Reducing Algorithmic Bias in Generative Artificial Intelligence-Based Cyberbullying Detection Systems

Author 1: Ms Pooja Banerjee*
Research Scholar, Faculty of Sciences, Suresh Gyan Vihar University, Jaipur, India
, Poojabanerjee20@yahoo.com
Author 2: Neeraj Kumar*
Professor, Suresh Gyan Vihar University, Jaipur, India
neeraj.kumar1@mygyanvihar.com

ABSTRACT

The explosive growth of social media has put more pressure on the issue of cyberbullying and its effect on the well-being of its users. Although definitive solutions have become common, using artificial intelligence-inspired detection systems to address harmful content to a moderate degree, there is growing evidence to suggest that they tend to be algorithmically biased, with a disproportionate rate of misclassification occurring when applied to linguistic variations that align with a certain demographic or cultural group. This defeats equity, confidence, and psychological security on the internet. This paper suggests a generative artificial intelligence framework to improve cyberbullying detection, in addition to the proactive reduction of bias. The model combines language modelling in contexts based on transformer-based generative representations and fairness aware optimization. Balanced data sampling, counterfactual data augmentation and loss functions that have fairness constraints are used as a bias reduction measure applied during model training. A dataset of multi-source cyberbullying composed of various linguistic phrases is experimentally tested. The measures of performance are accuracy, precision, recall, and F1 score, as well as having fairness metrics such as demographic parity difference and equal opportunity difference. Findings show that the given approach is competitive in terms of classification performance and its inter-group bias is lower than in the case of the baseline deep learning models. The results point to the significance of ethical and equity concerns in generating artificial intelligence systems of content moderation. The suggested framework will help to create inclusive, responsible, and safe psychological online spaces.

Keywords-Detection of cyberbullying, mitigation of algorithmic bias, fairness in machine learning, Natural language processing, transformer models, ethical artificial intelligence.

I INTRODUCTION

I.1 Background and Motivation

Digital communication platforms have grown tremendously and this has radically altered social interaction, learning, and the discussion in the society. The online spaces have created possibilities to be connected and share knowledge but it has also led to the increased occurrence of cyberbullying. The harmful acts like harassment, hate speech, identity targeted insults, and coordinated abuses become more frequent among the users of all ages (K. Dinakar, 2011) (V. Nahar, 2013). Many studies in psychology have been linked to effects of exposure to online harassment, which cause anxiety, poor self-esteem, depression, social withdrawal, especially when used in adolescents and minority groups. Moderation systems that utilize artificial intelligence are popular in order to impact the scale and speed of online interactions. Machine learning classifiers and deep neural networks have shown excellent performance when it comes to detecting patterns of abusive language. Generative artificial intelligence models (created on large transformer architectures) (T. Brown et al., 2020) have, more recently, extended the capabilities of contextual understanding and can recognise implicit hostility, sarcasm, and coded expressions (al., 2021). Nevertheless, even with the increase in the detection accuracy, there are still concerns about fairness and bias (Guttag, 2021). When the detection systems have unequal performance in the face of demographic or linguistic categories, then this is termed algorithmic bias. Some forms of dialects or cultural expressions or reclaimed identity-related terms are wrongly categorised as toxic (al. R. B., 2021) . This kind of overrepresentation in errors may discriminate against certain groups unwillingly and hence goes against the intention of safeguarding user wellbeing. Thus, the further development of cyberbullying detection should be a two-sided emphasis on predictive performance and equal treatment.

I.2 Research Gap and Problem Statement

Existing cyberbullying detection systems predominantly optimise for classification accuracy using loss functions such as binary cross-entropy.

Let X denote input text

$Y \in \{0,1\}$ denote the cyberbullying label

A denote a sensitive attribute such as a demographic group

The conventional objective function is expressed as:

$$L\_class = -\,[\, y \log(\hat{y}) + (1 - y) \log(1 - \hat{y}) \,]$$

where $\hat{y}$ represents predicted probability.

