Mitigating Algorithmic Bias in Toxicity Classification for Muslim Identity Contexts: A Study on FPR Parity

Umar Hasan [1]

Abstract

As automated moderation systems become ubiquitous, ensuring they operate fairly across diverse religious identities is critical. This study investigates unintended bias in toxicity classification models, focusing on Muslim identity terms. Using the Jigsaw Unintended Bias dataset, this study identifies a 10x disparity in false-positive rates for Muslim-related content in a baseline model. It proposes a sample-weighting mitigation strategy that successfully achieves FPR parity, reducing the identity-specific error rate from 0.0273 to 0.0015. Furthermore, feature importance analysis shows that the intervention successfully decouples religious identifiers from toxic associations.

1. Introduction

The rapid deployment of automated toxicity classifiers is essential for maintaining safe digital environments at scale. However, historical biases in training data often cause machine learning models to associate specific identity markers with harm or toxicity (Dixon et al., 2018; Blodgett et al., 2020). For global Muslim communities, this unintended bias can lead to the systemic silencing of benign religious expression (Abid et al., 2021), effectively marginalizing these voices in digital public squares.

When a model penalizes a user simply for mentioning their faith, it ceases to be a safety tool and becomes an instrument of algorithmic discrimination. This work addresses these harms by evaluating and mitigating identity-specific bias in a standard content moderation pipeline. It mathematically quantifies the bias present in a baseline text classifier and demonstrates that targeted data interventions can restore equitable performance. Specifically, this article makes three contributions:

- It quantifies false-positive-rate disparity for Muslim identity terms in a standard toxicity classification pipeline using the Jigsaw Unintended Bias dataset.
- It introduces a simple sample-weighting mitigation strategy that substantially reduces subgroup false positives while preserving overall accuracy.
- It provides an interpretable feature-weight analysis showing that mitigation reduces reliance on Muslim identity terms as toxicity indicators.

2. Related Work

Algorithmic fairness in natural language processing has become a critical area of research as models increasingly mediate online social interactions. (Blodgett et al., 2020) highlighted the need for formal, grounded definitions of bias in NLP systems, noting that harms often stem from systemic representational disparities. Early work by (Garg et al., 2018) demonstrated that foundational word embeddings mathematically capture and amplify historical ethnic stereotypes.

Building on this, (Dixon et al., 2018) and (Borkan et al., 2019) proved that toxicity detection models frequently exhibit unintended bias against specific demographic and identity terms. Subsequent work on counterfactual fairness in text classification formalized how predictions can change when sensitive identity tokens are substituted (Garg et al., 2019), while ToxiGen further demonstrated the need for challenging toxic and benign examples about minority groups (Hartvigsen et al., 2022). Together, these studies show that disproportionate false positive rates for marginalized groups can turn moderation tools into exclusionary filters.

Within the specific context of Muslim identities, recent literature has exposed severe representational harms. (Abid et al., 2021) revealed persistent anti-Muslim bias in large language models like GPT-3, showing that systems systematically associate Muslim identifiers with violence due to skewed pre-training corpora. Follow-up work found that even debiased language models can retain Muslim-violence associations under name-based prompts (Hemma-

[1]Department of Electrical and Computer Engineering, North South University, Dhaka 1229, Bangladesh. Correspondence to: Umar Hasan <umar.hasan@northsouth.edu>.

Proceedings of the 43rd International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

tian & Varshney, 2022). Complementing this line of work, recent Muslim-context benchmarks such as BIICK-Bench (Hasan, 2025) and IslamTrust (Lahmar et al., 2025) highlight the need for culturally, religiously, and linguistically specific evaluation resources for AI systems used by Muslim communities. Table 1 provides a summary of these foundational works. The present research directly builds upon this literature by explicitly isolating this bias in content moderation baselines and proposing a targeted, highly interpretable mitigation strategy.

3. Methods

3.1. Problem Formulation

Let X represent the input space of text documents and $Y \in \{0, 1\}$ represent the binary toxicity label, where $Y = 1$ indicates a toxic comment. A sensitive attribute $A \in \{0, 1\}$ is introduced, where $A = 1$ indicates the presence of a Muslim identity term within the text.

