International Journal of Machine Learning
P-ISSN: xxxx-xxxx
E-ISSN: xxxx-xxxx
Vol. 1, No. 1, June 2026, Page. 23-37
https://ijoml.com/
DOI: https://doi.org/10.52436/1.ijoml.1.1.id

RoBERTa with Sample Reweighting and Temperature Scaling for Imbalanced Toxicity Detection: A Performance–Fairness–Calibration Study

Lasmedi Afuan*[1], Nurul Hidayat[2], Abdul Karim[3]

[1,2]Department of Informatics, Universitas Jenderal Soedirman, Indonesia
[3]Department of Artificial Intelligence Convergence, HallymUniversity, Chuncheon 24252, Republic of Korea

Email: [1]lasmedi.afuan@unsoed.ac.id

Received: Jan 9, 2026; Revised: Feb 28, 2026; Accepted: Feb 28, 2026; Published: June 1, 2026

Abstract

Detecting toxic language at scale requires models that are not only accurate but also robust to demographic subgroup bias and reliable in their probability estimates; however, these objectives can conflict, especially under severe class imbalance. This study investigates the performance–fairness–calibration interplay in toxicity detection using the Jigsaw Unintended Bias dataset (124,858 comments; 5.99% toxic; identity annotations in 9.39% of samples). We aim to quantify how sample reweighting and imbalance-aware training affect global discrimination, worst-subgroup behavior, and probabilistic calibration, and to assess post-hoc temperature scaling on predicted probabilities. We compare a TF-IDF + logistic regression baseline against RoBERTa variants trained without mitigation, with sample reweighting, and with an imbalance-oriented loss, using multi-metric evaluation (AUC, Min/Worst-Subgroup AUC, ECE, and NLL). RoBERTa consistently improves global AUC over the baseline (≈0.96 vs 0.9155) while worst-subgroup AUC remains substantially lower and varies modestly across RoBERTa variants (≈0.7726–0.7813). Calibration results indicate a marked gap between models: the baseline achieves the lowest ECE (0.0052), whereas RoBERTa exhibits higher ECE (≈0.0257) that increases further under reweighting and imbalance-oriented training (≈0.0490–0.0866), with NLL not improving consistently. These findings contribute empirical evidence that fairness-oriented interventions can shift error and calibration profiles, motivating holistic evaluation and methods that jointly constrain subgroup fairness and probabilistic reliability.

Keywords: *Calibration, Fairness, Imbalanced classification, RoBERTa, Temperature scaling, Toxicity detection.*

1. INTRODUCTION

Automatic detection of toxic comments has become an essential component of online content moderation as user interactions on digital platforms continue to grow in volume [1], [2]. The massive scale of contemporary data renders manual moderation operationally infeasible, necessitating automated approaches to protect users from exposure to hate speech, harassment, and abuse [1], [3], [4], as well as to comply with regulatory requirements that mandate efficient handling of harmful content [5], particularly in safeguarding vulnerable groups[2].

From a technical perspective, toxicity detection is commonly formulated as a text classification task with probabilistic outputs [6], [7], [8]. Machine learning models are trained to produce probability scores that represent the likelihood of a comment violating community policies [9], [10]. These scores serve as the basis for moderation decisions, either through automatic filtering based on predefined thresholds or via escalation to human moderators for further review [10], [11].

A central challenge in developing such systems lies in class imbalance, where toxic comments constitute only a small fraction of real-world data [6], [12]. This long-tailed distribution biases models

toward the majority class, leading to degraded performance on minority classes [13], [14]. The consequences extend beyond reduced recall for rare toxic comments to distortions in standard evaluation metrics, which often fail to adequately capture the risks associated with errors in critical cases [15], [16].

These issues are further compounded when model errors are unevenly distributed and disproportionately affect specific identity subgroups [17], [18]. Models frequently learn spurious correlations between identity-related terms—such as those referring to race, gender, or sexual orientation—and toxicity due to biases present in training data [2], [8], [19]. As a result, non-toxic comments that mention minority identities tend to exhibit elevated false positive rates, while implicit or contextual attacks against these groups may go undetected [20], [21], [22]. This phenomenon highlights the inadequacy of aggregate metrics such as accuracy or F1-score, which can obscure systematic failures on worst-performing subgroups [17], [23]. Consequently, fair evaluation practices require reporting subgroup-level performance and worst-group metrics to ensure system reliability in sensitive deployment scenarios [17], [24].

Beyond classification performance, the reliability of predicted probabilities constitutes a critical concern. Many modern models produce poorly calibrated predictions, in which confidence scores do not accurately reflect true likelihoods [9], [25], [26]. Overconfident yet incorrect predictions—particularly under distributional shift—complicate the selection of safe decision thresholds and may exacerbate fairness disparities in high-stakes automated moderation systems [17], [27].

Although numerous approaches have been proposed, existing literature largely prioritizes global accuracy optimization. The interplay between performance, subgroup fairness [17], [28], and probabilistic calibration [27], [29] is often examined in isolation. While larger and more complex models tend to achieve higher performance, they also incur substantial computational costs, introducing practical trade-offs for large-scale deployment [6]. A comprehensive evaluation should therefore consider these dimensions jointly, alongside their computational efficiency implications.

Within this context, RoBERTa has been widely adopted as a strong baseline for text classification due to its capacity to capture nuanced linguistic patterns [7], [30], [31]. To address class imbalance, sample reweighting techniques can be employed to amplify the contribution of minority or difficult examples during training [17]. In parallel, temperature scaling offers an efficient post-hoc method for improving probabilistic calibration without altering predicted class labels [9], [32].

Motivated by these gaps, this study aims to compare baseline models with RoBERTa, evaluate the effects of reweighting strategies and probabilistic calibration, and report results holistically across global metrics, subgroup performance, and computational cost using the Jigsaw Unintended Bias in Toxicity Classification dataset. This approach is designed to provide a more comprehensive understanding of the trade-offs among accuracy, algorithmic fairness, and probabilistic reliability in imbalanced toxicity detection, serving as a foundation for the subsequent experimental methodology and results analysis.

