Proceedings of the Indian National Science Academy (2025) 91:891–904
https://doi.org/10.1007/s43538-024-00371-1

RESEARCH PAPER

# Bias mitigation in text classification through cGAN and LLMs

**Gunjan Kumar**[1] · **Jyoti Prakash Singh**[1]

Received: 14 June 2024 / Accepted: 6 November 2024 / Published online: 5 December 2024


**Abstract**
Investigating bias, fairness, and ethics in AI/ML models is essential in the digital era, as automated systems are increasingly used to enhance people's daily lives. Evaluating these systems in terms of fairness has become crucial. Our proposed work focuses on identifying and mitigating gender bias in text classification models. Specifically, we are evaluating a hate speech classification (HSC) model using fairness metrics and mitigating bias through pre- and in-processing debiasing techniques. In the pre-debiasing stage, we address bias by applying gender swap techniques and removing gender-related terms from the dataset. For in-processing debiasing, we use conditional generative adversarial networks (cGAN) and large language models (LLMs) to generate text similar to that of minority classes, helping balance the dataset and reduce bias during model training. We assess the impact of both methods on bias reduction in the HSC model, with each targeting bias at different stages of model development. While the classification performance of the HSC model showed a slight decrease, fairness evaluation metrics improved by 3.03% and 3.42%, demonstrating the methods' effectiveness. Addressing gender bias in HSC models applied to social media posts is vital for fostering a healthier digital society.

**Keywords** Text-classification · Bias · Fairness · LLMs · cGAN · Mitigation

## Introduction

Bias and fairness in artificial intelligence (AI) and machine learning (ML) models are prominent topics of discussion and debate, as well as a growing area of research. Identifying, analyzing, and mitigating bias, as well as enhancing fairness, are critical as AI-based systems are increasingly deployed across various sectors of society, such as healthcare applications (Wang et al. 2023; Zhao et al. 2023) and criminal justice are few examples. Since these intelligent models train on data collected from society to make decisions, imbalanced or insufficient datasets can lead to biased outcomes. For example, in healthcare applications, models may associate certain diseases with a higher likelihood of affecting specific genders or races. Similarly, it has been observed that for the same offense, black individuals received harsher punishments compared to white individuals[1] (Angwin et al. 2022).

The cases similar to this highlight the issue of bias in the learned models. A biased model often produces results that favor certain groups based on gender, race, or ethnicity (Zafar et al. 2017). The use of such biased models in critical applications like healthcare, legal systems, or financial services erodes public trust and harms businesses. Gender biases are frequently observed in Google searches and translations. For instance, as shown in Fig. 1, the term "homemaker" is associated with feminine terms like "*housewife, not a job, meaning, and wife*," despite the fact that men can also be homemakers. Similarly, as demonstrated in Fig. 2, the term "breadwinner" is linked to male-dominated terms such as "*husband, quotes, book, movie*".

Similarly, gender bias is also seen in Google Translate Fig. 2, where words like *nanny* or *nurses* are translated to feminine gender in Hindi, whereas the doctor is translated to the male gender. The nurses can be male, or the doctor can be female also.

A few research have tried to mitigate the issues of bias by (i) pre-processing the data, (ii) in-processing during training, and (iii) post-processing the result. The pre-processing of the dataset to reduce the bias was done by researchers such as Calmon et al. (2017), Kamiran and Calders (2012),

✉ Gunjan Kumar
gunjank.phd20.cs@nitp.ac.in

Jyoti Prakash Singh
jps@nitp.ac.in

[1] National Institute of Technology Patna, Patna, India

[1] https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.


**Fig. 1** Bias exists in Google searches while searching breadwinners and homemakers, which are seen in the given option

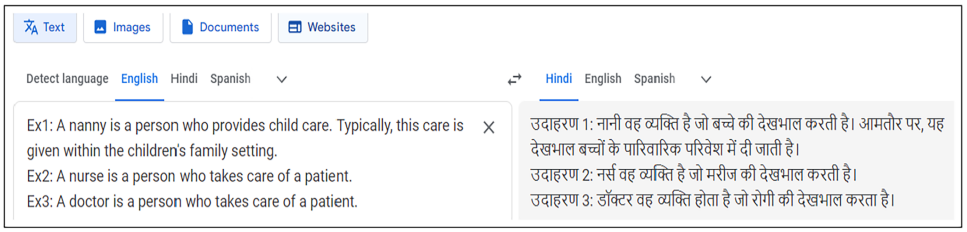


**Fig. 2** Bias seen in Google translator while translating the sentences using "nanny", "Doctor" and "Nurse"

and Park et al. (2018). Calmon et al. (2017), developed a method by pre-processing the data to reduce bias. They used a mathematical approach to achieve objectives like minimizing bias, mitigating distortion within individual data points, and preserving the data's utility. They tested their method on the criminal re-offending dataset and significantly reduced the discrimination with a small impact on the accuracy of predictions. Kamiran and Calders (2012) advocated suppressing the sensitive attribute and adjusting class labels to mitigate discrimination without altering individual instance labels. Whereas, Park et al. (2018) analyzed the effect of different pre-trained word embeddings on models trained on abusive language datasets and measured gender biases. They also used three bias mitigation methods: gender swap data augmentation, debias word embeddings, and fine-tuning with a larger corpus. They found that these methods reduce gender bias by 90–98%.

The in-processing debiasing method has also proven to be an effective technique, as reported by Nabi et al. (2019) and Feldman and Peake (2021). Nabi et al. (2019) address the challenge of preventing automated decision systems and learning algorithms from perpetuating societal injustices. They discuss strategies for making fair decisions by mitigating biases related to sensitive attributes such as gender, race, and disability. The authors propose a method that combines causal inference with constrained optimization to learn fair policies, validating their approach using both synthetic and real-world criminal justice data. Reducing bias in the dataset and optimizing the model's training process are crucial aspects of their work.

