Quantifying Bias in Hindi/Hinglish Toxicity Classification

Saurabh Gupta, Devansh Singh, Aviral Chandra, Rachit Mittal, Chanakya Nath
Department of Computer Science and Engineering
SRM Institute of Science and Technology,
Uttar Pradesh, India – 201204
saurabhg1@srmist.edu.in
ds2553@srmist.edu.in, ac5379@srmist.edu.in, rm8782@srmist.edu.in, cn2211@srmist.edu.in

Abstract

Toxicity classifiers trained on Hindi–English code-mixed text tend to over-flag content that simply mentions certain identity groups, producing unequal false positive rates across religions, castes, and genders. We audit a standard XLM-RoBERTa fine-tune on a 29,550-sample Hinglish hate speech corpus using three fairness metrics — false positive rate disparity, demographic parity difference, and a counterfactual fairness gap measured on identity-swapped sentence pairs — and find that the baseline model assigns a false positive rate of 0.42 to caste-mentioning text, nearly three times the rate for gender-mentioning text. We then compare two mitigation strategies under identical conditions: Counterfactual Data Augmentation (CDA), which adds identity-swapped copies of training examples, and adversarial debiasing, which uses a gradient reversal layer to discourage identity encoding. CDA reduces FPR disparity by 54% and demographic parity gap by 73% at a cost of only 1.8 F1 points, outperforming the adversarial approach on every fairness measure. The adversarial model improves group-level statistics but actually worsens counterfactual fairness relative to the baseline, suggesting that gradient reversal can produce shallow decorrelation rather than genuine identity invariance. We release our code, the augmented training corpus, and a counterfactual test set to support future fairness work in Indian-language NLP.

1 Introduction

Online platforms depend heavily on automated toxicity classifiers to keep abusive, hateful, and harassing content in check. These systems work well on average, but a growing line of research shows they can go wrong in a specific and damaging way: they learn to treat identity terms themselves as signals of toxicity. When a classifier picks up a spurious association between, say, the word "Muslim" and a toxic label, it starts flagging perfectly benign posts that happen to mention Muslims. The result is that content written about or by marginalised communities gets removed at higher rates than equivalent content about dominant groups, which is exactly the opposite of what fair moderation should look like.

Most of the work on this problem has been done in English, where large annotated datasets and well-established evaluation protocols make fairness auditing relatively straightforward. Hindi and Hinglish toxicity detection, by contrast, has focused almost entirely on pushing accuracy numbers higher, with little attention paid to whether models treat different identity groups equitably. This matters a great deal in the Indian context, where online conversations around religion, caste, and politics are both consequential and heavily moderated. A classifier that systematically over-censors mentions of Dalits or Muslims, for instance, can amplify the very marginalisation it is meant to prevent.

We set out to fill this gap. Starting from XLM-RoBERTa fine-tuned on a publicly available Hinglish hate speech corpus, we conduct what we believe is the first comprehensive bias audit of a Hindi/Hinglish toxicity classifier. Our evaluation covers four identity categories — religion, caste, gender, and region — and uses three complementary fairness metrics: false positive rate disparity across groups, demographic parity difference, and a counterfactual fairness gap computed on sentence pairs that differ only in an identity term. The baseline model turns out to have substantial problems. Caste-mentioning non-toxic text is misclassified as toxic at a rate of 0.42, nearly three times the rate for gender-mentioning text.

We then test two widely discussed debiasing strategies. Counterfactual Data Augmentation (CDA) generates identity-swapped copies of training examples and trains on the union of original and augmented data. Adversarial debiasing adds a gradient reversal layer that discourages the en-

coder from representing identity information. Both operate under identical hyperparameters and are evaluated on the same test set and counterfactual pairs.

The headline result is that CDA is the stronger method on every fairness metric we measure. It cuts FPR disparity by 54% and demographic parity gap by 73%, at a cost of just 1.8 F1 points. The adversarial model also improves group-level fairness, but it does something unexpected: it actually makes counterfactual fairness *worse* than the baseline. This suggests that gradient reversal removes identity signal on average without making the model's per-sentence decisions robust to identity swaps — a distinction that matters for deployment.