2. METHOD

To provide a clear overview of the experimental design, this section describes the end-to-end research workflow, including dataset formulation, preprocessing and deterministic stratified splitting, model training under controlled mitigation scenarios, post-hoc probability calibration, and evaluation of performance, subgroup fairness, calibration quality, and computational cost. The complete workflow and its main components are summarized in Figure 1 to clarify how each stage connects to the study objectives and ensures reproducibility.

International Journal of Machine Learning
P-ISSN: xxxx-xxxx
E-ISSN: xxxx-xxxx
Vol. 1, No. 1, June 2026, Page. 23-37
https://ijoml.com/
DOI: https://doi.org/10.52436/1.ijoml.1.1.id

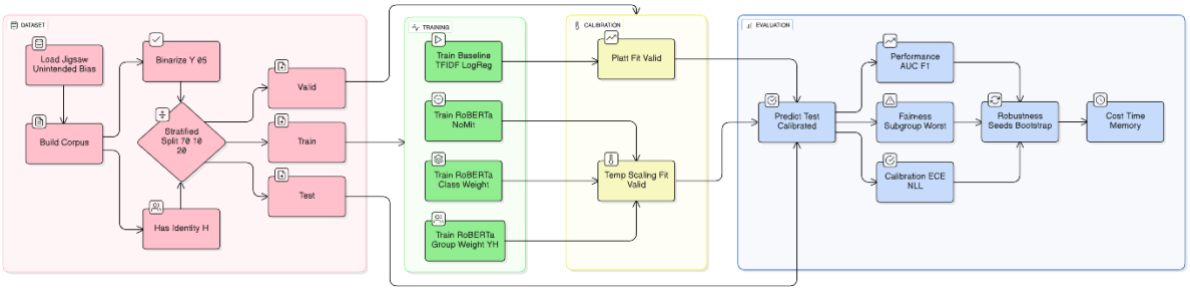

Figure 1. End-to-end workflow of the proposed

2.1. Dataset

This study employs the *Jigsaw Unintended Bias in Toxicity Classification* dataset obtained from Hugging Face, which provides demographic identity attribute annotations based on annotator perceptions and enables explicit evaluation of subgroup bias. All data were consolidated prior to splitting in order to preserve distributional consistency within this inherently imbalanced dataset.

The task is formulated as a binary classification problem, where the toxicity label y is derived from the continuous target score $s$using the standard threshold of 0.5, as defined in Equation (1):

$$y = 1[s \geq 0.5] \quad (1)$$

Demographic identity attributes a_k are binarized using the annotator consensus threshold, as specified in Equation (2):

$$z_k = 1[a_k \geq 0.5] \quad (2)$$

To identify comments that contain references to any identity, a composite indicator $h$is defined as shown in Equation (3):

$$h = 1[\sum_k z_k \geq 1] \quad (3)$$

This formulation enables systematic subgroup construction and supports fairness-aware evaluation under severe class imbalance.

2.2. Preprocessing and Data Splitting

The dataset was split using stratified sampling based on the target label to preserve consistent class proportions across all data subsets, thereby minimizing distributional bias in the evaluation of classification models [23], [31]. The data were partitioned into training, validation, and test sets following standard split ratios, such as 70:10:20 or 80:10:10, to ensure objective and reliable performance assessment [17], [23], [31]. All splitting procedures were conducted deterministically using a fixed random seed to guarantee experimental reproducibility and to prevent performance variability arising from stochastic initialization effects [33], [34].

Text preprocessing was applied to standardize the input data and reduce social media–specific noise without discarding semantically relevant information, a critical step in hate speech detection pipelines [12], [35]. This process included the removal of URLs, HTML elements, and user mentions, as well as text normalization through lowercasing [3], [34], [36]. In addition, non-informative non-alphanumeric symbols, irregular repeated characters, and irrelevant punctuation were eliminated to produce a consistent input representation [12]. Sequence length was further controlled through truncation, with a maximum input length set to 128 tokens. This configuration provides an effective

trade-off between contextual coverage and computational efficiency for Transformer-based models such as BERT and RoBERTa [24], [37], [38], [39].

2.3. Model Architectures

The methodological implementation begins with the development of a traditional machine learning baseline that combines statistical text representations using the Term Frequency–Inverse Document Frequency (TF–IDF) scheme with a limited n-gram range to capture salient lexical patterns [12]. This feature representation is modeled using Logistic Regression (LR), which approximates a linear relationship between the bag-of-words feature vector and the target class probability, as formulated in Equation (4):

$$P(y = 1 \mid x) = \sigma(\mathbf{w}^{\top}\phi(x)), \qquad (4)$$

where σdenotes the logistic function and $\phi(x)$represents the TF–IDF feature vector [40]. To improve the reliability of probabilistic estimates on unseen data, probability calibration is performed using Platt scaling on the validation set. This procedure maps the raw classifier score $f(x)$to calibrated probabilities via a sigmoid function, as shown in Equation (5):

$$P(y = 1 \mid f) = \frac{1}{1+\exp\,(A \cdot f(x)+B)} \qquad (5)$$

where the parameters $A$and $B$are optimized by minimizing the log-loss on the calibration set [26].

As a comparison to this statistical approach, the study employs the Transformer-based RoBERTa model, which represents the state of the art due to its large-scale pretraining and enhanced optimization strategies. Input text is processed using Byte-Pair Encoding (BPE) tokenization, followed by padding and truncation to a fixed maximum sequence length $L_{\max}$to ensure consistent dimensionality across mini-batches [41]. The contextualized representation of the special $[CLS]$token is then projected through a linear binary classification head to produce the output probability distribution.

2.4. Training Procedure and Mitigation Components

This study adopts RoBERTa-base as the core model and evaluates it under three controlled experimental scenarios designed to assess the mitigation of data imbalance and identity-based bias. The baseline scenario (NoMit) applies standard fine-tuning without any modification to the loss function, a setting known to be susceptible to spurious correlations in imbalanced data [17]. The second scenario (Imb) incorporates class weighting to correct the global label distribution, a strategy shown to be effective for text classification under long-tailed distributions [14]. The third scenario (Reweight) applies group-based sample reweighting to balance sample contributions across combinations of target labels and sensitive identity attributes, with the objective of reducing lexical bias against minority groups [42].