Feldman and Peake (2021) developed a model to reduce end-to-end bias in deep learning (DL) models by utilizing all three debiasing methods ('Pre', 'In', and 'Post') across two datasets: the University of California, Irvine (UCI) Adult dataset and the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) dataset. However, their model does not address debiasing in text classification or use a text-based dataset. In contrast, some researchers have worked with text datasets, generating text samples to mitigate bias in models (Rezaei et al. 2020; Akter et al. 2021). Text sample generation is critical in natural language processing (NLP), which involves creating samples that closely resemble textual data. Significant advancements in this area have been made by researchers using pre-trained models such as GPT-2 and GPT-3 (Chung et al. 2023; Li et al. 2022; Yoo et al. 2021). However, most existing pre-trained models are trained on large corpora derived from society, which are often biased in terms of gender, ethnicity, and stereotypes. Consequently, the performance of these models reflects the same biases, as shown in Figs. 1 and 2, which motivates us to use a GAN model to generate text samples aimed at debiasing text classification models. While existing GAN models have been employed to debias image classification models and generate text-to-image (Reed et al. 2016), image-to-text (Chen et al. 2022), and image-to-image (Rezaei et al. 2020) data, they have not been used to generate text-to-text samples.

Any model receiving biased input will propagate that bias in its performance, and its output also contains the same bias, which has motivated us to propose a model to identify and reduce bias in text classification models. Most existing models are trained on datasets originating from society and are prone to gender, race, and stereotype biases. We aim to

develop a model that can mitigate these biases in text classification models.

A dataset containing gender information is necessary to validate the proposed model. For this, we are using the Hate Speech Gender Bias (HSGB) dataset, which includes the gender of the targeted one. Hate speech classification (HSC) models have been found to suffer from gender bias (Fortuna and Nunes 2018; Mozafari et al. 2020). Additionally, several studies have demonstrated that the methods used for HSC are also biased (Park et al. 2018; Ryu et al. 2017; Dixon et al. 2018).

The major contributions to reducing bias in automated AI/ML-based systems are:

- Identifying and reducing gender bias in AI/ML-based text classification models using pre-processing methods such as gender swapping and removing gender-related terms.
- Reducing the bias of AI/ML models using in-processing debiasing methods by generating data samples using conditional generative adversarial networks (cGAN).
- Generating similar sentences of a particular class to balance the dataset using large language models (LLM) to validate the proposed model.
- Authenticating the AI-generated data samples by human evaluators.

The rest of the paper is structured as follows: "Related work" section discusses related work on fairness in AI models and gender bias in text classification and generation models. "Methodology" section describes the dataset and the proposed model. "Debiasing methods" section explains the debiasing methodology applied to the dataset. "Result" section presents the results of the proposed model. "Discussion" section outlines the key findings of the paper, and "Conclusion" section provides the conclusion.

## Related work

Over the past few years, the primary focus of many researchers has been the development of automated AI-based systems. We depend very heavily on AI/ML-based systems in our day-to-day lives, which are also used for decision-making systems. Therefore, fairness and bias of automated systems have become crucial and emerging research areas for researchers. Concerns have arisen regarding the accuracy and fairness of AI/ML models used in decision-making systems. Some authors have (Karnouskos 2020; Parra et al. 2021) raised awareness regarding the ethical implications of AI-based models and their impacts on society. Karnouskos (2020) focused on the social implications of AI in enhancing deepfakes. They say that deepfakes are made by modifying realistic digital products, and many such videos have appeared on social media in the past 2 years. Technical expertise and equipment to create deepfakes are optional, so anyone can easily produce and distribute such content on social media. They say that the social implications are significant and far-reaching, but it is too dangerous for social media. Parra et al. (2021) enrich our understanding of biases that exist in different areas by employing a scenario-based survey distributed to 387 participants in the United States. The results indicate that in scenarios related to human resource (HR) recruitment and financial product/service procurement, there is a greater propensity for questioning when perceived racial and gender bias is considered. In contrast, scenarios involving healthcare exhibit the lowest likelihood of questioning.

Various authors have extensively discussed the ethical implications, biases, and fairness aspects of AI-based automated systems across different societal domains (Peters et al. 2020; Zicari et al. 2021; Drozdowski et al. 2020; Khakurel et al. 2022). Peters et al. (2020) discuss the unintended consequences of artificial intelligence (AI), highlighting the absence of ethical implications in healthcare professional practices. They explore two complementary frameworks (The Responsible Design Process and The Spheres of Technology Experience) aimed at integrating ethical analysis into engineering practices to tackle this issue. Subsequently, they present the findings of an ethical analysis conducted within the context of Internet-delivered therapy in mental health, informed by these frameworks. Zicari et al. (2021) developed an inspection model to check the trustworthiness of the AI model based on ethics and named it Z-Inspection. The definition of trustworthy AI is taken from the high-level European Commission's expert group on AI. Z-Inspection is a simple inspection procedure applied to many domains where AI systems are used, including business, healthcare, the public sector, and numerous others. Drozdowski et al. (2020) examine the concerns surrounding biometric technologies in both public and academic spheres, particularly addressing the bias ingrained within automated decision systems, including biometrics. They thoroughly explore algorithmic bias within the context of biometrics, survey existing literature on bias estimation and mitigation, discuss relevant technical and social issues and outline remaining challenges and future directions for research, encompassing both technological and social aspects. Khakurel et al. (2022) investigated the benchmark Statlog "Australian Credit Approval" data set. They identified the bias existing in the dataset using the AIF360 tool and mitigated bias in both the data and the learning algorithms.

In the domain of automated decision algorithms, several factors can contribute to potential bias. Most prominently, the training data possibility be skewed, incomplete, outdated, disproportionate, or historically biased. These

factors affect the algorithmic training and propagate the biases from the data.GAN models balance the training data by generating various types of data, including text from text, text from images, and images from images. Although extensive research has been dedicated to generating images from images or text from images using GANs or similar models, less attention has been given to generat text from text. Kusner et al. (2017) introduced a model representing discrete data as parse trees and using a context-free grammar. They designed a variational autoencoder to encode and decode parse trees directly, ensuring that the generated outputs remain syntactically correct at all times. The model generated valid output, learnt a more coherent latent space in the nearby points, and decodes it to a similar discrete output. Chen et al. (2018) proposed a novel approach for text generation using feature-mover's distance (FMD), called feature-mover GAN (FM-GAN), and discussed the challenges of GANs to text generation due to the discrete nature of the text. They proposed an innovative approach inspired by optimal transport and matched the latent feature distributions of real and synthetic sentences using a novel metric called FMD. Their model showed its wide applicability and effectiveness in overcoming challenges associated with GAN-based text generation. To validate the model, they conducted many experiments, including unsupervised cipher cracking, style transfer from non-parallel text, and unconditional text generation. Fedus et al. (2018) introduce a GANs model by explicitly training the generator to produce high-quality samples and have a lot of success in image generation. However, using GANs for discrete language generation is a complex task. So, they introduced an actor-critic conditional GAN, which fills the missing text based on the contextual parameter but does not generate the discrete sentences. Their approach yielded more realistic conditional and unconditional text samples when compared to a model trained using maximum likelihood. Diao et al. (2021) developed a transformer based implicit latent GAN (TILGAN) model which merged a Transformer autoencoder and GAN in the latent space with a distribution matching based on the Kullback–Leibler (KL) divergence. They introduce a multi-scale discriminator to capture the semantic information at varying scales among the sequence of hidden representations encoded by the Transformer to increase the global coherence. The decoder is enhanced by an additional KL loss to be consistent with the latent generator. Still, they were not able to mimic the human writing style due to the discrete nature of the text.