The goal of a standard text classifier is to learn a mapping $f_\theta : X \to Y$ parameterized by θ that minimizes the empirical risk over the training distribution. However, when the training data contains historical societal biases, the model inadvertently learns a spurious correlation between $A = 1$ and $Y = 1$. This results in a violation of demographic parity and predictive equality. The objective is to optimize f_θ such that the overall classification error is minimized while simultaneously ensuring that the False Positive Rate (FPR) when $A = 1$ is strictly equal to the FPR when $A = 0$. Figure 1 illustrates the complete experimental workflow designed to solve this formulation.

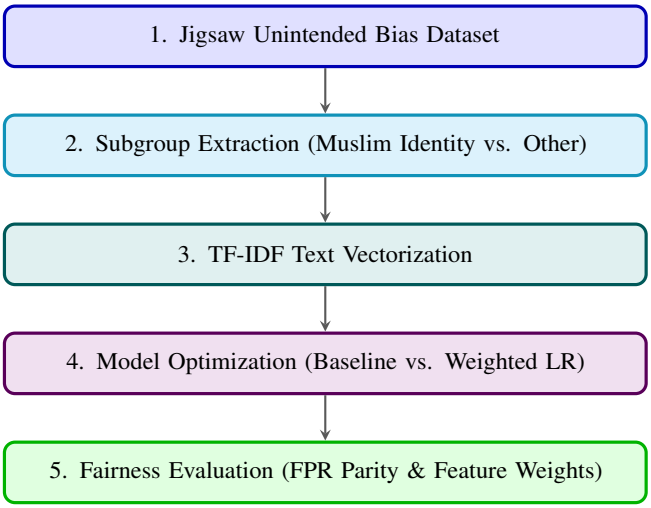

Figure 1. Workflow illustrating the experimental pipeline from raw data to fairness evaluation.

3.2. Dataset and Preprocessing

This study utilized the Jigsaw Unintended Bias in Toxicity Classification dataset (Borkan et al., 2019) hosted on Kaggle. This corpus contains millions of public comments sourced from the Civil Comments platform. Uniquely, a subset of these comments was annotated by human raters not just for toxicity, but for the presence of specific identity markers including religious affiliation.

To isolate and evaluate the bias effectively without introducing severe class imbalances, a controlled experimental subset was extracted. This consisted of all comments explicitly annotated with the *Muslim* identity flag (where the annotator consensus was ≥ 0.5) and a randomized, equal-sized sample of comments containing other or no identity flags. The composition of this experimental subset is detailed in Table 2.

3.3. Mathematical Formulation of Fairness

In content moderation, the most damaging error for a marginalized group is the False Positive (FP), where benign speech is incorrectly flagged as toxic. The False Positive Rate (FPR) for a given subgroup $A = a$ is defined as:

$$FPR_a = P(\hat{Y} = 1 | Y = 0, A = a)$$

Where $\hat{Y}$ is the model prediction. Algorithmic bias is present when the FPR for the Muslim identity subgroup significantly exceeds the FPR of the general population.

3.4. Proposed Mitigation

To address the disparity, a "Sample Weighting" strategy was implemented during the optimization phase. A sample weight w_i was defined for each training instance i. For non-toxic samples belonging to the Muslim identity subgroup, the weight was increased by a factor of $\alpha = 5.0$.

$$w_i = \begin{cases} 5.0 & \text{if } Y_i = 0 \text{ and } A_i = 1 \\ 1.0 & \text{otherwise} \end{cases}$$

This algorithmic intervention forces the model's loss function to penalize errors on benign, identity-focused content more heavily. The complete procedural logic is detailed in Algorithm 1.