The base loss function is Binary Cross-Entropy (BCE), formulated in Equation (6):

$$\ell_i = -\log\, p_\theta(y_i \mid x_i) \qquad (6)$$

In the Imb scenario, class weights $w_c$are computed in inverse proportion to class frequencies to prevent dominance by the majority class, yielding the weighted loss defined in Equation (7):

$$\ell_i^{cw} = w_{y_i}\, \ell_i, w_c = \frac{N}{K \cdot n_c} \qquad (7)$$

For more fine-grained bias mitigation, the Reweight scenario defines groups $g = (y, h)$ as combinations of the target label and identity indicator. This approach follows the principles of

Distributionally Robust Optimization (DRO), which prioritizes minimizing risk on the worst-performing groups to improve overall model fairness [28]. The aggregated loss is normalized to maintain gradient stability during training and is expressed in Equation (8):

$$\mathcal{L}(\theta) = \frac{\sum_i w_{g_i} \ell_i}{\sum_i w_{g_i}}, w_g = \frac{N}{G \cdot n_g} \quad (8)$$

As a post-training step, probabilistic calibration is performed using Temperature Scaling to enhance the reliability of prediction scores, particularly under distribution shift [43]. The temperature parameter T is optimized on the validation set by minimizing the Negative Log-Likelihood (NLL), as defined in Equation (9):

$$T^* = \arg\min_T \sum_{i \in \text{valid}} -\log \text{ softmax}\left(\frac{\mathbf{z}_i}{T}\right)_{y_i} \quad (9)$$

The calibrated logits $\mathbf{z}/T^*$ are then used for final inference, a procedure empirically shown to reduce Expected Calibration Error (ECE) in natural language understanding tasks [32].

2.5. Evaluation Protocol and Metrics

Performance evaluation is conducted using calibrated prediction probabilities to ensure the reliability of model confidence estimates [26]. The baseline model is calibrated via Platt scaling, while RoBERTa models employ temperature scaling, which has been shown to effectively align probability distributions without affecting classification accuracy [26]. Global performance is primarily assessed using the Area Under the Receiver Operating Characteristic Curve (ROC-AUC) as a threshold-independent discrimination metric, complemented by F1-score, precision, and recall computed at the standard decision threshold of 0.5 [12], [44].

Fairness evaluation is performed at the identity subgroup level by computing subgroup-specific ROC-AUC scores and analyzing error rates derived from the confusion matrix [44]. Two key indicators are the False Positive Rate (FPR) and False Negative Rate (FNR), defined in Equation (10):

$$FPR = \frac{FP}{FP+TN}, FNR = \frac{FN}{FN+TP} \quad (10)$$

where FP, TN, FN, and TP denote the components of the confusion matrix [2]. To summarize potential systemic bias, worst-group metrics are reported, including the minimum ROC-AUC and the maximum FPR and FNR observed across all subgroups [17].

Probabilistic calibration quality is assessed using binning-based Expected Calibration Error (ECE), which quantifies the discrepancy between empirical accuracy and predicted confidence, as formalized in Equation (11):

$$ECE = \sum_{b=1}^{B} \frac{|S_b|}{N} \left| acc(S_b) - conf(S_b) \right| \quad (11)$$

where B denotes the number of bins, N the total number of samples, and $\text{acc}(S_b)$ and $\text{conf}(S_b)$ represent the accuracy and average confidence within the b-th bin, respectively [45]. This calibration analysis is complemented by the Brier score and Negative Log-Likelihood (NLL) as indicators of sharpness and overall probabilistic quality [9].

To ensure robustness and generalizability, all RoBERTa experiments are conducted using multiple random seeds, with results reported as mean ± standard deviation [1]. Performance differences

across scenarios are statistically validated using bootstrapping on the test set to empirically estimate confidence intervals and uncertainty [46].

2.6. Measurement of Computational Cost and Reproducibility Protocol

Algorithmic efficiency is evaluated through systematic logging of computational costs to compare the baseline model and RoBERTa across different mitigation scenarios. These costs include training time, inference latency, and peak GPU memory usage, following standard practices in large-scale deep learning experiments[1], [17]. This evaluation is intended to ensure that improvements in model fairness do not incur disproportionate computational overhead, given the inherent trade-offs between model complexity and resource consumption in practical NLP applications [6], [26].

Reproducibility is maintained by using fixed random seeds for all data splits and multi-run training procedures to mitigate stochastic variance in Transformer-based models [37]. All experiments apply standardized preprocessing pipelines and a consistent maximum sequence length ($L_{\max}$) to minimize systematic bias [3], [47]. Calibration parameters are optimized exclusively on the validation set to prevent data leakage, while performance and fairness evaluations are reported under a unified protocol on a previously unseen test set [4], [42].

3. RESULT

The experiments were conducted on a dataset comprising 124,858 samples with a highly imbalanced label distribution, where 5.99% of instances were classified as toxic and 94.01% as non-toxic. Demographic identity attributes were available for 9.39% of the samples. The data were partitioned using a 70:10:20 train/validation/test split, while preserving a consistent proportion of toxic instances across all splits. A comprehensive summary of the dataset statistics, including label proportions per split, identity availability, and the extent of identity overlap, is presented in Table 1.

Table 1. Overall Dataset Summary and Split Statistics

Split	Count	Share of Dataset	Toxic (%)	Non-toxic (%)	Has identity (%)	Identity overlap (%)
Train	87,400	70.00%	5.99%	94.01%	9.44%	2.78%
Validation	12,486	10.00%	5.99%	94.01%	8.68%	2.56%
Test	24,972	20.00%	5.99%	94.01%	9.57%	2.89%
Total (All)	124,858	100.00%	5.99%	94.01%	9.39%	—

A comparison of global discriminative performance, fairness performance on the worst-performing subgroup—defined as the minimum AUC across identity subgroups—and computational cost across all modeling scenarios is summarized in Table 2. Consistently, all Transformer-based variants achieve higher global AUC than the TF–IDF + logistic regression baseline. The unmitigated RoBERTa model attains the highest mean global AUC (0.9618 ± 0.0012), followed by RoBERTa with class weighting (0.9612 ± 0.0004) and RoBERTa with group-based reweighting (0.9600 ± 0.0006), whereas the baseline records a substantially lower value of 0.9155 ± 0.0000. This pattern is further corroborated by the precision–recall curves in Figure 4(c), where the Transformer variants maintain higher precision over a broader range of recall compared with the baseline on the imbalanced test set.