Recently, Generative Pretrained Transformers (GPT) and Text-To-Text Transfer Transformers (T5) have been able to perform various NLP tasks, including text generation. Dale (2021) GPT-3 attracted significant attention in public media and received high attention in technical advancements in NLP. The main capability is to generate text that mimics the human writing style, which has captured people's fascination. However, they have not mentioned anything about the practical applications. After getting great attention on the GPT model, Ding et al. (2022) proposed a dataset annotation model. Dataset annotation is a very important process for the supervised learning task, the labelling/annotation of the data used to train ML and DL models. A high-quality annotation is crucial to get the best results from supervised learning. The performance of their model is better than that of traditional data annotation methods, and the output obtained is analyzed using various tasks. Zhuo et al. (2023) investigated the LLMs and observed that LLMs suffer from toxicity and social prejudice, introducing ethical and societal dangers leadings to irresponsible consequences. Their empirical investigations revealed few ethical difficulties in advanced LLMs, and very few systematic examination and user study of the risks and harmful behaviours of current LLMs usage have been done. To construct ethical LLMs responsibly, they perform a qualitative research method called "red teaming" on OpenAI's ChatGPT-1 to better understand the practical features of ethical dangers in recent LLMs. They analyze ChatGPT systematically concerning four perspectives: (1) Toxicity, (2) Reliability, (3) Robustness, and (4) Bias. Their findings shed light on AI ethics and harmful behaviours, offering insights for designing responsible LLMs to mitigate ethical hazards in LLM applications.

The research mentioned above highlights the various uses of LLM models and the bias of AI models. Both LLM models and AI models are trained on datasets collected from society; hence, they are bound to have biases. Hence, we need to build models that are bias-free. The limitations exist in the cGAN in generating the text, and the HSC model used for hate text classification needs to be fixed. It's important to generate the text and investigate the bias and fairness of the HSC models concerning gender attributes because they could have significant and different effects on various groups, regardless of specific definitions. We are using cGAN and LLMs to generate the text to reduce the bias of the models at the training time.

## Methodology

This section presents a model to identify gender bias in an automated HSC model based on social media posts (X). The proposed model is shown in Fig. 3. Section 3.1 outlines the collection, labelling, and description of the dataset. The detailed implementation of the proposed approach is provided in "Proposed approach" section, which explains each step involved in building the model. The obtained results, analyzed for fairness, are discussed in "Result" section. If the fairness evaluation indicates bias, a debiasing method is applied, as described in "Debiasing methods" section.

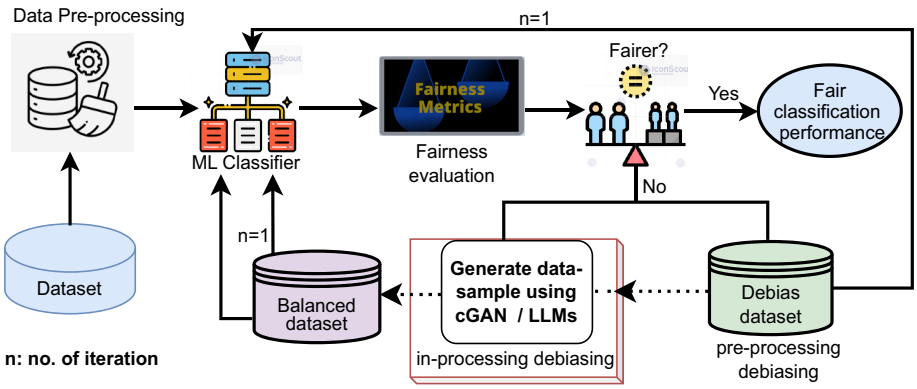


**Fig. 3** Flow diagram of the proposed model

## Data prepration

In this section, we discuss the creation of a new dataset derived from the MMHS150K dataset,[2] published by Gomez et al. (2020). The original dataset contains six classes: racist, homophobic, sexist, religion-based, not hate, and attacks on other communities. Data samples from the sexist class are utilized to validate the proposed model. This newly created dataset is named the hate speech gender bias (HSGB) dataset. Five annotators from different organizations were enlisted to label the data samples, focusing on gender as the primary concern. Of the five annotators, two were male, and three were female, with the intention of incorporating diverse perspectives on hate speech. The annotators were trained to label abusive tweets based on offensive language related to gender, as well as corresponding nouns and pronouns referring to males or females, as listed in Table 1. Labels were assigned according to the gender targeted by the abusive tweets: "0" for males, "1" for females, and "2" for tweets not related to any gender. The final label for each tweet was determined using a majority voting scheme. The HSGB dataset contains 2782 tweets, of which 1534 are labelled as not targeting any gender, 365 are labelled as targeting males, and 883 as targeting females.

## Proposed approach

The proposed model is used to evaluate the performance of the hate speech classification (HSC) model based on classification and fairness metrics. The flow diagram of the proposed model is shown in Fig. 3. The proposed model comprises multiple phases: the first involves data pre-processing, and feature extraction is in the second phase. In the third phase, classification is performed with ML and DL models. In the final phase, the classification results are analyzed with respect to fairness, and debiasing methods are applied if any bias exists.

### Data pre-processing

During data pre-processing, unnecessary special characters such as # $ % & ~ _ ^ { } \ were removed from the textual tweets. Stopwords were retained to preserve the original meaning of the sentences. Capitalized words were converted to lowercase to ensure syntactical consistency and to utilize the same pre-trained embedding values for words with similar meanings. Each comment was limited to a maximum sequence length of 25 words. Shorter sentences were post-padded, while longer sentences were truncated to retain only the first 25 words.

