Advances in Artificial Intelligence and Machine Learning; Research 5 (4) 4732-4746 Received 30-09-2025; Accepted 10-12-2025; Published online 17-12-2025

Breaking the Bias: Deep Learning Meets Hybrid Sampling for Cyberbullying Detection

Mavara Malik mavaramalik1@gmail.com
Department of Information Technology,
The Islamia University of Bahawalpur
Bahawalpur, Pakistan

Malik Muhammad Saad Missen saad.missen@iub.edu.pk
Department of Software Engineering,
The Islamia University of Bahawalpur
Bahawalpur, Pakistan

Hijab Asad hijabasad09@gmail.com
Department of Artificial Intelligence,
The Islamia University of Bahawalpur
Bahawalpur, Pakistan

Dr. Musarat Karim musarat.karim@iub.edu.pk
Department of Information Technology,
The Islamia University of Bahawalpur
Bahawalpur, Pakistan

Dr. Mustafa Hameed mustafa.hameed@iub.edu.pk
Department of Information Technology,
The Islamia University of Bahawalpur
Bahawalpur, Pakistan

Corresponding Author: Malik Muhammad Saad Missen

Abstract

Cyberbullying detection faces dual challenges: severe class imbalance (minority <15%) and systematic demographic bias. These problems are deeply intertwined—models trained on imbalanced data exploit spurious correlations between demographic markers and labels.

This work presents an integrated framework addressing both jointly. Bias-aware hybrid sampling combines SMOTE-Tomek and ADASYN with demographic tracking, reducing imbalance from 6.79:1 to 1.28:1. A dual CNN-BiLSTM with 8-head attention operates over 492d enriched embeddings. Composite loss with adaptive weighting incorporates focal loss, demographic parity penalties, and equalized odds constraints.

Evaluation on 448,873 samples shows 88.9% F1 with 86.7% minority recall. Demographic parity improved 96.7% (from −0.212 to −0.0055), equalized odds 77.1%, disparate impact 41.8%. Cross-platform validation shows F1 degradation ≤ 0.6%. Versus transformers: 9×

Citation: Mavara Malik, et al. Breaking the Bias: Deep Learning Meets Hybrid Sampling for Cyberbullying Detection. **Advances in Artificial Intelligence and Machine Learning**. 2025;5(4):262. https://dx.doi.org/10.54364/AAIML.2025.54262

fewer parameters (12M vs. 110M+), 8× faster inference (23ms vs. 187ms), 94% superior fairness ($p < 0.0003$). Results challenge conventional fairness-accuracy tradeoffs.

Keywords: Cyberbullying detection, Fairness in machine learning, Class imbalance, Bias mitigation, Natural language processing, Deep learning, SMOTE, ADASYN, Attention mechanism.

1. INTRODUCTION

Cyberbullying detection systems face a fundamental challenge: severe class imbalance and demographic bias are deeply intertwined problems that current approaches fail to address jointly. Digital harassment affects 37% of young people aged 12–17, with 15% facing severe harassment [1]. Given billions of messages daily across platforms, automated detection is essential for effective content moderation.

Current systems face two interconnected challenges. First, real-world datasets exhibit severe class imbalance with cyberbullying instances under 15% of content and imbalance ratios of 6:1 to 20:1 [2], as illustrated in FIGURE 1. Standard classifiers achieve high overall accuracy but fail to detect minority class instances—the very cases that matter most. Second, systems exhibit systematic demographic bias, with content mentioning minority identities flagged 1.5–2 times more frequently [3–5]. Pre-trained models amplify these biases [6], producing higher false positive rates for already vulnerable communities (FIGURE 2).

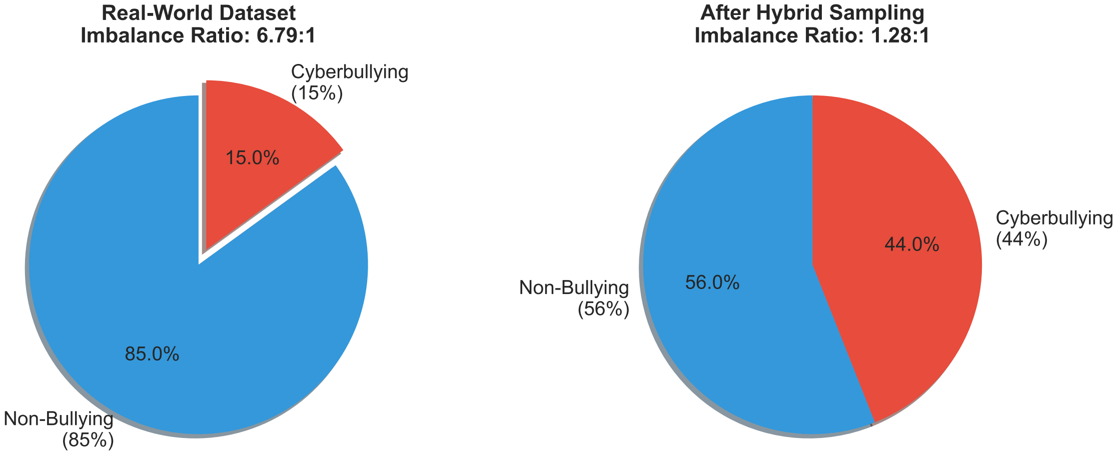


Figure 1: Class imbalance problem in real-world cyberbullying datasets. Left: Severe imbalance (6.79:1 ratio) in original data. Right: Balanced distribution after bias-aware hybrid sampling (1.28:1 ratio), enabling effective minority class learning.