Table 2. Performance, Fairness, and Computational Efficiency Across Modeling Scenarios

Scenario	AUC	Min Subgroup AUC	Training Time (s)	Peak VRAM (MB)	Inference Throughput (samples/s)
Baseline (TF–IDF)	0.9155 ± 0.0000	0.7632 ± 0.0000	33.9683	228.3242	4075.441

Vol. 1, No. 1, June 2026, Page. 23-37
https://ijoml.com/
DOI: https://doi.org/10.52436/1.ijoml.1.1.id

Transformer (Class Weighting)	0.9612 ± 0.0004	0.7813 ± 0.0239	1348.376	8034.606	667.1836
Transformer (No Mitigation)	0.9618 ± 0.0012	0.7726 ± 0.0384	1351.177	8034.606	670.6863
Transformer (Group Reweighting)	0.9600 ± 0.0006	0.7778 ± 0.0235	1351.07	8034.606	666.4366

From an efficiency perspective, the computational metrics reported in Table 2 and the visualizations in Figures 2(a)–(b) reveal a clear separation between the baseline and Transformer-based models. The baseline operates with training times on the order of tens of seconds and a memory footprint of several hundred megabytes, whereas all RoBERTa configurations require training times of approximately 1.35×10^3 seconds, with peak VRAM usage around 8.03×10^3 MB and inference throughput in the range of 6.66×10^2 to 6.71×10^2 samples per second. In Figure 2(c), the Pareto front illustrates the relationship between global AUC and worst-group AUC across the different modeling scenarios.

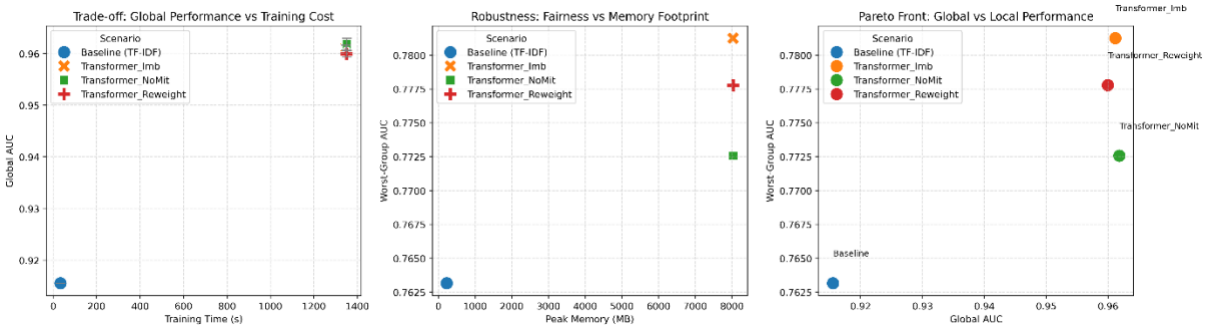

Figure 2. Trade-off analysis and Pareto front: (a) the relationship between global performance and training time, (b) model fairness with respect to memory footprint, and (c) the Pareto front illustrating the trade-off between global AUC and subgroup AUC.

The classification error behavior on the test set is depicted by the normalized confusion matrices in Figure 3. The baseline model achieves a sensitivity of 38.17% for the toxic class and a specificity of 99.14% for the non-toxic class. In contrast, all Transformer-based models exhibit substantially higher sensitivity: the unmitigated RoBERTa reaches 69.59%, RoBERTa with class weighting attains 85.03%, and RoBERTa with group-based reweighting achieves 76.74%. These gains in sensitivity are accompanied by corresponding changes in specificity, decreasing to 92.78% under class weighting and 95.35% under group-based reweighting, while the unmitigated configuration lies between these two extremes, as reflected in Figure 3.

International Journal of Machine Learning
P-ISSN: xxxx-xxxx
E-ISSN: xxxx-xxxx
Vol. 1, No. 1, June 2026, Page. 23-37
https://ijoml.com/
DOI: https://doi.org/10.52436/1.ijoml.1.1.id

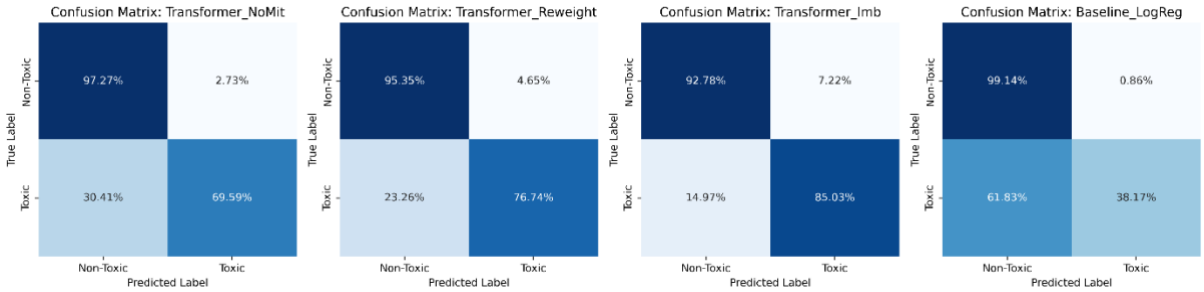

Figure 3. Normalized Confusion Matrices for All Modeling Scenarios

For fairness evaluation, the comparison of worst-group AUC (defined as the minimum subgroup AUC) in Table 2 indicates that class weighting yields the highest mean worst-group AUC (0.7813 ± 0.0239), followed by group-based reweighting (0.7778 ± 0.0235), while the unmitigated configuration attains 0.7726 ± 0.0384. Bootstrap-based tests applied to the fairness comparisons report no statistically significant differences between the mitigated and unmitigated models at the 0.05 significance level ($p > 0.05$). Detailed subgroup-wise AUCs are visualized in Figure 4(b), illustrating variability across identity groups for each modeling scenario, whereas Figure 2(b) shows that these fairness differences among Transformer variants occur under comparable peak GPU memory usage.