Our contributions are:

1. A first systematic bias audit of Hindi/Hinglish toxicity classification across religion, caste, gender, and regional identities.

2. A reusable counterfactual test set of identity-swapped sentence pairs for evaluating per-instance fairness.

3. A controlled comparison of CDA and adversarial debiasing showing that CDA provides more comprehensive bias reduction with a smaller accuracy cost.

4. Open-source code, augmented training data, and model checkpoints to support reproducibility and future fairness research in Indian-language NLP.

2 Related Work

2.1 Fairness and Bias in NLP

Fairness in machine learning has been studied along several complementary axes. Group fairness criteria like demographic parity and equalized odds ask whether model outputs or error rates look roughly the same across protected groups. Individual and counterfactual fairness take a different angle, asking whether similar people — or the same person under a hypothetical identity change — get similar predictions. Early NLP work on representational harm revealed that word vectors and contextual representations encode gender and occupational stereotypes, leading to intrinsic bias metrics like WEAT and StereoSet and various debiasing post-processing methods. More recently, the field has shifted toward examining end-to-end systems, showing that even highly accurate models can produce large error-rate gaps across demographic groups. This shift makes clear that embedding-level analyses are not enough; task-level auditing is essential.

2.2 Bias in Toxicity and Hate-Speech Detection

A substantial body of work has documented unintended identity bias in English toxicity classifiers. Audits of both commercial APIs and research models have shown that neutral sentences containing identity terms are often scored as highly toxic, and that false positive and false negative rates can differ sharply across racial, religious, and LGBTQ+ groups. Researchers have introduced metrics like subgroup AUC and background-positive-subgroup-negative AUC to capture these disparities, and have built controlled counterfactual test sets where identity tokens are swapped while the surrounding context stays the same. The consistent finding is that toxicity classifiers are especially prone to picking up spurious correlations between identity mentions and toxic labels in training data, and that careful fairness evaluation before deployment is not optional.

2.3 Hindi and Hinglish Toxicity Detection

Research on Hindi and Hinglish hate speech detection has been focused primarily on dataset construction and classification accuracy. Shared tasks like HASOC have provided labelled Hindi tweets for hate/offensive versus non-offensive classification, and follow-up work has introduced hierarchical annotation schemes and larger corpora. Code-mixed Hindi–English toxicity detection has benefited substantially from multilingual transformers like XLM-RoBERTa, sometimes combined with CNN or BiLSTM layers to handle informal and mixed-script input. To our knowledge, however, none of these efforts has systematically examined whether models behave differently across religion, caste, gender, or region, and none has provided identity-aware evaluation suites or fairness metrics for Hindi/Hinglish toxicity.

2.4 Debiasing Methods for Text Classification

Two broad families of debiasing strategies have emerged. Data-centric approaches — most notably Counterfactual Data Augmentation (CDA) — generate additional training examples by swapping

identity terms while preserving labels, discouraging the model from treating specific identity tokens as predictive. Representation-centric approaches, including adversarial debiasing, add an auxiliary adversary network that tries to predict protected attributes from the main model's representations; training then minimises task loss while maximising adversary loss, pushing the encoder toward identity-agnostic features. Both strategies have shown measurable bias reductions in English toxicity and sentiment tasks, generally at a modest accuracy cost. What has not been done is a direct comparison of these two approaches on Hindi/Hinglish data with multiple fairness metrics, which is what we provide.

2.5 Summary and Positioning

Table 1 positions our work within the existing literature. Fairness audits to date have been confined to English and do not address the sociolinguistic landscape of Indian online discourse. We fill this gap by conducting the first bias audit of a Hindi/Hinglish toxicity classifier using multiple group-fairness and counterfactual metrics, constructing a reusable counterfactual test set targeting four identity categories, and directly comparing CDA and adversarial debiasing under matched conditions.