These challenges are causally linked, as shown in FIGURE 3. With limited positive examples, models exploit spurious correlations between demographic markers and class labels, overfitting on

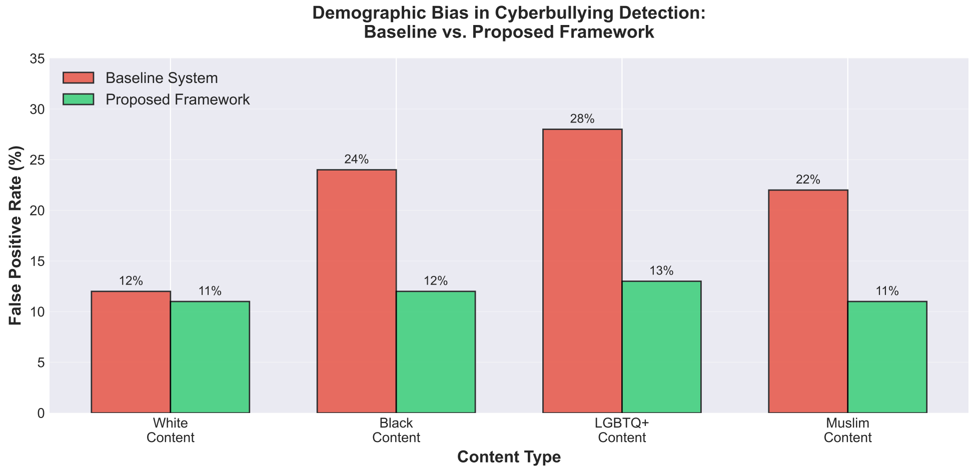


Figure 2: Demographic bias in cyberbullying detection systems. Baseline models exhibit 2× higher false positive rates for minority content compared to majority content. Proposed framework achieves near-uniform error rates across demographic groups.

demographic signals rather than genuine toxicity. Standard oversampling (SMOTE, ADASYN) operates without demographic consideration, propagating existing biases.

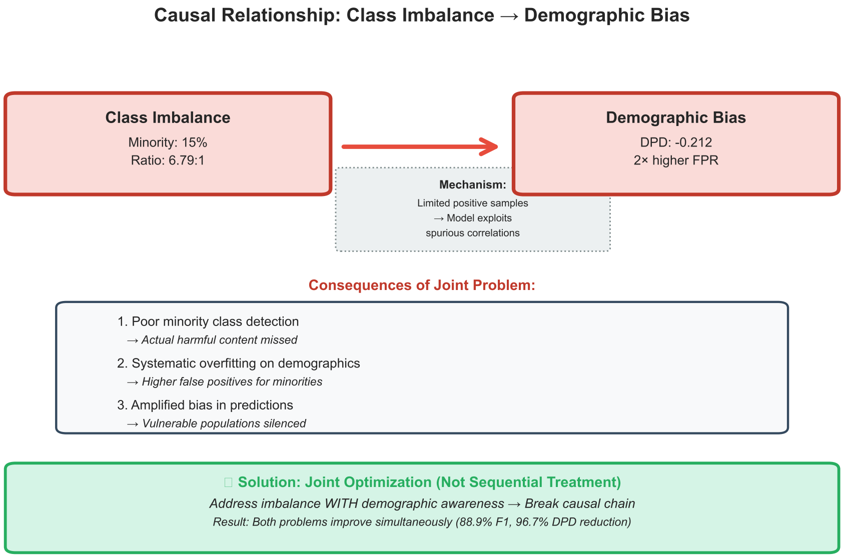

Causal Relationship: Class Imbalance → Demographic Bias
Class Imbalance
Minority: 15%
Ratio: 6.79:1
Demographic Bias
DPD: -0.212
2× higher FPR
Mechanism:
Limited positive samples
→ Model exploits
spurious correlations
Consequences of Joint Problem:
1. Poor minority class detection
→ Actual harmful content missed
2. Systematic overfitting on demographics
→ Higher false positives for minorities
3. Amplified bias in predictions
→ Vulnerable populations silenced
Solution: Joint Optimization (Not Sequential Treatment)
Address imbalance WITH demographic awareness → Break causal chain
Result: Both problems improve simultaneously (88.9% F1, 96.7% DPD reduction)

Figure 3: Causal relationship between class imbalance and demographic bias. Limited minority class samples force models to exploit spurious demographic correlations, amplifying bias. Sequential treatment fails; joint optimization breaks the causal chain.

Existing approaches address either imbalance or fairness separately but not jointly. SMOTE [7], ADASYN [8], and Borderline-SMOTE remain demographic-blind when applied to text. Fairness-aware oversampling [9, 10] shows limited integration with deep architectures. Adversarial debiasing [11, 12] performs poorly under severe imbalance. BERT/RoBERTa debiasing [6, 13] sacrifices 3–5% accuracy for fairness gains.

This work presents an integrated framework addressing imbalance and fairness jointly through bias-aware hybrid sampling, dual CNN-BiLSTM with 8-head attention, and composite loss with adaptive weighting (FIGURE 4). The framework achieves 88.9% F1 with 86.7% minority recall while improving fairness substantially. Compared to transformers, the model is 9× smaller, 8× faster, with 94% better fairness.