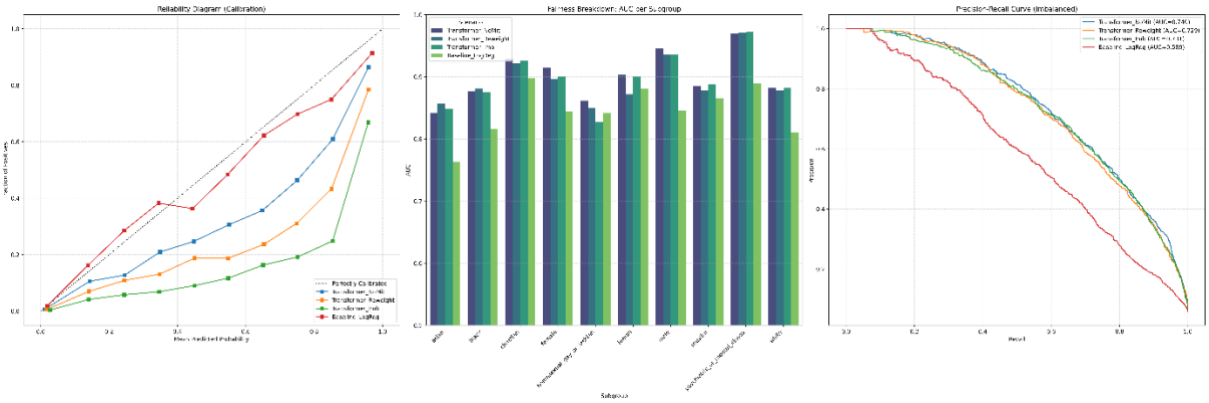

Figure 4. Evaluation of calibration, fairness, and discriminative performance: (a) reliability diagram illustrating probabilistic calibration, (b) subgroup-wise AUC comparison across identity groups, and (c) precision–recall curves on the imbalanced test set.

The reliability of predicted probabilities is summarized in Table 3 and visualized using reliability diagrams in Figure 4(a). The baseline model exhibits the lowest Expected Calibration Error (ECE) of 0.0052, with a corresponding negative log-likelihood (NLL) of 0.1339. In contrast, all Transformer-based models display higher ECE values: 0.0257 (NLL 0.1109) for the unmitigated RoBERTa, 0.0490 (NLL 0.1401) for group-based reweighting, and 0.0866 (NLL 0.1967) for class weighting. These deviations of the reliability curves from the ideal diagonal line are evident in Figure 4(a), with the largest deviation observed for the class-weighted configuration.

Table 3. Probability Calibration Metrics Across Modeling Scenarios

Model Scenario	ECE	NLL (Cross-Entropy)
Baseline (TF-IDF + Logistic Regression)	0.0052	0.1339
RoBERTa without Bias Mitigation	0.0257	0.1109
RoBERTa with Group-Based Reweighting	0.049	0.1401
RoBERTa with Class Weighting	0.0866	0.1967

4. DISCUSSIONS

The findings of this study reveal an empirical trade-off among global discriminative performance, subgroup fairness, and probabilistic reliability in highly imbalanced toxicity detection. Compared with the TF–IDF + logistic regression baseline (global AUC = 0.9155), all RoBERTa variants achieve substantially higher global AUC, with the unmitigated configuration attaining the highest value (0.9618 ± 0.0012). However, this improvement does not automatically translate into enhanced fairness under the definition adopted in this study—namely, the minimum AUC across identity subgroups (worst-group AUC). For the unmitigated RoBERTa, the worst-group AUC remains at 0.7726 ± 0.0384, indicating a persistent performance gap between aggregate metrics and the most disadvantaged subgroups. This pattern is consistent with concerns that pretrained language models trained on large-scale web corpora may reflect sociolinguistic biases and spurious associations related to identity mentions [48], [49]. Importantly, the present study does not investigate the causal mechanisms of such biases and restricts its assessment to measurable impacts via identity-based metrics.

The two mitigation strategies examined—class weighting and identity group–based reweighting—produce only modest shifts in worst-group AUC. Specifically, worst-group AUC increases from 0.7726 (unmitigated) to 0.7813 ± 0.0239 ($\Delta \approx +0.0087$) with class weighting, and to 0.7778 ± 0.0235 ($\Delta \approx +0.0052$) with group-based reweighting. However, the reported bootstrap tests indicate that these differences are not statistically significant at $\alpha = 0.05$ ($p > 0.05$). Under the experimental conditions considered, the mitigation strategies are therefore more appropriately characterized as inducing small shifts in worst-group performance rather than yielding statistically conclusive fairness improvements. The difficulty of achieving stable fairness gains, as well as their sensitivity to metric choice and evaluation procedures, has also been emphasized in prior fairness literature [18], [50].

In terms of global performance, the differences in AUC among RoBERTa variants are relatively small compared with the large improvement over the baseline, and they follow a pattern in which mitigation is accompanied by slight absolute reductions in AUC: 0.9618 ± 0.0012 (unmitigated) versus 0.9612 ± 0.0004 (class weighting; $\Delta \approx -0.0006$) and 0.9600 ± 0.0006 (group-based reweighting; $\Delta \approx -0.0018$). For threshold-dependent metrics, the confusion matrix results reveal more pronounced trade-offs on the imbalanced dataset. The baseline exhibits very high specificity (99.14%) but low sensitivity for the toxic class (38.17%), whereas RoBERTa substantially increases sensitivity (69.59% without mitigation; 85.03% with class weighting; 76.74% with group-based reweighting) at the cost of reduced specificity (92.78% and 95.35% for the two mitigation strategies). Because identity-conditional error analyses (e.g., FPR/TPR per subgroup or analyses restricted to samples with identity annotations) are not reported, these changes should be interpreted as shifts in the aggregate error profile rather than as direct evidence of error dynamics specific to identity-bearing texts.