3 Methodology

Our goal is to measure and reduce demographic biases in Hindi/Hinglish toxicity classifiers, focusing on disparities across religion, caste, gender, and regional identities. The subsections below describe each component of the pipeline.

3.1 Dataset

We use the Kaggle Code-Mixed Hinglish Hate Speech corpus (`combined_hate_speech_dataset.csv`), an aggregated public release that draws from multiple Hindi–English social media sources. The corpus contains 29,550 samples with binary labels (hate vs. non-hate), split 46.4% hate and 53.6% non-hate — a reasonably balanced distribution. We partition the data into training, development, and test sets using a 70/10/20 stratified split with seed 42. Table 2 gives the exact counts.

We originally planned to include the HASOC 2021 shared task data, the Hindi Hate Speech Dataset (HHSD), and the Hinglish Offensive Language Dataset (HOLD). Access to HASOC was never granted, and HHSD and HOLD were dropped during the course of the project. All results reported in this paper therefore use the Kaggle corpus only. We acknowledge this as a limitation and discuss it in Section 6.

In addition to the main corpus, two resources are generated as part of our pipeline: the Counterfactual Augmented Hindi/Hinglish (CAHH) training set and the Counterfactual Fairness Test (CFT) set. Both are described in Section 3.5.1.

3.2 Preprocessing and Identity Annotation

We apply lightweight normalisation to all texts: Unicode normalisation, removal of URLs and user handles, and whitespace cleanup. Devanagari text is additionally processed with the `indic-nlp-library` for script-specific normalisation, while Romanised Hinglish tokens are left unchanged. All text is lowercased for lexicon matching.

We construct an identity lexicon covering four socially salient categories in the Indian context, shown in Table 3. Each sample is annotated with the set of matched identity terms and a boolean flag indicating whether at least one identity mention is present. These annotations are used for group-wise error analysis and counterfactual example generation.

Table 4 shows how identity mentions are distributed across the corpus. About 23.6% of samples contain at least one identity term. Gender is the most common category, appearing in 16% of samples, followed by religion at 6%. Caste and region are much rarer — caste terms appear in only 1.4% of samples and region in just 0.2% (63 samples total). The region subgroup is too small for reliable statistical estimates, which we flag in the limitations.

3.3 Baseline Toxicity Classifier

Our baseline classifier fine-tunes `xlm-roberta-base` on the binary toxicity task using standard cross-entropy loss and AdamW optimisation. Input texts are tokenised with the XLM-RoBERTa subword tokeniser and truncated or padded to a fixed maximum sequence length of 128 tokens. The model adds a single linear classification head on top of the `[CLS]` representation. No identity signal is used during baseline training. Table 5 lists all hyperparameters, which are held constant across the baseline, CDA, and adversarial models to ensure a fair comparison.

Table 1: Representative prior work on toxicity/hate-speech fairness and Hindi/Hinglish detection. ✓ = addressed; – = not addressed; * = this work. CFT = counterfactual test set; CDA = counterfactual data augmentation.

Work	Language	Identity Categories	CFT Set	Fairness Metrics	Debiasing Method
Dixon et al. (2018)	English	Race, Religion, Gender	✓	Subgroup AUC, FPR	CDA
Borkan et al. (2019)	English	Race, Religion, LGBTQ+	✓	Subgroup AUC, BPSN	None
Zhao et al. (2018)	English	Gender	–	FPR/FNR disparity	CDA
Zhang et al. (2018)	English	Gender, Race	–	Demographic parity	Adversarial
HASOC (2019–21)	Hindi / Hinglish	None	–	Accuracy, F1	None
Mandl et al. (2021)	Hindi	None	–	Macro-F1	None
Kumar et al. (2021)	Hinglish	None	–	Accuracy, F1	None
Velankar et al. (2021)	Hindi	None	–	Accuracy, F1	None
Ours*	**Hindi / Hinglish**	**Religion, Caste, Gender, Region**	✓	**FPR Disp., Δ_{DP}, CFT Gap**	**CDA + Adversarial**

Table 2: Dataset split statistics. The corpus is an aggregated Hinglish hate speech collection from Kaggle.

