Yue Ning. 2020. Empirical analysis of multi-task learning for reducing identity bias in toxic comment detection. In *Proceedings of the International AAAI Conference on Web and Social Media*, 14(1):683-693.