With respect to probabilistic quality, the results further demonstrate that discriminative superiority does not necessarily imply reliable probability estimates. Under the temperature scaling procedure employed in this study (with the optimal temperature estimated on the validation set), the baseline retains the lowest Expected Calibration Error (ECE = 0.0052), whereas RoBERTa exhibits higher ECE values (0.0257 without mitigation; 0.0490 with group-based reweighting; 0.0866 with class weighting). Negative log-likelihood (NLL), however, reveals a different nuance: unmitigated RoBERTa achieves a lower NLL (0.1109) than the baseline (0.1339), while both mitigation strategies increase NLL (0.1401 and 0.1967). This divergence between ECE and NLL aligns with arguments that modern models may excel in ranking performance while remaining miscalibrated under certain reliability measures, underscoring the need to evaluate calibration using multiple metrics [26]. From a practical

perspective, these findings suggest that selecting training configurations and decision thresholds in moderation systems may require explicit trade-offs among sensitivity, score reliability, and operational objectives.

Several limitations constrain the scope of the conclusions. First, identity attributes are available for only approximately 9.39% of samples, with identity overlap of about 2.56–2.89% per split, which may render identity-based fairness estimates unstable, particularly for low-frequency subgroups—consistent with critiques of identity benchmarks that are sensitive to distributional sparsity [23]. Second, the evaluation is conducted on an English-language dataset, limiting cross-linguistic and cross-cultural generalization [12], [14]. Third, fairness is proxied solely through worst-group AUC and bootstrap testing. Future work should incorporate identity-conditional error analyses, compare more expressive mitigation approaches (e.g., adversarial or contrastive methods aimed at disentangling toxicity signals from identity tokens [18], [51], complement distribution-based evaluation with functional tests such as HateCheck to assess robustness beyond dataset artifacts [23], and strengthen audits through explainability analyses [52] and targeted strategies for small subgroups [42].

5. CONCLUSION

This study examines the interplay between discriminative performance, identity subgroup fairness, and probabilistic calibration in highly imbalanced toxicity detection. Consistently, RoBERTa outperforms the TF–IDF + logistic regression baseline in terms of global AUC (≈0.96 vs. 0.9155); however, performance on identity subgroups remains substantially lower, with minimum (worst-subgroup) AUC values ranging from approximately 0.7726 to 0.7813 across all RoBERTa variants. Interventions aimed at addressing imbalance (e.g., reweighting or loss modification) primarily shift the trade-off profile across evaluation dimensions rather than yielding uniform improvements in either global AUC or fairness metrics.

From a probabilistic reliability perspective, RoBERTa variants exhibit greater miscalibration than the baseline, as evidenced by increases in ECE from 0.0052 (baseline) to approximately 0.0257 (unmitigated) and further to about 0.0490–0.0866 under mitigation strategies, alongside NLL values that do not improve consistently. These findings underscore the necessity of holistic evaluations that jointly report performance, subgroup fairness, and calibration. The study is limited by its focus on a single monolingual dataset and identity definitions based on annotator perceptions; future research should assess cross-lingual and cross-domain generalization and develop objectives that explicitly integrate fairness and calibration considerations.

CONFLICT OF INTEREST

The authors declares that there is no conflict of interest between the authors or with research object in this paper.

ACKNOWLEDGEMENT

The authors would like to thank all individuals and institutions that provided guidance, support, and contributions during the research activities and manuscript preparation, whose input and cooperation were essential to the successful completion of this study. is only addressed to funders or donors and object of research.

REFERENCES

[1] K. Maity, A. s. Poornash, S. Saha, and P. Bhattacharyya, “ToxVidLM: A Multimodal Framework for Toxicity Detection in Code-Mixed Videos,” in *Findings of the Association for Computational Linguistics ACL 2024*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2024, pp. 11130–11142. doi: 10.18653/v1/2024.findings-acl.663.

[2] M. Mamta, R. Chigrupaatii, and A. Ekbal, "BiasWipe: Mitigating Unintended Bias in Text Classifiers through Model Interpretability," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2024, pp. 21059–21070. doi: 10.18653/v1/2024.emnlp-main.1172.

[3] T. A. Suman and A. Jain, "AStarTwice at SemEval-2021 Task 5: Toxic Span Detection Using RoBERTa-CRF, Domain Specific Pre-Training and Self-Training," in *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2021, pp. 875–880. doi: 10.18653/v1/2021.semeval-1.118.

[4] L. Pozzobon, B. Ermis, P. Lewis, and S. Hooker, "On the Challenges of Using Black-Box APIs for Toxicity Evaluation in Research," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2023, pp. 7595–7609. doi: 10.18653/v1/2023.emnlp-main.472.

[5] C. Luo, R. Bhambhoria, S. Dahan, and X. Zhu, "Legally Enforceable Hate Speech Detection for Public Forums," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2023, pp. 10948–10963. doi: 10.18653/v1/2023.findings-emnlp.730.

[6] Z. Hu, J. Piet, G. Zhao, J. Jiao, and D. Wagner, "Toxicity detection for free," in *Proceedings of the 38th International Conference on Neural Information Processing Systems*, in NIPS '24. Red Hook, NY, USA: Curran Associates Inc., 2024.

[7] D. Sarkar, M. Zampieri, T. Ranasinghe, and A. Ororbia, "fBERT: A Neural Transformer for Identifying Offensive Content," in *Findings of the Association for Computational Linguistics: EMNLP 2021*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2021, pp. 1792–1798. doi: 10.18653/v1/2021.findings-emnlp.154.

[8] G. Zhang, B. Bai, J. Zhang, K. Bai, C. Zhu, and T. Zhao, "Demographics Should Not Be the Reason of Toxicity: Mitigating Discrimination in Text Classifications with Instance Weighting," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2020, pp. 4134–4145. doi: 10.18653/v1/2020.acl-main.380.