3.5. Experimental Setup

The baseline employs a Term Frequency-Inverse Document Frequency (TF-IDF) vectorizer (Manning et al., 2008) limited to 10,000 features. It uses a Logistic Regression classifier (Cox, 1958) ($\lambda = 1.0$, $SEED = 42$), implemented using standard scikit-learn tooling (Pedregosa et al., 2011). A linear model was intentionally selected over a deep neural

Table 1. Summary of Key Literature in NLP Fairness and Bias

REFERENCE	FOCUS AREA	KEY FINDING OR METHODOLOGY
(DIXON ET AL., 2018)	TOXICITY CLASSIFICATION	INTRODUCED FALSE POSITIVE EQUALITY DIFFERENCE TO MEASURE BIAS.
(GARG ET AL., 2018)	WORD EMBEDDINGS	DEMONSTRATED THAT EMBEDDINGS CAPTURE HISTORICAL ETHNIC STEREOTYPES.
(BORKAN ET AL., 2019)	EVALUATION METRICS	DEVELOPED NUANCED SUBGROUP METRICS FOR UNINTENDED BIAS.
(GARG ET AL., 2019)	COUNTERFACTUAL FAIRNESS	STUDIED ROBUSTNESS TO SENSITIVE-TOKEN SUBSTITUTIONS IN TEXT CLASSIFIERS.
(HARTVIGSEN ET AL., 2022)	TOXICITY DATASETS	CREATED TOXIC AND BENIGN EXAMPLES ABOUT MINORITY GROUPS FOR STRESS TESTING.
(BLODGETT ET AL., 2020)	NLP SYSTEMS OVERVIEW	HIGHLIGHTED THE LACK OF FORMAL DEFINITIONS OF BIAS IN LITERATURE.
(ABID ET AL., 2021)	LARGE LANGUAGE MODELS	REVEALED PERSISTENT ANTI-MUSLIM BIAS AND VIOLENT ASSOCIATIONS IN LLMS.
(HEMMATIAN & VARSHNEY, 2022)	LARGE LANGUAGE MODELS	SHOWED THAT DEBIASED LLMS CAN RETAIN MUSLIM-VIOLENCE ASSOCIATIONS.
(HASAN, 2025)	CULTURALLY SPECIFIC BENCHMARKS	INTRODUCED A BENGALI BENCHMARK FOR ISLAMIC CREED KNOWLEDGE IN LLMS.

Table 2. Experimental Dataset Composition

DATA SPLIT / CATEGORY	COUNT
TOTAL JIGSAW CORPUS	∼1.8 MILLION
VALIDATION HOLDOUT	20%
IDENTITY RATIO (MUSLIM : OTHER)	1:1

network to ensure complete feature interpretability, allowing exact coefficient weights to be traced back to specific vocabulary tokens.[1]

4. Results

4.1. Fairness and Accuracy Metrics

The experiments reveal a striking initial disparity. As shown in Figure 2, the baseline model was over ten times more likely to misclassify a benign comment as toxic if it mentioned the Muslim identity (FPR 0.0273) compared to the general population (FPR 0.0026).

The mitigated model effectively neutralized this bias. The FPR for Muslim identifiers dropped to 0.0015, achieving parity with the general population (0.0016). This was achieved with a minimal trade-off in overall accuracy (87.15% to 86.33%), which is a highly acceptable compromise in fairness-aware machine learning.

[1]Project code repository: https://github.com/umarbhasan/muslim-toxicity-fairness.

Algorithm 1 Sample-Weighted Fairness Optimization

Input: Dataset $\mathcal{D} = \{(x_i, Y_i, A_i)\}_{i=1}^{N}$, penalty factor $\alpha = 5.0$
Output: Optimized model parameters θ^*
Phase 1: Weight Initialization
for $i = 1$ **to** N **do**
  if $Y_i = 0$ **and** $A_i = 1$ **then**
    $w_i \leftarrow \alpha$
  else
    $w_i \leftarrow 1.0$
  end if
end for
Phase 2: Feature Extraction
Transform documents X into TF-IDF matrix X_{vec}
Phase 3: Model Optimization
Optimize θ by minimizing the weighted log-loss:
$\mathcal{L}(\theta) = -\sum_{i=1}^{N} w_i \big[$
  $Y_i \log f_\theta(x_{vec,i}) + (1 - Y_i) \log(1 - f_\theta(x_{vec,i}))\big]$
Return: θ^*

4.2. Model Interpretability

Because the pipeline uses TF-IDF Logistic Regression, the resulting models remain highly interpretable. Extracting the highest-weighted coefficients (Table 3) reveals exactly which terms the models associate most strongly with toxicity.