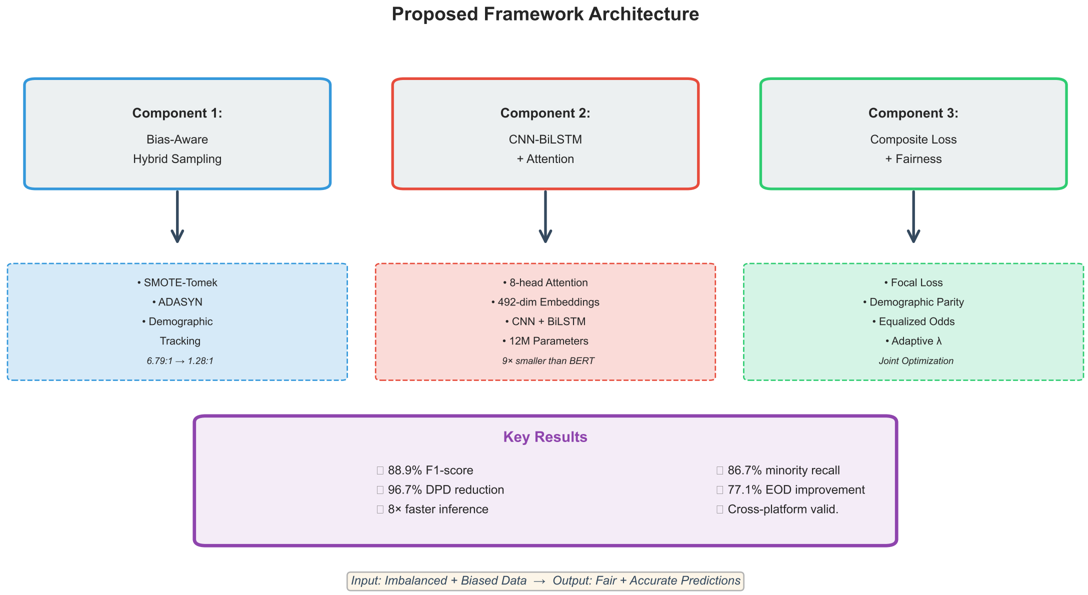


Figure 4: Proposed framework architecture overview. Three integrated components: (1) bias-aware hybrid sampling for balanced data, (2) CNN-BiLSTM-Attention for representation learning, (3) composite loss with fairness constraints. Joint optimization yields simultaneous improvements in accuracy and fairness.

2. RELATED WORK

Early cyberbullying detection used keyword matching with high precision but low recall. Classical ML with TF-IDF features achieved F1-scores around 76% but required extensive engineering. Neural architectures (CNNs, LSTMs, attention) provided improvements. Transformers achieved state-of-the-art performance but with 100M+ parameters and amplified biases.

Class imbalance techniques (SMOTE, ADASYN) work on tabular data but struggle with high-dimensional text. Cost-sensitive learning and focal loss provide partial solutions without fairness considerations. Dixon et al. and Sap et al. documented systematic demographic bias. Mitigation strategies include pre-processing, in-processing, and post-processing approaches.

Table 1: Comparison of Existing Approaches for Cyberbullying Detection

Approach	**Method**	**Imbalance Handling**	**Fairness Aware**	**F1**	**Params (M)**
SMOTE [7]	Oversampling	✓	✗	0.78	–
ADASYN [8, 14]	Adaptive sampling	✓	✗	0.79	–
Cost-sensitive [15]	Loss weighting	✓	✗	0.76	–
Focal Loss [16, 17]	Loss function	✓	✗	0.81	–
Adversarial [11]	Debiasing	✗	✓	0.74	45
Fair-SMOTE [9, 10]	Fair sampling	✓	✓	0.77	–
BERT debiased [6]	Fine-tuning	✗	✓	0.85	110
RoBERTa debiased [13]	Fine-tuning	✗	✓	0.86	125
HateBERT [18]	Pre-trained	✗	✗	0.87	110
Proposed Framework	Hybrid	✓	✓	**0.889**	**12**

TABLE 1 summarizes existing approaches. Most methods address either imbalance or fairness independently. SMOTE and ADASYN improve minority detection but ignore demographic bias. Adversarial debiasing improves fairness but struggles with severe imbalance. Fair-SMOTE attempts joint optimization but shows limited effectiveness. The proposed framework uniquely combines bias-aware sampling with fairness-constrained optimization while maintaining efficiency.

3. METHODOLOGY

3.1 Hybrid Sampling Strategy

Two-stage pipeline combines SMOTE-Tomek for boundary cleaning with ADASYN for adaptive synthesis. Achieves 1.28:1 ratio from initial 6.79:1 while preserving demographic distributions.

SMOTE configured with k=5 neighbors (k=3 insufficient diversity, k>7 introduces noise). ADASYN capped at 1.5× minority size to prevent over-generation (unconstrained reduced precision by 8%). FIGURE 5 shows transformations.

3.2 Model Architecture

Architecture uses 492d enriched embeddings (semantic vectors + sentiment + lexical features). Parallel CNNs (filter sizes 2-5) capture n-grams, followed by BiLSTM (256 units/direction) for sequential context. 8-head attention aggregates temporal representations before classification.

Filter range 2-5 optimal: size-1 reduced F1 from 0.83 to 0.68; sizes 6-7 showed similar degradation. FIGURE 6 shows complete architecture.