[9] Y. Xiao, P. P. Liang, U. Bhatt, W. Neiswanger, R. Salakhutdinov, and L.-P. Morency, "Uncertainty Quantification with Pre-trained Language Models: A Large-Scale Empirical Analysis," in *Findings of the Association for Computational Linguistics: EMNLP 2022*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2022, pp. 7273–7284. doi: 10.18653/v1/2022.findings-emnlp.538.

[10] M. D. Muralikumar, Y. S. Yang, and D. W. McDonald, "A Human-centered Evaluation of a Toxicity Detection API: Testing Transferability and Unpacking Latent Attributes," *ACM Transactions on Social Computing*, vol. 6, no. 1–2, pp. 1–38, Jun. 2023, doi: 10.1145/3582568.

[11] H. Park, H. Shim, and K. Lee, "Uncovering the Root of Hate Speech: A Dataset for Identifying Hate Instigating Speech," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2023, pp. 6236–6245. doi: 10.18653/v1/2023.findings-emnlp.412.

[12] M. S. Jahan and M. Oussalah, "A systematic review of hate speech automatic detection using natural language processing," *Neurocomputing*, vol. 546, p. 126232, Aug. 2023, doi: 10.1016/j.neucom.2023.126232.

[13] J. Xin, R. Tang, Y. Yu, and J. Lin, "The Art of Abstention: Selective Prediction and Error Regularization for Natural Language Processing," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2021, pp. 1040–1051. doi: 10.18653/v1/2021.acl-long.84.

[14] Y. Huang, B. Giledereli, A. Köksal, A. Özgür, and E. Ozkirimli, "Balancing Methods for Multi-label Text Classification with Long-Tailed Class Distribution," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2021, pp. 8153–8161. doi: 10.18653/v1/2021.emnlp-main.643.

[15] P. Xu, L. Xiao, B. Liu, S. Lu, L. Jing, and J. Yu, "Label-Specific Feature Augmentation for Long-Tailed Multi-Label Text Classification," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 9, pp. 10602–10610, Jun. 2023, doi: 10.1609/aaai.v37i9.26259.

[16] S. Henning, W. Beluch, A. Fraser, and A. Friedrich, "A Survey of Methods for Addressing Class Imbalance in Deep-Learning Based Natural Language Processing," in *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2023, pp. 523–540. doi: 10.18653/v1/2023.eacl-main.38.

[17] S. Sagawa, P. W. Koh, T. B. Hashimoto, and P. Liang, "Distributionally Robust Neural Networks for Group Shifts: On the Importance of Regularization for Worst-Case Generalization," 2020. [Online]. Available: https://arxiv.org/abs/1911.08731

[18] X. Zhou, M. Sap, S. Swayamdipta, Y. Choi, and N. Smith, "Challenges in Automated Debiasing for Toxic Language Detection," in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2021, pp. 3143–3155. doi: 10.18653/v1/2021.eacl-main.274.

[19] I. Sen, M. Samory, C. Wagner, and I. Augenstein, "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*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2022, pp. 4716–4726. doi: 10.18653/v1/2022.naacl-main.347.

[20] M. ElSherief *et al.*, "Latent Hatred: A Benchmark for Understanding Implicit Hate Speech," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2021, pp. 345–363. doi: 10.18653/v1/2021.emnlp-main.29.

[21] V. Raman, E. Fleisig, and D. Klein, "Centering the Margins: Outlier-Based Identification of Harmed Populations in Toxicity Detection," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2023, pp. 9316–9329. doi: 10.18653/v1/2023.emnlp-main.579.

[22] J. Schäfer, U. Heid, and R. Klinger, "Hierarchical Adversarial Correction to Mitigate Identity Term Bias in Toxicity Detection," in *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2024, pp. 35–51. doi: 10.18653/v1/2024.wassa-1.4.

[23] P. Röttger, B. Vidgen, D. Nguyen, Z. Waseem, H. Margetts, and J. Pierrehumbert, "HateCheck: Functional Tests for Hate Speech Detection Models," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2021, pp. 41–58. doi: 10.18653/v1/2021.acl-long.4.

[24] N. Sahoo, H. Gupta, and P. Bhattacharyya, "Detecting Unintended Social Bias in Toxic Language Datasets," in *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2022, pp. 132–143. doi: 10.18653/v1/2022.conll-1.10.

[25] A. Gasparin and G. Detommaso, "Distance-aware Calibration for Pre-trained Language Models," in *Findings of the Association for Computational Linguistics: EMNLP 2024*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2024, pp. 12434–12447. doi: 10.18653/v1/2024.findings-emnlp.725.

[26] S. Desai and G. Durrett, "Calibration of Pre-trained Transformers," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2020, pp. 295–302. doi: 10.18653/v1/2020.emnlp-main.21.

[27] G. He, J. Chen, and J. Zhu, "Preserving Pre-trained Features Helps Calibrate Fine-tuned Language Models," 2023. [Online]. Available: https://arxiv.org/abs/2305.19249

[28] S. S. Ghosal and Y. Li, "Distributionally Robust Optimization with Probabilistic Group," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 10, pp. 11809–11817, Jun. 2023, doi: 10.1609/aaai.v37i10.26394.

[29] J. Geng, F. Cai, Y. Wang, H. Koeppl, P. Nakov, and I. Gurevych, "A Survey of Confidence Estimation and Calibration in Large Language Models," in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2024, pp. 6577–6595. doi: 10.18653/v1/2024.naacl-long.366.

[30] J. Lu *et al.*, "Take Its Essence, Discard Its Dross! Debiasing for Toxic Language Detection via Counterfactual Causal Effect," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, and N. Xue, Eds., Torino, Italia: ELRA and ICCL, May 2024, pp. 15566–15578. [Online]. Available: https://aclanthology.org/2024.lrec-main.1353/

[31] R. Steed, S. Panda, A. Kobren, and M. Wick, "Upstream Mitigation Is Not All You Need: Testing the Bias Transfer Hypothesis in Pre-Trained Language Models," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2022, pp. 3524–3542. doi: 10.18653/v1/2022.acl-long.247.

[32] S. Y. Park and C. Caragea, "On the Calibration of Pre-trained Language Models using Mixup Guided by Area Under the Margin and Saliency," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2022, pp. 5364–5374. doi: 10.18653/v1/2022.acl-long.368.