Split	Samples	Hate (1)	Non-Hate (0)
Train	20,684	9,607	11,077
Dev	2,955	1,373	1,582
Test	5,911	2,745	3,166
Total	**29,550**	**13,725**	**15,825**

Table 3: Identity lexicon categories and example terms (Hindi/Hinglish).

Category	Example Terms
Religion	Hindu, Muslim, Christian, Sikh, Jain, Buddhist; *musalman, isai, baudh*
Caste	Dalit, Brahmin, OBC, Vaishya, Kshatriya; *savarna, achhoot*
Gender	woman, man, *mahila, purush, aurat, ladki, ladka*
Region	North Indian, South Indian, Bihari, Punjabi, Madrasi; *bhaiya, madrasi*

3.4 Bias Metrics

We measure bias using three complementary metrics, summarised in Table 6.

False Positive Rate (FPR) Disparity. For each identity group g, we compute the false positive rate on truly non-toxic examples:

$$\mathrm{FPR}(g) = \frac{\mathrm{FP}_g}{\mathrm{FP}_g + \mathrm{TN}_g}.$$

The disparity is the range across groups:

$$\text{FPR-Disparity} = \max_g \mathrm{FPR}(g) - \min_g \mathrm{FPR}(g).$$

A large value means that non-toxic content about some groups gets incorrectly flagged far more often than content about others.

Demographic Parity Difference (Δ_{DP}).

$$\Delta_{\mathrm{DP}} = \max_g P(\hat{Y}{=}1 \mid g) - \min_g P(\hat{Y}{=}1 \mid g),$$

Table 4: Identity term coverage in the full corpus.

Group	Samples	% of corpus
Gender	4,722	16.0%
Religion	1,777	6.0%
Caste	411	1.4%
Region	63	0.2%
Any identity	6,973	23.6%
None	22,577	76.4%

Table 5: Training hyperparameters (shared across all three models).

Hyperparameter	Value
Pretrained checkpoint	`xlm-roberta-base`
Max sequence length	128 tokens
Batch size	32
Learning rate	2e-5
Epochs	5
Warmup ratio	0.1
Optimiser	AdamW (β_1=0.9, β_2=0.98)
Weight decay	0.01
Gradient clipping	1.0
Random seed	42
Best-model selection	Highest dev accuracy
Loss function	Cross-entropy

where $\hat{Y}$ is the model prediction. This captures whether some groups are classified as toxic at substantially higher rates than others, regardless of the ground-truth label.

Counterfactual Fairness Gap (CFT Gap). For each counterfactual pair (x, x') differing only in an identity term:

$$\text{Gap} = \left| s(x) - s(x') \right|,$$

where $s(\cdot)$ is the predicted toxicity probability. We report the mean over all pairs. A high value means the model's output is sensitive to which identity term appears, even when nothing else about the sentence has changed.

Table 6: Bias metrics at a glance.

Metric	What It Measures	Ideal
FPR Disparity	Spread in false-positive rates across groups	0
Δ_{DP}	Spread in positive-prediction rates across groups	0
CFT Gap	Avg. score change from swapping one identity term	0

3.5 Debiasing Methods

We investigate two debiasing strategies applied on top of the baseline model.

3.5.1 Counterfactual Data Augmentation (CDA)

For each training example that contains an identity term, we generate an additional example by swapping that term with another term from the same lexicon category while keeping the original label. For instance:

> *Original:* "Those Muslims are criminals." (*toxic*)
>
> *Augmented:* "Those Hindus are criminals." (*toxic*)

We apply this procedure across all four identity categories (religion, caste, gender, region) with a 1:1 augmentation ratio, producing the CAHH corpus. The augmented data are merged with the original training set and used to fine-tune XLM-RoBERTa under the same hyperparameters as the baseline. The intuition is straightforward: by showing the model that the same sentence is toxic (or not) regardless of which identity term appears, we break the spurious correlation between identity tokens and the toxicity label.