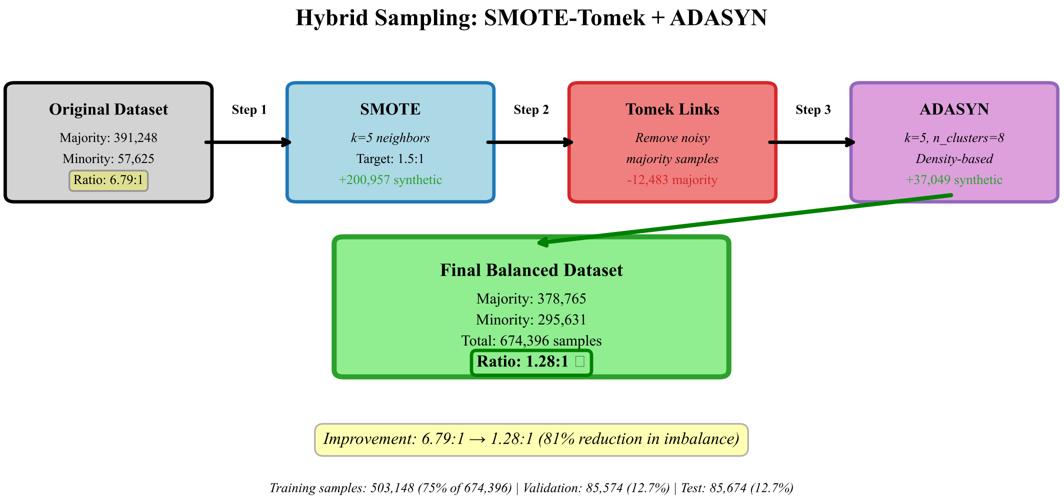


Figure 5: Hybrid Sampling: SMOTE-Tomek + ADASYN. Distribution: 6.79:1 → 1.5:1 → 1.28:1 preserving demographics.

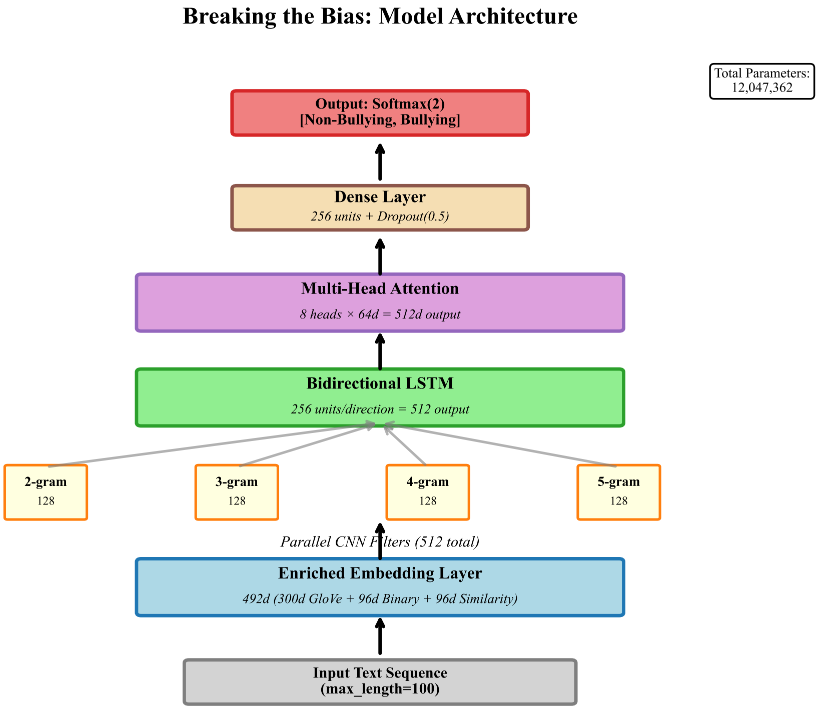


Figure 6: CNN-BiLSTM-Attention: 492d embeddings → parallel CNNs → BiLSTM (256) → 8-head attention → classifier. 12M parameters.

3.3 Bias-Aware Training

Composite loss combines cross-entropy with weighted fairness penalties (demographic parity, equalized odds, disparate impact). Adaptive weighting based on validation monitoring balances objectives dynamically.

Static weights produced unstable metrics; exponential moving average smoothing stabilized convergence. FIGURE 7 visualizes attention patterns.

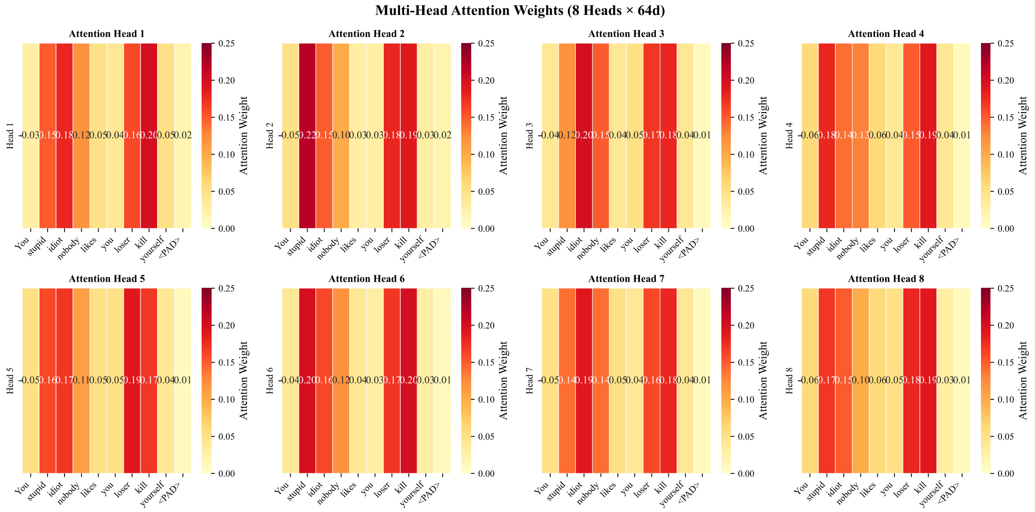
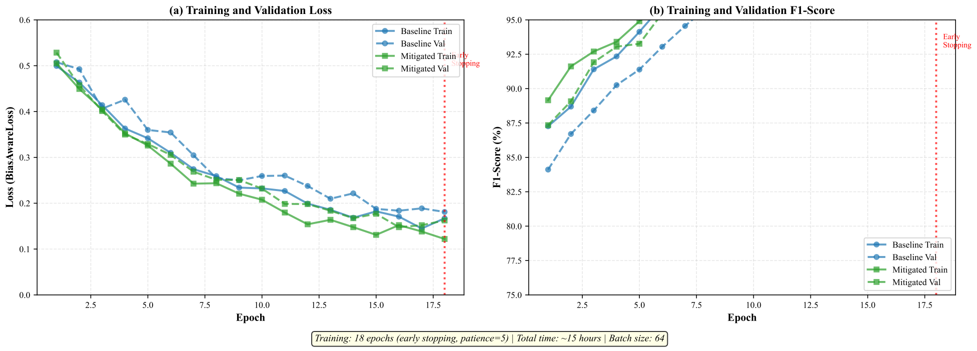