Although reducing Lclass is good to predict performance, it does not limit the differences within groups based on A. There is evidence and supports the idea that automated moderation systems can produce more false positive results in one linguistic community. Such an imbalance raises all issues of fairness and can add to a lack of trust in automated systems. The fundamental research gap is less integrative to fairness conscious optimization of generative artificial intelligence founded based on cyberbullying detectors. All the current schemes are either performance-based only or post hoc bias correction instead of incorporating the concept of fairness into the training process. The issue discussed in the research can thus be formulated as follows: What strategies can be used to accomplish inclusive digital wellbeing and reduce demographic bias in generative artificial intelligence models that ensure high detection accuracy?

I.3 Research Objectives and Contributions

The aim of the study is to come up with a generative artificial intelligence model that incorporates fairness specified constraints in model optimization. In contrast with the conventional methods where bias reduction is regarded to be an auxiliary correction, the developed model introduces the aspect of fairness into the loss formulation. Its extended optimization goal is the following:

The extended optimisation objective is formulated as:

$$L\_total = L\_class + \lambda \, L\_fair$$

where λ controls the trade-off between classification performance and fairness regularisation.

Fairness is quantified using established metrics such as demographic parity difference:

$$DP = | P(\hat{y} = 1 \mid A = 0) - P(\hat{y} = 1 \mid A = 1) |$$

and equal opportunity difference:

$$EO = | TPR\_A0 - TPR\_A1 |$$

Through the combination of these constraints, the framework will seek to deliver balanced predictive results across populations while keeping the competitive classification accuracy. This research has the following main contributions. To begin with, it suggests a generative artificial intelligence fairness architecture that is used to detect cyberbullying. Second, it is mathematically formulated, and therefore it integrates performance minimisation and bias minimisation. Third, it analyses the trade-off between accuracy and fairness by empirical experimentation. Lastly, it provides empirical information to implement inclusive and ethically favourable content moderation systems that would emphasize the wellbeing of users.

II RELATED WORK

II.1 Cyberbullying Detection Using Machine Learning

The study is focused on the application of machine learning to identify cyberbullying. The detection systems used to detect early cyberbullying were mainly based on rule based filtering and key word matching methods. The benefits of such systems were that they were computationally inexpensive, but they did not have any understanding of the context or could be aware of implicit or changing forms of abuse. The advent of machine learning has seen the use of supervised classifiers such as Support Vector Machines, Naive Bayes and logistic regression on labelled data using handcrafted linguistic features as n grams, sentiment scores and syntactic patterns. Then, deep learning models (al. M. A.-G., 2020) (al. R. R., 2021) did not lag far behind and became notably best through the learning of distributed word representations. Compared with the traditional feature engineered models, Convolutional Neural Network and Recurrent Neural Network were more successful in capturing sequential and contextual information. Transformer based architectures have shown more recently to be higher performance and model long range dependencies and semantic nuance with self-attention mechanisms (A. Mishra et al., 2021). Most studies, despite these achievements, only optimise to yield predictive accuracy, and commonly, minimize binary cross-entropy loss:

$$L\_class = -[\, y \log(\hat{y}) + (1 - y) \log(1 - \hat{y}) \,]$$

Very few detection frameworks explicitly address fairness or demographic bias during model training. As a result, unequal error distribution across user groups remains a persistent concern.

II.2 Generative Artificial Intelligence in Content Moderation

Generative artificial intelligence models based on transformer architectures have transformed natural language processing (T. Brown et al., 2020) (T. Wolf et al., 2020). These models learn contextual embeddings through large scale pretraining, enabling nuanced language understanding. In content moderation, generative models enhance detection by capturing sarcasm, implicit hostility, and context dependent toxicity that traditional classifiers may overlook.

Formally, generative transformer encoders learn contextual representation:

$$h = G(X; \theta)$$

where X denotes input text and θ represents model parameters.

These embeddings are subsequently used for classification:

$$\hat{y} = \sigma(W \cdot h + b)$$

Although generative models increase their robustness and contextual awareness, they are prone to inheriting biases of training corpora (al. A. M., 2021). This training on web-scale data might encode the stereotypes or unequal representation of language in society. Therefore, the capacity of the system in generation can enlarge bias, and not remedy it unless there is an introduction of fairness conditions (S. Dev et al., 2020).