### Embeddings

Embeddings represent words as vectors in a real-valued space, where syntactically or semantically similar words are positioned closer to one another. Embedding vectors have been created using the character n-gram when $1 \leq n \leq 6$ embeddings. The GloVe (Pennington et al. 2014) embedding of size 50 was used for word embedding in text generation. The text was generated using conditional GANs (cGAN) to balance the dataset during in-processing debiasing. The character n-gram features (with a vector size of 100,000) capture nuances like slang, acronyms, abbreviations, and misspellings common in social media platforms (X), while

**Table 1** Pairs of nouns and generic pronouns representing a female and a male person used to label the dataset

| Gender | Noun, generic-pronouns and abusive words |
|---|---|
| Male | Man, men, boy, brother, son, bro..., husband, boyfriend, father, uncle, dad, grandfather, male, nigga, cock, ass, dick, his, him, he, himself, bf, dude, foggot |
| Female | Woman, women, girl, sister, sissy..., daughter, wife, girlfriend, mother, aunt, mom, grandmother, tits, female, bitch, milf, twat, pussy, cunt, babe, her, she, herself, gf |

[2] https://gombru.github.io/2019/10/09/MMHS/.

the GloVe embeddings handle out-of-vocabulary (OOV) words.

### ML classifier

Several machine learning (ML) classifiers, including support vector machine (SVM), random forest (RF), linear regression (LR), decision tree (DT), gradient boosting (GB), and adaptive boosting (AdaBoost) were used to classify the textual tweets into three categories. Features extracted using character n-grams, with $1 \leq n \leq 6$, were used to train the ML classifiers. The dataset was split, with 80% of the tweets used for training and 20% for testing. The experimental results are presented in Table 4. After debiasing, the same ratio was maintained to validate the model across all three scenarios: (i) pre-processing debiasing, (ii) in-processing debiasing, and (iii) a combination of both.

### Fairness evaluation

In this section, fairness metrics are defined, drawing from the fairness literature (Kwiatkowska 1989; Mehrabi et al. 2021), and are calculated based on the confusion matrix of the classification performance. Four metrics are used to assess classification performance: precision (P), recall (R), F1-score (F1), and accuracy (Acc.). Additionally, four metrics are used for evaluating fairness: (i) overall accuracy equality (OAE), (ii) statistical parity (StPs), (iii) conditional procedure accuracy equality, and (iv) treatment equality (ToE).

### Debiasing methods

Debiasing methods help mitigate inherent biases in AI/ML models and promote fair outcomes. These biases often arise from imbalances in training data that reflect historical or societal prejudices. Debiasing techniques aim to correct these biases and improve fairness in the classification process. In this study, both pre-processing and in-processing debiasing methods are applied to reduce bias in the HSC model, as explained below:

1. **Pre-processing debiasing**: Two methods are used in pre-processing debiasing: (i) swapping the gender (GS) and (ii) removing the gender term (RGT). In the first method, we are swapping the gender terms with the opposite gender or their corresponding noun and pronouns. It reduces the correlation between the label and the data sample. The second debiasing method is RGT, in which the gender-related nouns and pronouns have been removed from the dataset.
2. **In-processing debiasing**: In the in-processing debiasing methods, debiasing is done during training time. The data samples of the minority class were generated and made equivalent to those of the majority class during training time. We have balanced the data sample for all the minority classes using cGAN and a large language model (LLM). After balancing the dataset, two new datasets were created. The first dataset was obtained after balancing with cGAN, named BGAN, and the second one was balanced with LLM, named BLLM. Examples of some data samples generated using cGAN and LLM are listed in Table 3.

    - **cGAN:** The architecture of the cGAN used to generate text is shown in Fig. 4. The cGAN consists of two convolutional networks: a generator and a discriminator. A latent vector (label) of size 100 was used in the generator module to produce corresponding text. The generator comprises three transposed convolution layers and two convolution layers, which work together to generate text. It trains iteratively, generating text until the discriminator can no longer distinguish between the generated text and real data samples. The generated and real text samples were fed into the discriminator to intentionally confuse the generator and classify the generated text as real text. The discriminator had two convolution layers, followed by a flattened layer and a dense layer con-

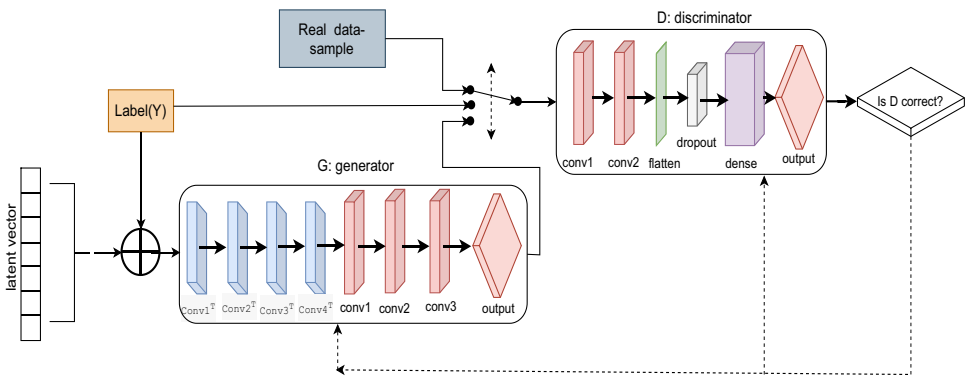


**Fig. 4** Architecture of cGAN

necting the flattened layer to the output layer. The hyperparameters used to train the cGAN model are listed in Table 2.

- **LLM:** LLMs are highly powerful and capable models and effectively generate text similar to human-written text because they are trained on large datasets. We use the Text-to-Text Transfer Transformer (T5), an LLM developed by Google (Raffel et al. 2020), to generate text samples targeted at both males and females.

  We were generating the text to balance a dataset to reduce bias, for which we identified the target classes that we needed to balance. It needed to determine the minority classes (female and male) of the dataset and generate the sample up to the majority class (not targeted to any gender). To generate the text using the T5 model, a few-shot prompts engineering was used to guide the LLMs model. Prompt engineering helps to elicit desired output from language models and ensures that the generated text is accurate, relevant, and aligned with the intended objective. Using this approach, we generated synthetic examples for the minority classes to balance the dataset.