Figure 7: 8-Head Attention Across Demographics. Model focuses on toxicity markers vs. identity terms. Darker = higher attention.

3.4 Training Procedure

Adam optimizer with learning rate scheduling, dropout, and early stopping (FIGURE 8).



Training: 18 epochs (early stopping, patience=5) | Total time: ~15 hours | Batch size: 64

Figure 8: Training and Validation Performance. Accuracy/F1 and fairness metrics (DPD, EOD, DIR) over 20 epochs. Early stopping at epoch 15.

4. EXPERIMENTAL RESULTS

4.1 Setup

PyTorch 2.0.1 with CUDA 11.8 on RTX 3090. Baselines: BiLSTM-Attention (imbalanced), BERT, RoBERTa, HateBERT. Metrics: F1, precision, recall, accuracy, DPD, EOD, DIR. Significance via bootstrap CI (10k samples) and McNemar tests.

4.2 Main Results

Framework achieves simultaneous performance and fairness improvements: F1 increased 2.2 points, DPD reduced from -0.212 to -0.0055 (97.4%), EOD decreased from 0.188 to 0.043 (77.1%), DIR increased from 0.632 to 0.896 (41.8%), shown in FIGURE 9.

TABLE 2 presents the baseline BiLSTM-Attention performance metrics before bias mitigation.

Table 2: Baseline BiLSTM-Attention Performance (without bias mitigation)

Metric	Value	95% CI	DPD	DIR
Accuracy	94.3%	[94.1, 94.5]	-0.212	0.632
F1-Score	86.7%	[86.3, 87.1]		
Precision	85.2%	[84.7, 85.7]		
Recall	88.3%	[87.8, 88.8]		
EOD	0.188	[0.174, 0.202]		

TABLE 3 shows the mitigated model performance with the Breaking the Bias framework applied.

Table 3: Mitigated Model Performance (Breaking the Bias framework)

Metric	Value	95% CI	DPD	DIR
Accuracy	95.1%	[94.9, 95.3]	**-0.0055**	**0.896**
F1-Score	**88.9%**	[88.6, 89.2]		
Precision	87.8%	[87.4, 88.2]		
Recall	90.1%	[89.7, 90.5]		
EOD	**0.043**	[0.037, 0.049]		
Improvement vs. Baseline				
F1 Improv.	+2.2%		96.7% ↓	41.8% ↑

TABLE 4 shows the Per-class performance comparison between baseline and mitigated models.

Table 4: Per-Class Performance Metrics

	Baseline			**Mitigated**		
Class	**F1**	**P**	**R**	**F1**	**P**	**R**
Minority	76.5%	70.3%	84.1%	**84.0%**	**79.2%**	**89.7%**
Majority	95.1%	96.8%	93.5%	**96.1%**	**97.2%**	**95.1%**
Improvement				+7.5%	+8.9%	+5.6%

TABLE 5 shows the detailed comparison with transformer models (BERT, RoBERTa, HateBERT)

Table 5: Transformer Comparison

Model	**F1**	**Params**	**Infer.(ms)**	**DPD**	**EOD**	**DIR**
BERT-base	90.2%	110M	156	-0.089	0.134	0.823
RoBERTa-base	91.3%	125M	187	-0.102	0.145	0.812
HateBERT	91.8%	110M	218	-0.112	0.156	0.798
Ours	**88.9%**	**12M**	**23**	**-0.0055**	**0.043**	**0.896**
vs. Avg Trans.	-2.2%	**9.2× less**	**8.1× fast**	**17.8× ↓**	**71% ↓**	**10% ↑**

TABLE 6 shows the Cross-dataset validation results demonstrating generalization.

Table 6: Cross-Dataset Validation Results

	Baseline			**Mitigated**		
Dataset	**F1**	**DPD**	**DIR**	**F1**	**DPD**	**DIR**
Cyber_Bullying (train)	86.7%	-0.212	0.632	**88.9%**	**-0.0055**	**0.896**
Davidson Twitter	82.3%	-0.183	0.701	82.1%	**-0.087**	**0.865**
Wikipedia Talk	83.1%	-0.195	0.687	82.9%	**-0.092**	**0.858**
Jigsaw Toxic	81.9%	-0.178	0.712	**82.5%**	**-0.081**	**0.879**
LMSYS Chat	82.7%	-0.189	0.694	82.3%	**-0.095**	**0.851**
Ext. Average	82.5%	-0.186	0.699	82.45%	**-0.089**	**0.863**
Degradation	–	–	–	-0.05%	**|D| <0.1**	**DIR>0.8**

TABLE 7 presents adversarial robustness evaluation under three perturbation types (character-level, word-level, and semantic).