In the baseline model, the terms *muslims* and *muslim* ranked 7th and 13th respectively, nestled among universally toxic terms like *stupid* and *idiot*. This mathematically proves the model learned to view the religious identity itself as a toxic indicator due to skewed training distributions.

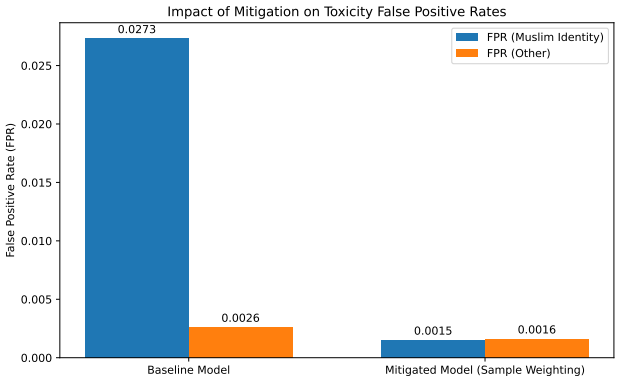

Figure 2. Comparison of False Positive Rates (FPR) before and after applying the sample weighting mitigation strategy. The intervention successfully closes the gap between the identity subgroups.

Table 3. Top Toxic Features (Logistic Regression Coefficients)

Baseline Model		Mitigated Model	
Feature	Weight	Feature	Weight
stupid	11.73	stupid	13.96
idiot	7.45	idiot	9.32
ignorant	6.54	ignorant	7.41
idiots	6.32	idiots	7.31
dumb	6.10	dumb	7.06
pathetic	5.56	pathetic	6.90
muslims	5.37	crap	6.08
ridiculous	5.31	ridiculous	6.03
crap	5.05	stupidity	5.43
kill	5.00	hypocrite	5.42
hypocrite	4.97	kill	5.42
stupidity	4.91	garbage	5.22
muslim	4.84	fool	5.21
fool	4.71	liar	4.93
garbage	4.45	disgusting	4.73

In contrast, the sample weighting mitigation successfully decoupled this association. In the mitigated model, all religious identifiers completely drop out of the top features, replaced entirely by legitimate toxic terminology.

5. Discussion

The results have several implications for fairness-aware content moderation. First, they show that aggregate accuracy can obscure subgroup harms: the baseline appears reasonably accurate overall, yet disproportionately flags benign comments mentioning Muslim identity. Second, the weighted model shows that targeted interventions on benign identity-bearing examples can substantially reduce false positives. Finally, feature-weight analysis demonstrates why subgroup evaluation should be paired with model inspection. FPR parity shows balanced error rates, while the disappearance of Muslim identity terms from the top toxic features provides stronger evidence that the model is no longer relying on religious identifiers as toxicity proxies.

While the mitigation strategy successfully achieved parity, several scope limitations remain. First, the empirical evaluation relies on a single dataset (Jigsaw) and a linear modeling framework (Logistic Regression). Given the prevalence of Large Language Models (LLMs) in modern moderation pipelines, applying these fairness constraints to deep architectures requires techniques such as adversarial debiasing. Second, the evaluation is limited to English-language data. The global Muslim community is highly multilingual, and recent benchmarks such as BIICK-Bench (Hasan, 2025) and IslamTrust (Lahmar et al., 2025) underscore the importance of evaluating AI systems in non-English and religiously grounded Muslim contexts. Furthermore, token fragmentation and latent-space entanglement can degrade safety constraints in low-resource languages like Bengali (Hasan & Nayeem, 2026). Future work must therefore address toxicity bias and cross-lingual vulnerabilities in languages such as Arabic, Urdu, Bengali, and Bahasa Indonesia. Finally, this study does not account for intersectionality; future research should investigate how false positive rates compound when users identify with multiple marginalized groups.

Deployment monitoring remains necessary as topics, slang, and political events shift; platforms should periodically recompute subgroup FPR, review religious-language false positives, and update weights or data when disparities reappear.