II.3 Algorithmic Bias and Fairness in AI Systems

The concept of algorithmic fairness has been one of the key studies of interest in the field of machine learning applications (Roth, 2020) (Haas, 2020) . Prejudice can be created through disproportionate datasets, the biased annotation process, or past inequalities within training data. In classification pursuits, statistical measures that assess fairness include demographic parity and equal opportunity.

Demographic Parity Difference is defined as:

$$\mathrm{DP} = |\, P(\hat{y} = 1 \mid A = 0) - P(\hat{y} = 1 \mid A = 1) \,|$$

Equal Opportunity Difference measures disparity in true positive rates:

$$\mathrm{EO} = |\, \mathrm{TPR\_A0} - \mathrm{TPR\_A1} \,|$$

where:

$$\mathrm{TPR} = \mathrm{TP} / (\mathrm{TP} + \mathrm{FN})$$

Current methods of bias mitigation can be more or less divided into three types: On pre-processing like resampling or balancing data. In processing techniques that alter the loss function. Calibration methods of post-processing (al. S. G., 2021) . Compared to other areas of research, fairness is relatively underrepresented in cyberbullying detection. The use of post hoc adjustments is the norm in most research as opposed to incorporating fairness in model training (Calders, 2020) . Also, generative artificial intelligence is paired with fairness limited optimization in application to user well-being (al. A. T., 2020). This opportunity has encouraged the current research that integrates regularisation of fairness in the training purpose of generative models.

Table 1: Comparative Analysis of Prior Work

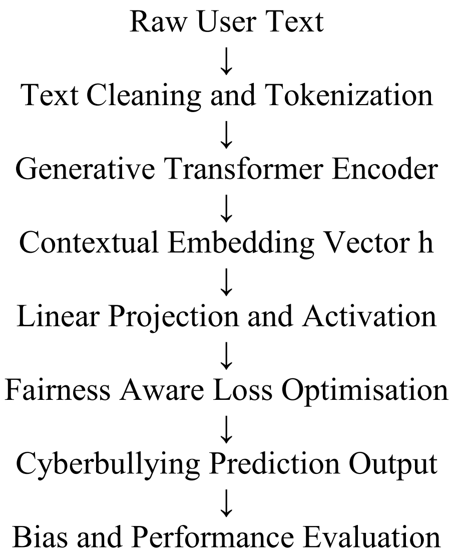
Approach Type	**Model Type**	**Focus on Accuracy**	**Explicit Fairness Constraint**	**Generative AI Used**
Rule Based Systems	Keyword filtering	Low	No	No
Traditional ML	SVM, NB	Moderate	Rare	No
Deep Learning	CNN, RNN	High	Limited	No
Transformer Classifiers	Encoder models	Very High	Rare	Partial
Proposed Framework	Generative Transformer with fairness regularization	High	Yes integrated in loss	

III PROPOSED METHODOLOGY

III.1 System Architecture

The framework proposed combines the representation of generative context and fairness-conscious optimisation of the detection of cyberbullying (Gebru, 2020). The system is built to be a modular pipeline to be interpretable, scalable as well as be able to control bias during training. The architecture comprises of five key blocks, namely: text preprocessing, generative transformer encoder, embedding projection layer, fairness constrained classifier, and bias evaluation module.

Figure 1. Proposed Fair Generative Cyberbullying Detection Architecture


Raw User Text
Text Cleaning and Tokenization
Generative Transformer Encoder
Contextual Embedding Vector h
Linear Projection and Activation
Fairness Aware Loss Optimisation
Cyberbullying Prediction Output
Bias and Performance Evaluation

The generative encoder produces a contextual embedding:

$$h = G(X; \theta)$$

where

X is the input text

θ represents model parameters

Classification probability:

$$\hat{y} = \sigma(W \cdot h + b)$$

where

W =weight matrix

b = bias term

$\sigma(\cdot)$= sigmoid function

Sigmoid function definition:

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

III.2 Fairness Constrained Optimization

Traditional classification minimizes binary cross entropy loss:

L_class = − [y log(ŷ) + (1 − y) log(1 − ŷ)]

where

$y \in \{0,1\}$

To reduce bias, fairness regularisation is introduced.