During this stage, we also generate the CFT Test Set — roughly 800 counterfactual identity-swapped pairs drawn from the original data — which is used exclusively for evaluation. No CFT pair overlaps with any training example.

3.5.2 Adversarial Debiasing

We extend the baseline with a two-head architecture. The first head predicts toxicity as before. The

second head is an adversary that tries to predict the identity group (4-way classification: religion, caste, gender, region) from the encoder's intermediate representation. A Gradient Reversal Layer (GRL) sits between the encoder and the adversary, flipping the gradient sign during backpropagation. The combined objective is:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{tox}} - \lambda \mathcal{L}_{\text{adv}},$$

where $\mathcal{L}_{\text{tox}}$ is cross-entropy for toxicity, $\mathcal{L}_{\text{adv}}$ is cross-entropy for identity-group prediction, and $\lambda = 0.5$ controls the debiasing strength. The adversary has a hidden dimension of 128. By penalising the encoder for making identity information accessible, this setup pushes the learned representations toward identity invariance — at least in principle.

3.6 Evaluation Protocol

We evaluate three systems, summarised in Table 7, on two fronts. For utility, we report accuracy, precision, recall, F1, and macro-F1 on the full held-out test set. For fairness, we compute FPR disparity, Δ_{DP}, and CFT Gap on the identity-containing subset of the test set and on the CFT Test Set of identity-swapped pairs.

Table 7: Systems compared in experiments.

System	Description	Key Tunable
Baseline	XLM-RoBERTa fine-tuned on original data	LR, epochs
CDA	Fine-tuned on original + CAHH augmented data	Augment. ratio
Adversarial	Baseline + identity-adversary head (GRL)	λ

3.7 Implementation Details

All models are implemented with the HuggingFace `transformers` library and trained on a single GPU (Apple Silicon MPS; the code also supports CUDA and CPU). Best checkpoints are selected by highest development-set accuracy. We publicly release our code, trained model checkpoints, the CAHH augmentation corpus, and the CFT Test Set.[1]

[1]Code and data: https://github.com/DevanshSrajput/Hindi-Toxicity-Bias

4 Results

4.1 Overall Performance and Fairness

Table 8 presents the main results. All three models achieve similar overall accuracy (0.73–0.75), but they differ sharply on fairness.

The baseline achieves the highest F1 (0.703) but also the worst fairness scores across the board, with an FPR disparity of 0.421 and a demographic parity gap of 0.505. CDA brings FPR disparity down to 0.196 (a 54% reduction) and demographic parity gap to 0.136 (a 73% reduction), while losing only 1.8 F1 points. The adversarial model falls between the two on group-level metrics — FPR disparity of 0.256, DP gap of 0.294 — and preserves more accuracy than CDA, losing just 0.4 F1 points.

The most striking result concerns counterfactual fairness. CDA reduces the CFT Gap from 0.052 to 0.042, a 19% improvement. The adversarial model, however, pushes it in the wrong direction: from 0.052 to 0.068, a 31% worsening. We return to this finding in the Discussion.

4.2 Per-Group False Positive Rates

Table 9 breaks down false positive rates by identity group. In the baseline, caste-mentioning text has by far the highest FPR at 0.421, followed by religion at 0.283 and gender at 0.151. Region has an FPR of 0.000, but this number is meaningless given only 63 region-tagged samples in the entire corpus.

CDA brings the caste FPR down from 0.421 to 0.342, the largest absolute reduction for any group. The adversarial model makes a smaller improvement on caste (to 0.395) but achieves the lowest religion FPR (0.249) and the lowest gender FPR (0.139). Both debiased models assign a non-zero FPR to the region group, which had been zero in the baseline; however, as noted, this subgroup is too small to draw conclusions from.