3. **Pre-processing and in-processing debiasing**: In this section, we employed both pre-processing and in-processing debiasing methods to debias the HSC model. For in-processing debiasing, cGAN and LLM techniques were used separately to generate additional data. The generated samples were combined with the original data, creating two distinct datasets: one balanced with GAN (BGAN) and the other balanced with LLMs (BLLMs). Next, pre-processing debiasing methods gender swapping (GS) and removing gender terms (RGT) were applied to both the BGAN and BLLMs datasets. As a result, four datasets were formed: (i) GS+BGAN, (ii) RGT+BGAN, (iii) GS+BLLMs, and (iv) RGT+BLLMs.

**Table 2** Hyper-parameter settings for the proposed modal

| Hyper-parameter | Value |
|---|---|
| Maximum sequence length | 25 |
| Embedding vector length | 50 |
| No. of filter | 128 |
| Filter size | 3, 4 |
| Strides | 2, 4, 8 |
| No. of classes | 3 |
| latent-dim | 100 |
| Activation Function | LeakyReLU, sigmoid |
| Learning-rate | 0.0002 |
| Dropout Rate | 0.2, 0.4 |
| Epochs | 100 |
| Batch size | 10 |
| Loss | Binary-crossentropy |
| Optimizer | Adam |

## Results

This section presents the results obtained from the proposed models using various ML classifiers and the corresponding fairness metrics. The classification performance of the model is evaluated using the metrics described in "Classification performance evaluation metrices" section, while the fairness evaluation metrics are detailed in "Fairness evaluation metrices" section.

### Classification performance evaluation metrices

Three metrics are used to evaluate the performance of the ML classifiers: recall (R), precision (P), and the $F_1$-score ($F_1$). The performance metrics for the men's category are defined below, with similar definitions applied to the other classes.

$$Precision_{(male)} = \frac{\#\,\text{accurately predicted tweets for male}}{\#\,\text{predicted tweets for male}} \quad (1)$$

$$Recall_{(male)} = \frac{\#\,\text{accurately predicted tweets for male}}{\#\,\text{actual tweets for male}} \quad (2)$$

$$F_1 score = \frac{2 * Precision_{(male)} * Recall_{(male)}}{Precision_{(male)} + Recall_{(male)}} \quad (3)$$

### Fairness evaluation metrices

Four fairness metrics are used for evaluation: (i) overall accuracy equality (OAE), (ii) statistical parity (StPs), (iii) conditional procedure accuracy equality, and (iv) treatment of equality (ToE).

$TP_m$ = Number of accurately predicted male targeted tweets
$TN_m$ = Number of accurately predicted none class tweets w.r.t male
$FP_m$ = Number of none targeted tweets predicted as male
$FN_m$ = Number of male targeted tweets predicted as none
$N_m$ = Total numbers of tweets belongings to male and none class
$TP_f$ = Number of accurately predicted female targeted tweets

$TN_f$ = Number of accurately predicted none class tweets w.r.t female
$FP_f$ = Number of none targeted tweets predicted as female
$FN_f$ = Number of female target tweets predicted as none
$N_f$ = Total numbers of tweets belongs to female and none class

Overall Accuracy Equality (OAE) is attained when the proposed model demonstrates equal accuracy across all protected group categories (Male, Female).

$OAE_m$ = Overall Accuracy Equality (OAE) for male
$OAE_f$ = Overall Accuracy Equality (OAE) for female

The overall accuracy equality for male class is calculated as below. The OAE for the female class can be calculate similarly.

$$OAE_m = \frac{\|TP_m + TN_m\|}{N_m} \tag{4}$$

OAED = Overall Accuracy Equality Difference

$$\text{OAED} = \|OAE_m - OAE_f\| \tag{5}$$

Statistical Parity (StPs) is achieved when the proportion of tweets belonging to predicted classes is the same for both protected groups. The disparity between the protected groups should be zero to achieve fairness.

$SP_m$ = Statistical Parity for male class
$SP_f$ = Statistical Parity for female class

$$SP_m = \frac{\|TN_m + FN_m\|}{N_m} \tag{6}$$

$$SP_f = \frac{\|TN_f + FN_f\|}{N_f} \tag{7}$$

$$\text{Statistical Parity Difference (StPsD)} = \|SP_m - SP_f\| \tag{8}$$

Conditional Procedure Accuracy Equality is achieved when either the false negative rate (FNR) and the false positive rate (FPR) will be same for the protected group categories.

– False negative rate for male ($FNR_m$) is defined as the number of male targeted tweets correctly predicted out of the total number of male targeted tweets.

$$FNR_m = \frac{TP_m}{TP_m + FN_m} \tag{9}$$

$$\text{False negative rate difference (FNRD)} = \|FNR_m - FNR_f\| \tag{10}$$

- False positive rate ($FPR_m$) for males is the proportion of tweets where a model incorrectly predicts a positive outcome for males when the true outcome is negative.

$$FPR_m = \frac{TN_m}{FP_m + TN_m} \tag{11}$$

$$\text{False Positive Equality Difference (FPED)} = \|FPR_m - FPR_f\| \tag{12}$$

Similarly we calculate $FNR_f$ and $FPR_f$ for female class.

Treatment of Equality (ToE) is achieved when the ratio of the number of tweets belonging to protected (male, female) group is wrongly predicted and the number of tweets belonging to unprotected (neutral) group.

$$\frac{FN_m}{FP_m} = \frac{FN_f}{FP_f} \tag{13}$$

The fairness evaluation metrices before and after applying the de-biasing methods are listed in the Tables 5 and 6 respectively.

## Classification performance and fairness evaluation with HSGB dataset

Six ML classifiers were applied to the character $n$-gram features, with $1 \leq n \leq 6$ extracted from the initial HSGB dataset. The $F_1$-score, precision, and recall of the top-performing models are listed in Table 4. Gradient Boosting (GB) achieved the highest weighted average (Wt. avg) precision, recall, and $F_1$-score of 0.85, 0.84, and 0.84, respectively, as highlighted in the table.