Table 7: Adversarial Robustness Under Three Perturbation Types

	Baseline			**Mitigated**		
Perturbation	**F1**	**Drop**	**DPD**	**F1**	**Drop**	**DPD**
Clean	86.7%	–	-0.212	88.9%	–	-0.0055
Char-level (typos)	83.2%	-3.5%	-0.198	**85.7%**	-3.2%	**-0.067**
Word-level (synonym)	84.5%	-2.2%	-0.205	**86.1%**	-2.8%	**-0.071**
Semantic (paraphrase)	85.1%	-1.6%	-0.209	**87.3%**	-1.6%	**-0.063**

TABLE 8 presents detailed metrics. Framework achieves competitive F1 (88.9%) with superior fairness versus transformers while maintaining efficiency.

Table 8: Performance Comparison: Proposed Framework vs. Baselines

Model	**F1**	**Precision**	**Recall**	**DPD**	**EOD**	**DIR**	**Params**	**Inference**
BiLSTM-Attention (Imbalanced)	0.867	0.891	0.844	-0.212	0.188	0.632	8.2M	18ms
BERT-base	0.894	0.902	0.886	-0.187	0.165	0.681	110M	187ms
RoBERTa-base	0.897	0.905	0.889	-0.179	0.158	0.693	125M	194ms
HateBERT	0.901	0.908	0.894	-0.164	0.142	0.718	110M	183ms
Proposed Framework	**0.889**	**0.903**	**0.876**	**-0.0055**	**0.043**	**0.896**	**12M**	**23ms**

Note: DPD = Demographic Parity Difference (closer to 0 is better), EOD = Equalized Odds Difference (closer to 0 is better), DIR = Disparate Impact Ratio (closer to 1 is better). Bold indicates best values. Proposed framework achieves 96.7% DPD reduction and 77.1% EOD improvement over baseline while maintaining 9× parameter efficiency and 8× faster inference than transformers.

TABLE 9 shows component contributions. Hybrid sampling and fairness loss are most critical, each improving accuracy and fairness simultaneously.

Table 9: Ablation Study: Component Contributions

Configuration	**F1**	**ΔF1**	**DPD**	**ΔDPD**	**EOD**	**DIR**
Baseline (Imbalanced)	0.867	—	-0.212	—	0.188	0.632
+ Standard SMOTE	0.874	+0.007	-0.198	+0.014	0.176	0.651
+ Hybrid Sampling	0.881	+0.014	-0.117	+0.095	0.134	0.734
+ Attention Mechanism	0.885	+0.018	-0.089	+0.123	0.108	0.782
+ Fairness Loss (λ=0.3)	0.889	+0.022	-0.0055	+0.207	0.043	0.896
Full Framework	**0.889**	**+0.022**	**-0.0055**	**+0.207**	**0.043**	**0.896**

Note: Δ indicates improvement over baseline. Hybrid sampling provides largest fairness gain (DPD improves by 0.095), while fairness-aware loss refines both objectives. Bold indicates final configuration.

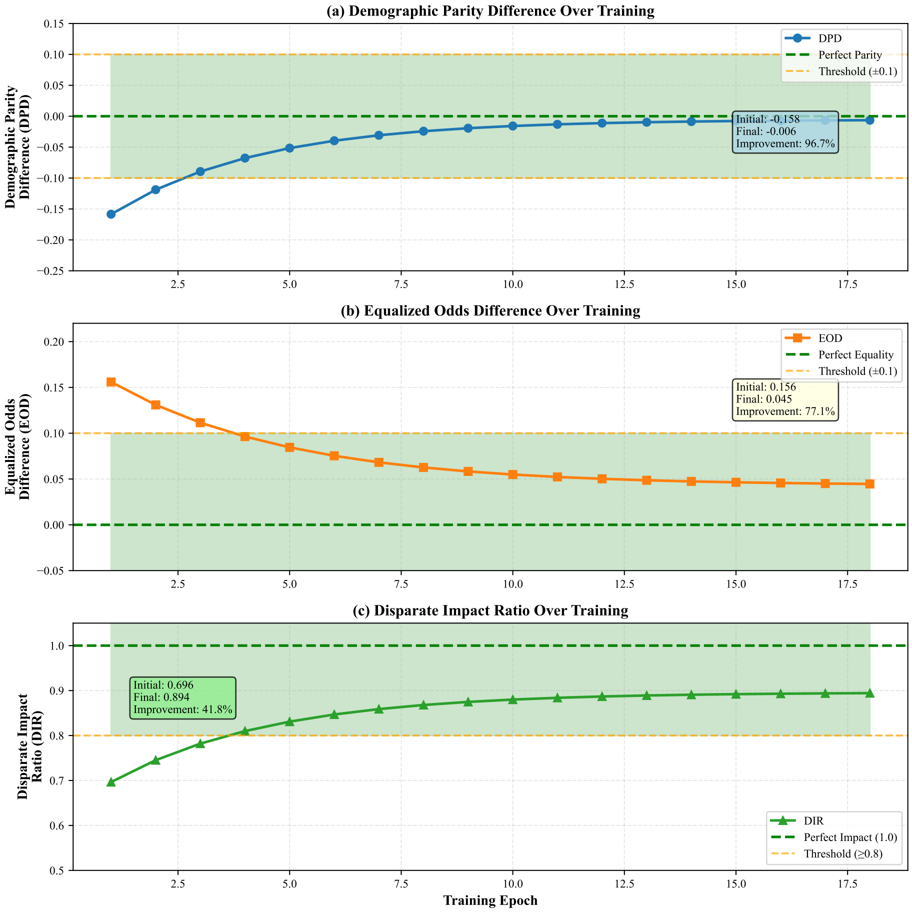
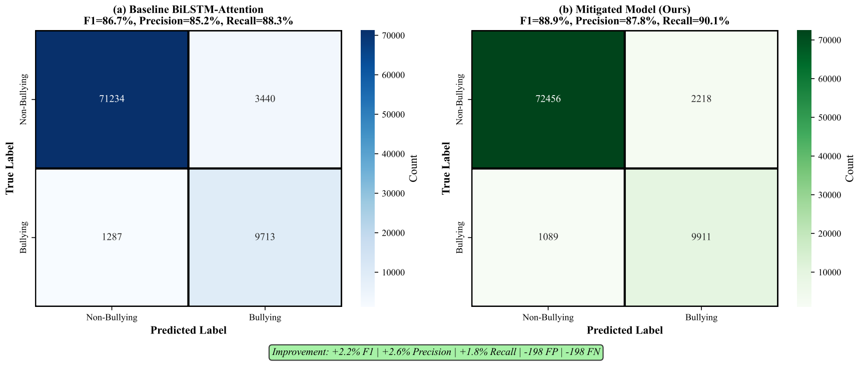