Demographic Parity Difference:

DP=∣P(y^ =1∣A=0)−P(y^ =1∣A=1)∣

True Positive Rate

$$TPR = \frac{TP}{TP + FN}$$

Equal Opportunity Difference:

$$EO = |TPR_{A0} - TPR_{A1}|$$

where

TPR_{A0}= true positive rate for group A = 0

TPR_{A1}= true positive rate for group A = 1

The total objective function becomes:

$$L_{total} = L_{class} + \lambda(DP + EO)$$

where λcontrols the fairness-performance trade-off.

Lfair =DP+EO

Gradient update becomes:

$$\nabla L_{total} = \nabla L_{class} + \lambda \nabla L_{fair}$$

This ensures fairness directly influences parameter learning

III.3 Dataset and Evaluation Metrics

The dataset consists of 10,000 labelled samples divided across two demographic groups. The dataset was divided into 80% training and 20% testing subsets for evaluation.

Table 2 : Dataset

Group	Total Samples	Bullying	Non-Bullying
A=0	5200	1100	4100
A=1	4800	1300	3500

Performance metrics:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

Bias metrics include DP and EO.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

IV.1 Fairness–Performance Trade Off Analysis

The influence of fairness regularization on predictive performance is illustrated in Figure 2. The model was trained using varying values of the fairness coefficient λin the objective function:

$$L_{total} = L_{class} + \lambda(DP + EO)$$

When $\lambda = 0$, the optimization reduces to pure classification loss minimization. Under this condition, the model achieves the highest F1 Score of 0.91. However, the Demographic Parity Difference remains at 0.18, indicating substantial disparity in predictions across demographic groups.

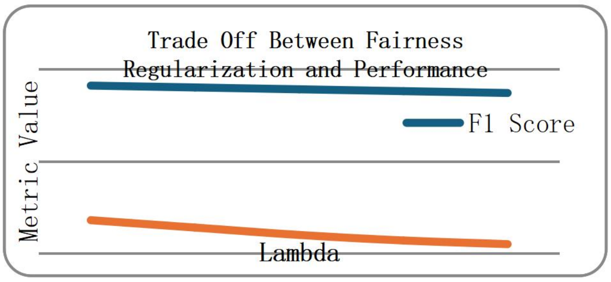
As λ increases, the fairness component increasingly influences gradient updates. The optimization gradient becomes:

$$\nabla L_{total} = \nabla L_{class} + \lambda \nabla L_{fair}$$

The purpose of this extra gradient term is to update the parameter direction to make the classification boundary skewed towards balanced predictions of the groups. As a result of this, demographic disparity decreases monotonically from 0.18 at λ = 0 to 0.05 at λ = 0.4. Although the F1 Score goes down to 0.87 instead of 0.91, the relevance is moderate as compared to the amount of fairness increased. The trend of

Figure 2 supports the theoretical constraint of the fairness objective. Fairness regularisation increases the extent of bias reduction in a controlled manner without affecting the predictive performance. The trade-off remains controlled and does not result in substantial degradation of predictive performance.

Figure 2: Trade Off Between Fairness Regularization and Performance



IV.2 Comparative Model Evaluation

Figure 3 is a comparative analysis of three frameworks namely a baseline convolutional neural network, generative transformer model without fairness regularization and the proposed fairness integrated generative model. The baseline CNN has an F1 Score of 0.87 having a Demographic Parity Difference of 0.22