After obtaining the classification performance metrics for the HSGB dataset from the ML classifiers, we evaluate the fairness metrics to analyze the fairness of the proposed model. The difference in fairness evaluation metrics of both genders is closer to zero or equal to zero and is considered optimal, indicating a higher degree of fairness. Conversely, if the values of the fairness evaluation metrics are far from zero or higher, this indicates a lower degree of fairness. The fairness evaluation metrics are detailed in "Fairness evaluation metrices" section, and the results are shown in Table 5. The best-achieved fairness evaluation metric values are highlighted in Table 5. The DT performs best with OAED and StPsD, achieving values of 3.13 and 3.24, respectively, which are the lowest among all classifiers; however, these values still indicate unfairness since they are not equal to zero. The AdaBoost classifier achieved the lowest value of 0.2 on the ToED fairness evaluation metric, while the FPR difference is 0.01, the same as that achieved by the Gradient

Boosting (GB) classifier. However, the existing fairness evaluation matrices reveal bias in the proposed HSC model, as shown in Table 5.

### Classifier performance and fairness evaluation with debiased dataset

After evaluating the fairness metrics, we observe that the proposed model is unfair. To address this, we debiased the HSC by applying pre-processing and in-processing debiasing methods. In the pre-processing phase, we used the gender swap (GS) and remove gender-related term (RGT) methods. Both methods effectively mitigate gender bias in classification models by eliminating the correlation between data samples and their respective labels, as noted in the study by Mozafari et al. (2020). After debiasing the corpus, we recalculated the classification performance metrics. We observed that the model's fairness improved after applying the pre-processing debiasing methods; however, the model's accuracy decreased by 3–4%. After employing the pre-processing debiasing methods, we found that gender bias was reduced by up to 0.04% in AOED, 0.02% in StPsD, and 0.41% in ToED, whereas the FPRD increased by 0.02%. While the reduction in bias is positive, it is relatively small and may not be considered significant or fair in most scenarios. Therefore, further refinement or exploration of more effective debiasing techniques was necessary.

After applying the in-processing debiasing method with cGAN, the obtained classification performances are listed in Table 4. Once again, the GB classifier performs the best, achieving the highest weighted precision of 0.74, a weighted recall of 0.74, and an $F_1$-score of 0.73, as highlighted in Table 4. The in-processing debiasing methods with the LLM model create a new balanced dataset (BLLM), with the classification performance and fairness evaluation metrics shown in Tables 4 and 6, respectively. GB again demonstrated superior performance compared to all other ML classifiers, achieving the highest weighted precision, recall, and $F_1$-score of 0.83, 0.82, and 0.82, respectively, as highlighted in Table 4. After applying both the in-processing debiasing techniques using cGAN and LLMs, the resulting fairness evaluation metrics were listed in Table 6. The performance of the DT was fairer than that of the other applied ML classifiers. While both in-processing debiasing methods increased the fairness of the proposed model, cGAN performed better than LLMs, can also be seen in Fig. 6.

After applying both debiasing methods individually, we found that the performance of in-processing debiasing with cGAN was better than that of the LLM. The model performed best after applying both the debiasing methods (pre- and in-processing) together. First, we de-biased the dataset using both the pre-debiasing methods and then applied the in-processing debiasing. After applying both debiasing methods and the best-performing models and their corresponding fairness evaluation methods, the results of fairness evaluation are listed in Table 7. GS+LLM performed the best and reduced the bias to almost zero in three fairness matrices: AOED, STPSD, and FPRD.

## Discussion

This study provides systematic mitigation methods for gender bias in the datasets, hate speech and text-based classification models.

Our significant finding is that none of the ML classifiers employed in the proposed model demonstrates fair classification, as shown in Table 5. After examining the fairness evaluation metrics for each classifier, we identified that the DT was relatively fair; however, none achieved complete fairness. This lack of fairness is primarily attributed to training the ML model on an unbalanced and biased dataset.

The second finding is that after applying the debiasing methods, the classification performances of all employed ML models declined, but their fairness improved. Despite this decrease in classification performance, our primary focus is on the model's fairness, which notably increased when utilizing DT classifiers. When applying the in-processing methods exclusively, we observed a discernible reduction in fairness metrics such as StPsD, AOED, and FPRD by 0.29%, 2.48%, and 0.0%, respectively, along with a minimal increase in ToED by 0.06%, as shown on the left side of Table 6. The third finding is that when both debiasing methods were applied simultaneously, the combination of GS and LLM performed better than all the others, as shown in Fig. 7. The in-processing with the LLM model effectively balanced the dataset when the GS pre-processing debiasing method was applied. The primary factor behind this improvement is that the pre-trained LLM (T5) model facilitated text generation and was trained on a large dataset, as shown in Table 3. In contrast, the cGAN model was trained on our created HSGB dataset, which is significantly smaller compared to the dataset used for the pre-trained model. Some examples of generated data samples from the cGAN and LLM models are shown in Table 3.

However, using LLMs to generate abusive comments similar to those in the HSGB dataset is complex because the LLM (T5) models do not generate abusive language. To overcome this issue, we employed prompt engineering by providing abusive words in the prompt and requesting the model to generate sentences using those words. In a zero-shot learning scenario, the model must tackle tasks it hasn't encountered during training. Consequently, the results are less influenced by changes in prompts and more by the training sets on which the model was built. While greater emphasis on the prompt yields slightly better results, significant

advancements can only be achieved through fine-tuning for more specific tasks or through few-shot learning with instructions of that kind, as demonstrated in Fig. 5.

Our fourth finding is that cGAN is more efficient in debiasing the model during training, as shown on the right side of Table 6. The performance of the fairness evaluation improves when using the in-processing debiasing methods, with cGAN outperforming LLM. The LLM-based T5 is trained on a large corpus for text generation only, not specifically for hate speech generation, and it does not generate sentences using the tokens from our HSGB dataset. In contrast, the cGAN is trained on our HSGB dataset to generate sentences using tokens similar to those in the HSGB dataset, but the sentences were often grammatically incorrect.

Our next finding is the in-processing debiasing methods were more effective than the pre-processing debiasing method, as shown in Table 6. The in-processing debiasing method de-bias the model at training time, and training is a very important part of any ML, DL and AI-based models. Addressing the bias at this stage will stop the propagation of the biases in the model because it stops it at training time. The last finding is that the model performs better when both debiasing methods are applied together than when the pre-processing and in-processing debiasing methods are employed individually. This combined approach reduces the model's bias at both the data and algorithmic levels. The pre-processing debiasing method addresses bias at the data level; however, the gender-swap (GS) method outperformed the remove gender term (RGT) method in our proposed work. The in-processing debiasing methods mitigate bias at the algorithmic level, with both cGAN and LLMs demonstrating similar performance, as shown in Table 7.