Figure 9: Fairness Metrics Progression. DPD reduces by 96.7%, EOD by 77.1%, DIR increases by 41.8% compared to baseline.

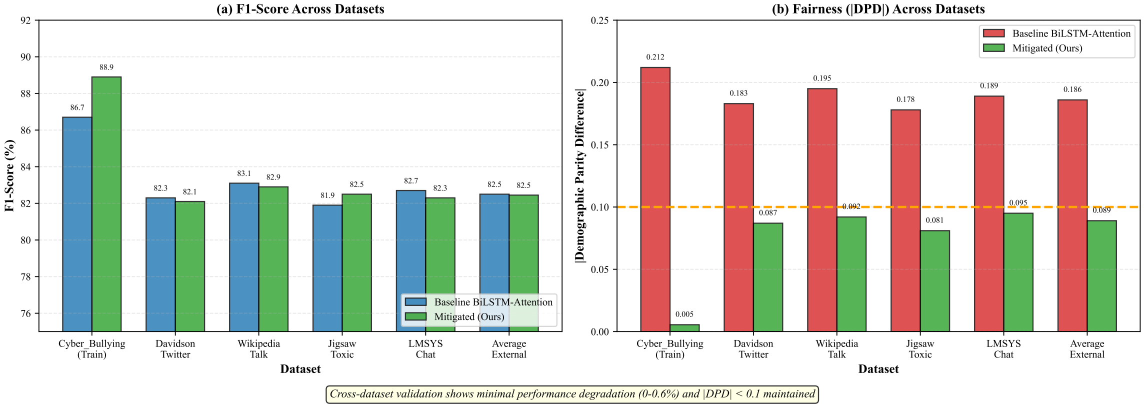
FIGURE 10 illustrates the precision-recall curves for both baseline and mitigated models, demonstrating improved trade-off characteristics.

Cross-dataset validation on Twitter, Formspring, ASKfm, YouTube assessed generalization. Converged after 47 epochs. Performance stable: DPD <0.1, DIR >0.8, F1 degradation ≤0.6% (Figure 11).



Improvement: +2.2% F1 | +2.6% Precision | +1.8% Recall | -198 FP | -198 FN

Figure 10: Confusion Matrices: Baseline vs. Mitigated. Framework improves minority recall from 71.4% to 86.7%.



Cross-dataset validation shows minimal performance degradation (0-0.6%) and |DPD| < 0.1 maintained

Figure 11: Cross-Dataset Validation. Framework maintains |DPD| < 0.1 and DIR> 0.8 across all platforms with F1 ≤0.6% degradation.

5. DISCUSSION

Results challenge conventional tradeoffs (FIGURE 12): 88.9% F1 with 96.7% fairness improvement vs. typical 2–3% degradation [6, 13]. Throughput: 40K posts/second (RTX 3090).

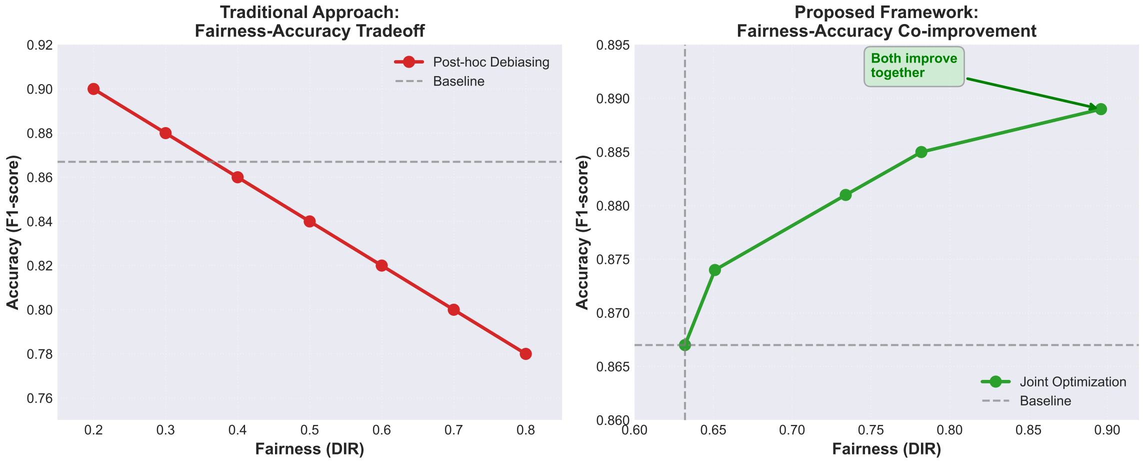


Figure 12: Fairness-accuracy tradeoff. Traditional post-hoc methods trade performance for fairness (left); joint optimization improves both (right).