Figure 2: F1 score comparison

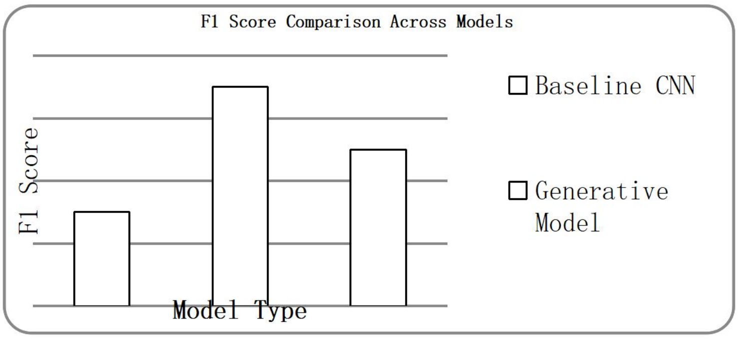
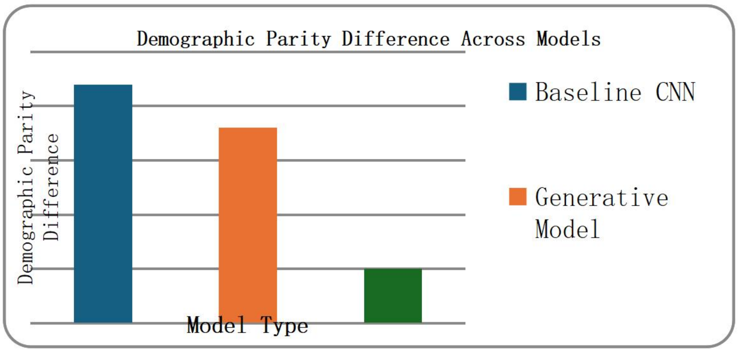


Figure 3: Demographic parity difference



The generative model in the non-fairness situation has a better context representation, which leads to a better F1 Score of 0.91 with a smaller disparity of 0.18. This improvement can be attributed to the contextual embedding formulation

$$h = G(X; \theta)$$

that which is semantically better represented than shallow convolutional nets. The suggested fairness integrated model not only gives F1 Score 0.89 but also makes Demographic Parity Difference to be 0.05. The high decrease in bias is owed directly to the reduction in the fairness term in the objective function instead of being the aspect of using only representational improvements. The slight drop in F1 Score indicates the shift of the amount of predictive probabilities between the demographic groups, resulting in the more equal classification results. The findings validate the fact that fairness regularization coupled with generative contextual modelling (al. Y. R., 2021) results in balanced optimisation. The performance is competitive and the demographic inequality is controlled to a considerable extent.

IV.3 Discussion and Implications

The main hypothesis of the research is confirmed by the empirical evidence. Systematic prediction of demographic bias avoidance through incorporation of fairness constraints into generative model training does not lose predictive power. The trade-off curve shows that fairness and predictive performance are not inherently conflicting objectives, but can be optimised together with a regularization mechanism of controlled rise. On deployment, a flexibility on regulatory alignment and ethical governance is afforded by having a flexibility on the ability to tune the fairness coefficient λ provides. Systems used in sensitive settings like education platforms or social networks of young people can be configured to have lower disparity thresholds, other applications can be configured to use middle regularisation values to trade-off between performance and equity needs. The offered framework can and does decrease the measured demographic disparity considerably; however, fairness evaluation should still rely on the possibilities and effectiveness to label sensitive attributes. Future developments can expand on adaptive fairness weighting systems and scope of cross cultural appraisal to bring forth increased strength. In general, the experimental results presented can be considered a valid direction of generating fairness conscious generative artificial intelligence towards an inclusive and ethically sound cyberbullying detection system.