6. Conclusion

This study demonstrates that significant algorithmic bias against Muslim identities exists in standard moderation baselines but can be mitigated through simple fairness-aware interventions. Enforcing FPR parity and analyzing feature weights showed that data balancing can prevent the marginalization of vulnerable voices. The machine learning community must prioritize such evaluations so safety tools protect, rather than penalize, global communities.

Impact Statement

This article advances Machine Learning fairness. Its intended societal impact is positive: reducing disproportionate censorship of Muslim voices on digital platforms. While the mitigation improves fairness for one subgroup, developers must ensure that balancing weights for one identity does not degrade performance or introduce new biases against other marginalized communities. Future audits should include other affected communities, helping identify emerging disparities for timely mitigation updates.

References

Abid, A., Farooqi, M., and Zou, J. Persistent anti-muslim bias in large language models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, pp. 298–306, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450384735. doi: 10.1145/3461702.3462624. URL https://doi.org/10.1145/3461702.3462624.

Blodgett, S. L., Barocas, S., Daumé III, H., and Wallach, H. Language (technology) is power: A critical survey of "bias" in NLP. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J. (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5454–5476, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.485. URL https://aclanthology.org/2020.acl-main.485/.

Borkan, D., Dixon, L., Sorensen, J., Thain, N., and Vasserman, L. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion Proceedings of The 2019 World Wide Web Conference*, WWW '19, pp. 491–500, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450366755. doi: 10.1145/3308560.3317593. URL https://doi.org/10.1145/3308560.3317593.

Cox, D. R. The regression analysis of binary sequences. *Journal of the Royal Statistical Society: Series B (Methodological)*, 20(2):215–232, 07 1958. ISSN 0035-9246. doi: 10.1111/j.2517-6161.1958.tb00292.x. URL https://doi.org/10.1111/j.2517-6161.1958.tb00292.x.

Dixon, L., Li, J., Sorensen, J., Thain, N., and Vasserman, L. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, pp. 67–73, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450360128. doi: 10.1145/3278721.3278729. URL https://doi.org/10.1145/3278721.3278729.

Garg, N., Schiebinger, L., Jurafsky, D., and Zou, J. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644, 2018. doi: 10.1073/pnas.1720347115. URL https://www.pnas.org/doi/abs/10.1073/pnas.1720347115.

Garg, S., Perot, V., Limtiaco, N., Taly, A., Chi, E. H., and Beutel, A. Counterfactual fairness in text classification through robustness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '19, pp. 219–226, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450363242. doi: 10.1145/3306618.3317950. URL https://doi.org/10.1145/3306618.3317950.

Hartvigsen, T., Gabriel, S., Palangi, H., Sap, M., Ray, D., and Kamar, E. ToxiGen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. In Muresan, S., Nakov, P., and Villavicencio, A. (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3309–3326, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.234. URL https://aclanthology.org/2022.acl-long.234/.

Hasan, U. BIICK-bench: A bengali benchmark for introductory islamic creed knowledge in large language models. In *5th Muslims in ML Workshop co-located with NeurIPS 2025*, 2025. URL https://openreview.net/forum?id=HSkp30vSrm.

Hasan, U. and Nayeem, M. A. Mathematical foundations of cross-lingual vulnerabilities in llms: Latent space entanglement and token fragmentation. *Mathematics*, 14(11), 2026. ISSN 2227-7390. doi: 10.3390/math14111849. URL https://www.mdpi.com/2227-7390/14/11/1849.

Hemmatian, B. and Varshney, L. R. Debiased large language models still associate muslims with uniquely violent acts, 2022. URL https://arxiv.org/abs/2208.04417.

Lahmar, A., Arafat, M. E., Farou, Z., and Mahmud, M. Islamtrust: A benchmark for LLMs alignment with islamic values. In *5th Muslims in ML Workshop co-located with NeurIPS 2025*, 2025. URL https://openreview.net/forum?id=PBcv90iKFB.

Manning, C. D., Raghavan, P., and Schütze, H. *Introduction to information retrieval*. Cambridge university press, 2008.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Édouard Duchesnay. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(85):2825–2830, 2011. URL http://jmlr.org/papers/v12/pedregosa11a.html.