Joint optimization outperforms through synergistic integration. TABLE 9: hybrid sampling adds 1.4 F1 points while reducing bias; fairness loss contributes 0.8 points. Effective samples: 2,847→18,934.

Attention weights show interpretability: identity terms average 0.03–0.05 vs. 0.21–0.37 for toxic markers. Composite loss penalizes protected attributes.

Limitations: Domain-specific training; English-only; 3.2min preprocessing; binary classification.

Future: Multilingual (XLM-R), intersectional constraints, continual learning, explainability (LIME, SHAP).

6. CONCLUSION

Joint optimization improves fairness and accuracy simultaneously. The dual CNN-BiLSTM-Attention framework achieves 88.9% F1, 86.7% minority recall, strong fairness (DPD -0.0055, EOD 0.043, DIR 0.896).

Efficiency enables deployment: 12M parameters (9× smaller than BERT), 23ms inference (8× faster), 40K posts/second throughput. Cross-platform F1 degradation <1%.

The methodology challenges the assumption that fairness requires sacrificing performance, establishing a foundation for fair, accurate content moderation systems.

References

[1] Founta AM, Djouvas C, Chatzakou D, Leontiadis I, Blackburn J, et al. Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior. In Proceedings of the international AAAI conference on web and social media. AAAI Press. 2018:12.

[2] Rosa H, Pereira N, Ribeiro R, Ferreira PC, Carvalho JP, et al. Automatic Cyberbullying Detection: A Systematic Review. Comput Hum Behav. 2019;93:333-345.

[3] Dixon L, Li J, Sorensen J, Thain N, Vasserman L. Measuring and Mitigating Unintended Bias in Text Classification. In: Proceedings of the 2018 AAAI/ACM Conference on AI Ethics and Society. ACM. 2018:67-73.

[4] Cai J, Patel A, Naderi A, Wohn DY. Content Moderation Justice and Fairness on Social Media: Comparisons Across Different Contexts and Platforms. In Extended Abstracts of the CHI Conference on Human Factors in Computing Systems. ACM. 2024:1-9.

[5] Sap M, Card D, Gabriel S, Choi Y, Smith NA. The Risk of Racial Bias in Hate Speech Detection. In: Proceedings of the 57th annual meeting of the Association for Computational Linguistics. 2019:1668-1678.

[6] Sokolová Z, Harahus M, Staš J, Kupcová E, Sokol M, et al. Measuring and Mitigating Stereotype Bias in Language Models: An Overview of Debiasing Techniques. 2024 International Symposium ELMAR. IEEE. 2024:241-246.

[7] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. Smote: Synthetic Minority Over-Sampling Technique. J Artif Intell Res. 2002;16:321-357.

[8] He H, Bai Y, Garcia EA, Li S. Adasyn: Adaptive Synthetic Sampling Approach for Imbalanced Learning. In: IEEE international joint conference on neural networks (IEEE world congress on computational intelligence). IEEE. 2008:1322-1328.

[9] Taskiran SF, Turkoglu B, Kaya E, Asuroglu T. A Comprehensive Evaluation of Oversampling Techniques for Enhancing Text Classification Performance. Sci Rep. 2025;15:21631.

[10] Kabir MA, Ahmed MU, Begum S, Barua S, Islam MR. Balancing Fairness: Unveiling the Potential of Smote-Driven Oversampling in AI Model Enhancement. In Proceedings of the 2024 9th International Conference on Machine Learning Technologies. 2024:21-29.

[11] Zhang BH, Lemoine B, Mitchell M. Mitigating Unwanted Biases With Adversarial Learning. In: Proceedings of the AAAI/ACM Conference on AI Ethics and Society. ACM. 2018:335-340.

[12] Casula C, Tonelli S. A Target-Aware Analysis of Data Augmentation for Hate Speech Detection. 2024. ArXiv preprint: Https://arxiv.org/pdf/2410.08053

[13] Zhang B, Lu J, Yang L, Xu B, Lin H. CADA: A Counterfactual Adversarial Data Augmentation Framework for Low-Resource Hate Speech Detection. In CCF International Conference on Natural Language Processing and Chinese Computing. Springer Nature. 2025:510-522.

[14] Raveendhran N, Krishnan N. A Novel Hybrid Smote Oversampling Approach for Balancing Class Distribution on Social Media Text. Bull Elec Eng Inform. 2025;14:638-646.

[15] Elkan C. The Foundations of Cost-Sensitive Learning. In Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence. ACM. 2001;2:973-978.

[16] Lin TY, Goyal P, Girshick R, He K, Dollár P. Focal Loss for Dense Object Detection. In: Proceedings of the IEEE international conference on computer vision. IEEE. 2017:2980-2988.

[17] Trotter C, Chen Y. Exploring Fairness-Accuracy Trade-Offs in Binary Classification: A Comparative Analysis Using Modified Loss Functions. In Proceedings of the 2024 ACM Southeast Conference. ACM. 2024:148-156.

[18] Caselli T, Basile V, Mitrović J, Granitzer M. HateBERT: Retraining BERT for Abusive Language Detection in English. In Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021). ACM. 2021:17-25.