5 Discussion

5.1 CDA Wins on Every Fairness Metric

The core result is that CDA outperforms adversarial debiasing on all three fairness measures. By directly exposing the model to identity-swapped variants of training examples, CDA breaks the spurious association between identity tokens and the toxicity label at the level of the training data itself. The model learns that a sentence is toxic or not regardless of whether it says "Hindu" or "Muslim," "Dalit" or "Brahmin." This is reflected not just in

Table 8: Test set results ($n = 5{,}911$). **Bold** = best per column. ↓ indicates lower is better.

Model	Acc	Prec	Rec	F1	Macro F1	FPR Δ ↓	DP Δ ↓	CFT Gap ↓
Baseline	0.749	0.778	0.642	**0.703**	0.743	0.421	0.505	0.052
CDA	0.734	0.761	0.624	0.685	0.728	**0.196**	**0.136**	**0.042**
Adversarial	0.747	0.781	0.633	0.699	0.741	0.256	0.294	0.068

Table 9: Per-group false positive rates on the test set. Region values are unreliable ($n = 63$).

Group	Baseline	CDA	Adversarial
Religion	0.2825	0.2825	0.2486
Caste	0.4211	0.3421	0.3947
Gender	0.1505	0.1465	0.1386
Region	0.0000	0.1667	0.1667

the group-level metrics (FPR disparity down 54%, DP gap down 73%) but also in the per-instance counterfactual gap (down 19%). The tradeoff is small: 1.8 F1 points, which for most real-world moderation settings is a cost worth paying.

5.2 Adversarial Debiasing Has a Hidden Failure Mode

The adversarial model tells a more complicated story. On group-level metrics — FPR disparity and DP gap — it improves over the baseline, though not as much as CDA. But on counterfactual fairness, it actually gets *worse*: the CFT Gap rises from 0.052 to 0.068, meaning that for any given sentence, swapping the identity term changes the model's predicted toxicity score by more than it did before debiasing.

This makes sense if you think about what gradient reversal actually does. The GRL pushes the encoder to produce representations from which a linear adversary cannot predict the identity group. But this is a statistical constraint on the representation *distribution* across the training set — it does not guarantee that any individual sentence's prediction is invariant to identity swaps. The encoder can learn to remove identity signal on average while still encoding it in ways that affect per-instance decisions, especially for sentences where identity terms interact with surrounding context. We call this *shallow decorrelation*: the aggregate statistics look better, but the model has not actually learned to ignore identity when making decisions about specific inputs.

This is arguably the most important finding of the paper. It suggests that practitioners evaluating debiasing methods should not rely on group-level metrics alone. A method can look good on FPR disparity or demographic parity while making counterfactual fairness worse, which means the model is still making identity-sensitive decisions — just in a less statistically detectable way.

5.3 Caste Is the Most Biased Subgroup

The baseline FPR for caste-mentioning text is 0.421 — nearly three times the FPR for gender (0.151) and substantially higher than religion (0.283). This means that almost half of all non-toxic text mentioning caste terms is incorrectly classified as toxic. Caste-based discrimination is a persistent issue in Indian society, and the finding that toxicity classifiers amplify this by disproportionately censoring caste-related content is concerning. CDA brings the caste FPR down to 0.342, a meaningful improvement though still the highest among all groups. This highlights the need for more caste-aware data and evaluation in Indian-language NLP, which remains understudied relative to religion and gender.

5.4 The Fairness–Utility Tradeoff Is Favourable

A common concern about debiasing methods is that they harm overall performance. Our results suggest this concern is overstated, at least in this setting. CDA achieves a 54% reduction in FPR disparity at a cost of 1.8 F1 points — from 0.703 to 0.685. The adversarial model loses even less (0.4 F1 points) but buys less fairness improvement. For deployment scenarios where unfair false positives on identity-mentioning content carry real harm — silencing marginalised voices, reinforcing existing power imbalances — a 1.8-point accuracy trade is

clearly worthwhile.