[33] P. Giovannotti, " Calibration of Natural Language Understanding Models with Venn–ABERS Predictors," in *Proceedings of the Eleventh Symposium on Conformal and Probabilistic Prediction with Applications*, U. Johansson, H. Boström, K. An Nguyen, Z. Luo, and L. Carlsson, Eds., in Proceedings of Machine Learning Research, vol. 179. PMLR, Dec. 2022, pp. 55–71. [Online]. Available: https://proceedings.mlr.press/v179/giovannotti22a.html

[34] R. Sridhar and D. Yang, "Explaining Toxic Text via Knowledge Enhanced Text Generation," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2022, pp. 811–826. doi: 10.18653/v1/2022.naacl-main.59.

[35] J. Lu, B. Xu, X. Zhang, C. Min, L. Yang, and H. Lin, "Facilitating Fine-grained Detection of Chinese Toxic Language: Hierarchical Taxonomy, Resources, and Benchmarks," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2023, pp. 16235–16250. doi: 10.18653/v1/2023.acl-long.898.

[36] X. Cui, "Addressing Data Imbalance in Transformer-Based Multi-Label Emotion Detection with Weighted Loss," in *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, S. Rosenthal, A. Rosá, D. Ghosh, and M. Zampieri, Eds., Vienna, Austria: Association for Computational Linguistics, Jul. 2025, pp. 247–256. [Online]. Available: https://aclanthology.org/2025.semeval-1.37/

[37] S. Panda, A. Kobren, M. Wick, and Q. Shen, "Don't Just Clean It, Proxy Clean It: Mitigating Bias by Proxy in Pre-Trained Models," in *Findings of the Association for Computational Linguistics: EMNLP 2022*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2022, pp. 5073–5085. doi: 10.18653/v1/2022.findings-emnlp.372.

[38] J. Pavlopoulos, L. Laugier, A. Xenos, J. Sorensen, and I. Androutsopoulos, "From the Detection of Toxic Spans in Online Discussions to the Analysis of Toxic-to-Civil Transfer," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2022, pp. 3721–3734. doi: 10.18653/v1/2022.acl-long.259.

[39] I. Chalkidis, T. Pasini, S. Zhang, L. Tomada, S. Schwemer, and A. Søgaard, "FairLex: A Multilingual Benchmark for Evaluating Fairness in Legal Text Processing," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2022, pp. 4389–4406. doi: 10.18653/v1/2022.acl-long.301.

[40] I. Sen, M. Samory, F. Flöck, C. Wagner, and I. Augenstein, "How Does Counterfactually Augmented Data Impact Models for Social Computing Constructs?," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2021, pp. 325–344. doi: 10.18653/v1/2021.emnlp-main.28.

[41] A. Vazhentsev *et al.*, "Uncertainty Estimation of Transformer Predictions for Misclassification Detection," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2022, pp. 8237–8252. doi: 10.18653/v1/2022.acl-long.566.

[42] T. Garg, S. Masud, T. Suresh, and T. Chakraborty, "Handling Bias in Toxic Speech Detection: A Survey," *ACM Comput Surv*, vol. 55, no. 13s, pp. 1–32, Dec. 2023, doi: 10.1145/3580494.

[43] J. Zhang, W. Yao, X. Chen, and L. Feng, "Transferable Post-hoc Calibration on Pretrained Transformers in Noisy Text Classification," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 11, pp. 13940–13948, Jun. 2023, doi: 10.1609/aaai.v37i11.26632.

[44] B. Mathew, P. Saha, S. M. Yimam, C. Biemann, P. Goyal, and A. Mukherjee, "HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 17, pp. 14867–14875, May 2021, doi: 10.1609/aaai.v35i17.17745.

[45] Y. Chen, L. Yuan, G. Cui, Z. Liu, and H. Ji, "A Close Look into the Calibration of Pre-trained Language Models," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2023, pp. 1343–1367. doi: 10.18653/v1/2023.acl-long.75.

[46] C. Si, C. Zhao, S. Min, and J. Boyd-Graber, "Re-Examining Calibration: The Case of Question Answering," in *Findings of the Association for Computational Linguistics: EMNLP 2022*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2022, pp. 2814–2829. doi: 10.18653/v1/2022.findings-emnlp.204.

[47] M. Nadeem, A. Bethke, and S. Reddy, "StereoSet: Measuring stereotypical bias in pretrained language models," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2021, pp. 5356–5371. doi: 10.18653/v1/2021.acl-long.416.

[48] S. L. Blodgett, S. Barocas, H. Daumé III, and H. Wallach, "Language (Technology) is Power: A Critical Survey of 'Bias' in NLP," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2020, pp. 5454–5476. doi: 10.18653/v1/2020.acl-main.485.

[49] S. Gehman, S. Gururangan, M. Sap, Y. Choi, and N. A. Smith, "RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2020, pp. 3356–3369. doi: 10.18653/v1/2020.findings-emnlp.301.

[50] X. Han and Y. Tsvetkov, "Fortifying Toxic Speech Detectors Against Veiled Toxicity," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2020, pp. 7732–7739. doi: 10.18653/v1/2020.emnlp-main.622.

[51] L. Cheng, A. Mosallanezhad, Y. N. Silva, D. L. Hall, and H. Liu, "Bias Mitigation for Toxicity Detection via Sequential Decisions," in *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, New York, NY, USA: ACM, Jul. 2022, pp. 1750–1760. doi: 10.1145/3477495.3531945.

[52] M. Sap, S. Swayamdipta, L. Vianna, X. Zhou, Y. Choi, and N. Smith, "Annotators with Attitudes: How Annotator Beliefs And Identities Bias Toxic Language Detection," in

Vol. 1, No. 1, June 2026, Page. 23-37
https://ijoml.com/
DOI: https://doi.org/10.52436/1.ijoml.1.1.id

Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Stroudsburg, PA, USA: Association for Computational Linguistics, 2022, pp. 5884–5906. doi: 10.18653/v1/2022.naacl-main.431.