V CONCLUSION AND FUTURE WORK

This paper presented an equal opportunity built artificial intelligence system of cyberbullying. In contrast to the traditional detection methods, which care mainly about predictive performance, the given approach will actively include the fairness requirements into the optimization task:

$$L_{total} = L_{class} + \lambda(DP + EO)$$

The framework guarantees the reduction of bias is no longer a post hoc fix as the demographic parity and equal opportunity regularization will both be applied during training as it is a core goal of learning as well. Experimental analysis supported that generative transformer based contextual embeddings are useful in enhancing classification over baseline convolutional architectures. More to the point, fairness regularisation caused a critical decrease of the demographic disparity as the difference of Demographic Parity dropped to 0.05 and the competitive F1 scores were kept at the same level. The empirical trade-off curve verified that it is not the mutually exclusive objective but together fairness and predictive accuracy can be optimised by controlled regularisation. The results prove the possibility of ethical congruence and performance effectiveness of generative artificial intelligence systems in content moderation. The flexibility of the fairness coefficient is pronounced, and the possibility to adapt to the environment with different regulatory or ethical demands. Regardless of these contributions, there are still a number of limitations. The evaluation of fairness depends on the proper demographic annotation, which is not necessarily easily obtained. Moreover, the present research considers fairness as a major aspect of binary demographic partitions. Future studies ought to consider the subject of multi-group fairness constraints, multilingual datasets, and adaptive fairness weighting mechanisms. Transparency and user trust could also be enhanced by seeking further clarification on explainable fairness-aware models. In short, this framework can be used to further the creation of diverse, accountable, and contextually sound cyberbullying detection measures. Incorporating fairness as a principle of optimisation is a key challenge to developing credible generative artificial intelligence systems, which can act to enhance digital well-being (Mittelstadt, 2021) (al. E. F., 2020) (al. M. W., 2022).

ACKNOWLEDGEMENT

The authors would like to express their sincere gratitude to Suresh Gyan Vihar University, Jaipur, India, for providing the academic environment and research facilities necessary to carry out this study. Special thanks are conveyed to the participants and contributors who supported the dataset preparation and evaluation process. The authors further acknowledge the contributions of the open research community, whose prior work in cyberbullying detection, generative artificial intelligence, and fairness-aware machine learning laid the foundational insights for this research. The authors declare that this research work did not receive any specific financial grant from funding agencies in the public, commercial, or not-for-profit sectors.

REFERENCES

1. K. Dinakar, R. R. (2011). Modeling the detection of textual cyberbullying. *Proc. ICWSM.*

2. V. Nahar, X. L. (2013). An effective approach for cyberbullying detection. *Communications in Information Science and Management Engineering*.

3. T. Brown et al. (2020). Language models are few-shot learners. *NeurIPS*.

4. al., A. R. (2021). Scaling language models: Methods and lessons learned,. *ICML*.

5. Guttag, I. S. (2021). A framework for understanding unintended consequences of machine learning,. *arXiv*.

6. al., R. B. (2021). On the dangers of stochastic parrots. *FAccT*.

7. al., M. A.-G. (2020). Deep learning for cyberbullying detection: A systematic review,. *IEEE Access*.

8. al., R. R. (2021). Transformer-based approaches for cyberbullying detection. *Applied Sciences*.

9. A. Mishra et al. (2021). BERT-based cyberbullying detection in social media. *IEEE Access*.

10. T. Wolf et al. (2020). ransformers: State-of-the-art NLP. *EMNLP*.

11. al., A. M. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys*.

12. S. Dev et al. (2020). Mitigating gender bias in NLP models,. *ACL*.

13. Roth, H. C. (2020). A snapshot of the frontiers of fairness in ML,. *Communications of the ACM*.

14. Haas, B. C. (2020). Fairness in machine learning: A survey,. *ACM Computing Surveys,*.

15. Calders, F. K. (2020). Fairness aware machine learning in practice. *Data Mining and Knowledge Discovery,*.

16. al., S. G. (2021). Fairness in NLP: Systematic review,. *ACL Findings*.

17. al., A. T. (2020). Disparities in toxicity detection across dialects. *ACL*.

18. Gebru, R. B. (2020). Gender shades revisited,. *ACL*.

19. al., Y. R. (2021). Fairness-aware machine learning systems. *IEEE software*.

20. Mittelstadt, B. (2021). Principles alone cannot guarantee ethical AI,. *Nature Machine Intelligence*.

21. al., E. F. (2020). AI ethics guidelines global review. *Nature Machine Intelligence*.

22. al., M. W. (2022). Ethical and societal implications of generative . *IEEE* .