6 Limitations

Several limitations should be kept in mind when interpreting these results. First, all experiments use a single dataset — the Kaggle Code-Mixed Hinglish Hate Speech corpus. We originally planned to include HASOC, HHSD, and HOLD, but access issues and scope constraints meant we could not. Whether our findings generalise to other Hindi/Hinglish corpora, other code-mixed language pairs, or other domains remains an open question.

Second, the region subgroup contains only 63 samples (0.2% of the corpus), which is far too few for reliable bias estimates. We report region-level numbers for completeness but caution against interpreting them.

Third, our identity detection is lexicon-based. It catches explicit mentions but misses implicit references, slang, sarcasm, and informal spellings. This introduces selection bias toward samples with overt identity terms and likely undercounts the true extent of identity-related content in the data.

Fourth, the corpus uses a binary toxicity label (hate vs. non-hate) with no distinction between hate speech, casual profanity, and offensive humour. More fine-grained labelling might reveal bias patterns that a binary setup obscures.

Fifth, all results come from a single random seed. Multi-seed runs with statistical significance testing would make the claims more robust.

Sixth, the CFT test set is generated by template-based identity swapping. The quality of counterfactual pairs depends on lexicon coverage, and some swaps may not produce fully fluent or natural-sounding Hinglish. More sophisticated generation methods, potentially involving human validation, would strengthen the counterfactual evaluation.

7 Conclusion

We have presented the first systematic bias audit of a Hindi/Hinglish toxicity classifier, covering religion, caste, gender, and regional identities with three complementary fairness metrics. The baseline XLM-RoBERTa model exhibits substantial identity-correlated bias, particularly against caste-mentioning content, which faces a false positive rate nearly three times that of gender-mentioning content.

Comparing two debiasing strategies under matched conditions, we find that Counterfactual Data Augmentation is the more effective and more reliable method. It reduces FPR disparity by 54% and demographic parity gap by 73% at a modest cost of 1.8 F1 points, and it improves counterfactual fairness as well. Adversarial debiasing improves group-level fairness metrics but worsens per-instance counterfactual fairness, revealing that gradient reversal can produce shallow decorrelation rather than genuine identity invariance. This finding has practical implications: debiasing methods should be evaluated at multiple levels of granularity, and group-level improvements alone are not sufficient evidence that a model has actually become fair.

We release our code, the CAHH augmented corpus, the CFT test set, and trained checkpoints to support future fairness research in Indian-language NLP.

Acknowledgments

We thank the creators of the Kaggle Code-Mixed Hinglish Hate Speech Detection Dataset for making their data publicly available.

References

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion Proceedings of The 2019 World Wide Web Conference*, pages 491–500.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

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.

Yaroslav Ganin and Victor Lempitsky. 2015. Unsupervised domain adaptation by backpropagation. In *International Conference on Machine Learning*, pages 1180–1189.

Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, pages 3315–3323.

Ritesh Kumar, Atul Kr. Ojha, Shervin Malmasi, and Marcos Zampieri. 2021. Benchmarking aggression identification in social media. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying*, pages 1–11.

Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. 2020. Gender bias in neural natural language processing. In *Logic, Language, and Security*, pages 189–202. Springer.

Thomas Mandl, Sandip Modha, Gautam Kishore Shahi, Hiren Madhu, Shrey Satapara, Prasenjit Majumder, Johannes Schäfer, Tharindu Ranasinghe, Marcos Zampieri, Durgesh Nandini, and Amit Kumar Jaiswal. 2021. Overview of the HASOC track at FIRE 2021: Hate speech and offensive content identification in English and Indo-Aryan languages. In *Forum for Information Retrieval Evaluation*, pages 1–3.

Alex Beutel, Jilin Chen, Zhe Zhao, and Ed H. Chi. 2017. Data decisions and theoretical implications when adversarially learning fair representations. *arXiv preprint arXiv:1707.00075*.

Abhishek Velankar, Hrushikesh Patil, and Raviraj Joshi. 2021. Hate and offensive speech detection in Hindi and Marathi. *arXiv preprint arXiv:2110.12200*.

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

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 15–20.