The proposed model can be used to identify and mitigate bias in AI/ML models used for text classification. Social media administrators can utilize it to validate their models for identifying hate speech, aggression, abusive language, and fake news. Additionally, it can validate models used to classify digester-related tweets from social media texts, where gender may be a hidden key attribute. The methods employed can be beneficial in identifying and mitigating gender bias present in any AI/ML-based classification model where gender is a factor, such as loan approval systems, software used in the admissions process, or job recruitment systems.

The limitation of the proposed work is that the debiasing methods are specific to the English language. The

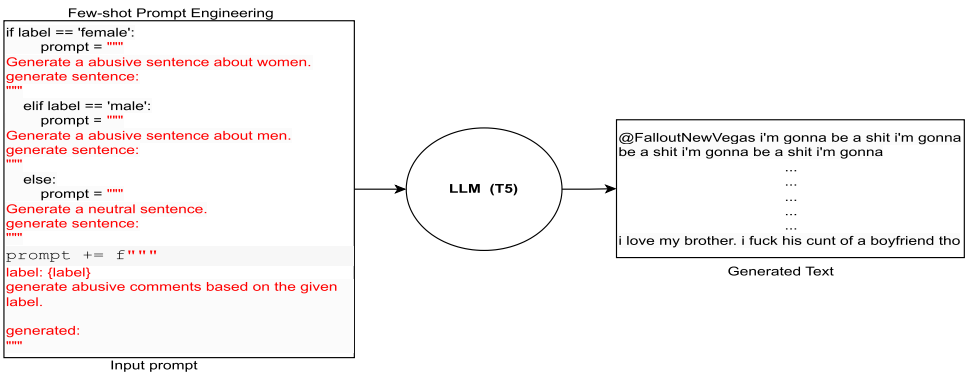


**Fig. 5** Implication of few-shot prompt engineering with LLM (T5)

**Table 3** Genarated samples using cGAN and LLM

| Text generated using cGAN | Class |
|---|---|
| 1. One come his talking still | 0 |
| 2. Booty that for man hair, see nigga | 0 |
| 3. Nothing the jap a grandma other soft stuff thick man his em over came keep instead | 0 |
| 4. Guv successfully but you you this ok which b**ch | 1 |
| 5. Gal magat kissed cunt character tooth but cunt? | 1 |
| **Text generated using LLM** | **Class** |
| 1. @ItsAllBollocks Protect yourself against your son's hobby Mrs. T*at! | 0 |
| 2. I'm gonna f*ck this faggo | 0 |
| 3. I'm a gay sisy faggo with gorgeous feminine legs | 1 |
| 4. I still a b*tch ass nigga | 1 |
| 5. Samaichii ichiro also says she's a c*nt | 1 |

**Table 4** ML Classifiers performance before (HSGB) and after (BGAN and BLLM) applying in-processing debiasing methods

| ML classifier | Class | Before debiasing | | | After debiasing | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | HSGB Dataset | | | BGAN Dataset | | | BLLM Dataset | | |
| | | Precision | Recall | $F_1$-score | Precision | Recall | $F_1$-score | Precision | Recall | $F_1$-score |
| | Male | 0.88 | 0.16 | 0.27 | 0.57 | 0.41 | 0.48 | 0.85 | 0.12 | 0.22 |
| SVM | Female | 0.90 | 0.68 | 0.78 | 0.73 | 0.60 | 0.66 | 0.89 | 0.66 | 0.76 |
| | None | 0.73 | 0.98 | 0.84 | 0.71 | 0.95 | 0.81 | 0.73 | 0.97 | 0.83 |
| | Wt. avg | 0.81 | 0.78 | 0.75 | 0.69 | 0.69 | 0.68 | 0.79 | 0.77 | 0.73 |
| | Male | 0.88 | 0.35 | 0.50 | 0.63 | 0.41 | 0.50 | 0.85 | 0.25 | 0.38 |
| RF | Female | 0.90 | 0.76 | 0.82 | 0.75 | 0.65 | 0.70 | 0.86 | 0.74 | 0.80 |
| | None | 0.78 | 0.96 | 0.86 | 0.72 | 0.97 | 0.83 | 0.77 | 0.95 | 0.85 |
| | Wt. avg | 0.81 | 0.78 | 0.75 | 0.71 | 0.72 | 0.70 | 0.81 | 0.79 | 0.77 |
| | Male | 0.73 | 0.47 | 0.57 | 0.57 | 0.54 | 0.55 | 0.72 | 0.40 | 0.52 |
| LR | Female | 0.81 | 0.76 | 0.79 | 0.73 | 0.66 | 0.69 | 0.81 | 0.75 | 0.78 |
| | None | 0.81 | 0.91 | 0.86 | 0.80 | 0.90 | 0.84 | 0.81 | 0.93 | 0.86 |
| | Wt. avg | 0.80 | 0.81 | 0.80 | 0.72 | 0.72 | 0.72 | 0.80 | 0.80 | 0.79 |
| | Male | **0.76** | **0.56** | **0.64** | **0.61** | **0.56** | **0.58** | **0.78** | **0.45** | **0.57** |
| GB | Female | **0.94** | **0.77** | **0.84** | **0.76** | **0.65** | **0.70** | **0.91** | **0.75** | **0.82** |
| | None | **0.82** | **0.96** | **0.88** | **0.79** | **0.95** | **0.86** | **0.80** | **0.95** | **0.87** |
| | Wt. avg | **0.85** | **0.84** | **0.84** | **0.74** | **0.74** | **0.73** | **0.83** | **0.82** | **0.82** |
| | Male | 0.63 | 0.62 | 0.62 | 0.55 | 0.61 | 0.58 | 0.69 | 0.47 | 0.56 |
| DT | 0.83 | 0.77 | 0.80 | 0.80 | 0.70 | 0.65 | 0.67 | 0.73 | 0.80 | 0.76 |
| | None | 0.84 | 0.88 | 0.86 | 0.83 | 0.84 | 0.84 | 0.83 | 0.85 | 0.84 |
| | Wt. avg | 0.81 | 0.81 | 0.81 | 0.72 | 0.71 | 0.72 | 0.71 | 0.71 | 0.71 |
| | Male | 0.73 | 0.51 | 0.60 | 0.53 | 0.62 | 0.57 | 0.65 | 0.46 | 0.54 |
| Ada | Female | 0.87 | 0.73 | 0.79 | 0.73 | 0.54 | 0.62 | 0.87 | 0.69 | 0.77 |
| Boost | None | 0.79 | 0.93 | 0.86 | 0.79 | 0.92 | 0.85 | 0.79 | 0.93 | 0.85 |
| | Wt. avg | 0.81 | 0.81 | 0.80 | 0.71 | 0.71 | 0.71 | 0.80 | 0.79 | 0.79 |

The bold values in table represent the highest scores achieved among the applied ML classifiers

effectiveness of these techniques may differ when applied to a hate speech classification model in another language, as the pre-processing required for English differs from that for other languages. The absence of cross-linguistic analysis in this study may limit its applicability to the challenges posed by gender bias in hate speech in languages other than English. Additionally, the lack of diverse datasets may restrict the exploration of debiasing methods across different languages and hinder a comprehensive understanding of the issue. This work focuses solely on English data, making it inapplicable to other languages. Furthermore, the in-processing debiasing methods concentrate on generating similar text during training but generating text in languages other than English presents a significant challenge.

**Table 5** Results of fairness evaluation before debiasing methods (HSGB dataset)

| ML Classifier | StPsD | OAED | FPRD | ToED |
|---|---|---|---|---|
| SVM | 4.63 | 5.61 | 0.04 | 31.71 |
| RF | 4.6 | 5.36 | 0.03 | 11.2 |
| LR | 4.07 | 4.58 | 0.07 | 1.69 |
| GB | 3.76 | 4.17 | **0.02** | 2.38 |
| DT | **3.13** | **3.42** | 0.03 | 0.41 |
| Ada boost | 3.89 | 4.29 | 0.03 | **0.2** |

The lowest values in Table indicate the highest fairness

## Conclusion

In this work, we study the effect of debiasing methods to reduce gender bias in the HSC model applied to social media posts. We implemented two debiasing methods concerning gender bias: i) pre-processing and ii) in-processing, using the HSGB dataset. For the pre-processing debiasing methods, we applied two techniques: the first removes gender-related terms from the dataset, while the second swaps gender terms. Next, we applied in-processing debiasing methods, employing two techniques for text generation to

**Table 6** Evaluation of fairness outcomes after implementing pre-processing and in-processing debiasing methods saperately

| ML-Classifier | Pre-processing | OAED | StPsD | FPRD | ToED | In-processing | OAED | StPsD | FPRD | ToED |
|---|---|---|---|---|---|---|---|---|---|---|
| SVM | GS | 4.54 | 5.54 | 0.04 | 69.3 | cGAN | 1.13 | 1.44 | 0.06 | 0.19 |
| | RGT | 4.54 | 5.45 | 0.04 | 69.0 | LLM | 4.2 | 5.24 | 0.04 | 34.94 |
| RF | GS | 4.47 | 5.36 | 0.47 | 21.21 | cGAN | 1.38 | 1.8 | 0.09 | 1.34 |
| | RGT | 4.47 | 5.36 | 0.06 | 21.21 | LLM | 4.19 | 5.11 | 0.04 | 14.6 |
| LR | GS | 3.79 | 4.3 | 0.05 | 1.65 | cGAN | 0.94 | 1.14 | 0.04 | 0.28 |
| | RGT | 3.79 | 4.3 | 0.05 | 1.65 | LLM | 3.73 | 4.31 | 0.06 | 2.3 |
| GB | GS | 3.74 | 4.1 | 0.02 | 0.4 | cGAN | 1.15 | 1.4 | 0.05 | 0.06 |
| | RGT | 3.74 | 4.1 | 0.02 | 0.4 | LLM | 3.69 | 4.24 | 0.02 | 1.27 |
| DT | **GS** | **3.09** | **3.4** | **0.04** | **0.0** | cGAN | **0.84** | **0.94** | **0.03** | **0.47** |
| | **RGT** | **3.09** | **3.4** | **0.04** | **0.0** | LLM | 3.69 | 4.24 | 0.02 | 1.27 |
| Ada | GS | 3.67 | 3.92 | 0.02 | 0.45 | cGAN | 0.98 | 1.11 | 0.03 | 0.51 |
| Boost | RGT | 3.67 | 3.92 | 0.02 | 0.45 | LLM | 3.37 | 4.25 | 0.1 | 1.8 |

The lowest values in Table indicate the highest fairness

**Table 7** Evaluation of fairness outcomes after implementing both pre-processing and in-processing debiasing methods

| ML-classifier | In-processing ⇒ | cGAN | | | | LLMs | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Pre-processing ⇓ | OAED | StPsD | FPRD | ToED | OAED | StPsD | FPRD | ToED |
| SVM | GS | 0.14 | 0.32 | 0.05 | 4.02 | 0.14 | 0.32 | 0.05 | 4.02 |
| | RGT | 0.05 | 0.11 | 0.04 | 1.66 | 0.05 | 0.11 | 0.04 | 1.66 |
| RF | GS | 0.54 | 1.1 | 0.15 | 7.39 | 0.09 | 0.22 | 0.05 | 2.64 |
| | RGT | 0.57 | 1.13 | 0.16 | 8.71 | 0.1 | 0.03 | 0.03 | 1.53 |
| LR | GS | 0.04 | 0.15 | 0.01 | 0.25 | 0.22 | 0.16 | 0.03 | 0.34 |
| | RGT | **0.01** | **0.1** | **0.0** | **0.12** | 0.22 | 0.09 | 0.0 | 0.63 |
| GB | GS | 0.61 | 1.16 | 0.18 | 8.71 | 0.11 | 0.31 | 0.05 | 3.14 |
| | RGT | 0.57 | 1.1 | 0.16 | 6.32 | 0.03 | 0.11 | 0.04 | 1.59 |
| DT | **GS** | **0.0** | **0.13** | **0.01** | **0.26** | **0.1** | **0.0** | **0.0** | **0.42** |
| | **RGT** | 0.11 | 0.28 | 0.04 | 0.54 | 0.24 | 0.11 | 0.01 | 0.38 |
| Ada | GS | 0.55 | 1.28 | 0.26 | 4.89 | 0.56 | 0.91 | 0.16 | 4.39 |
| Boost | RGT | 0.62 | 1.35 | 0.27 | 6.62 | 0.19 | 0.62 | 0.12 | 2.45 |

The lowest values in Table indicate the highest fairness

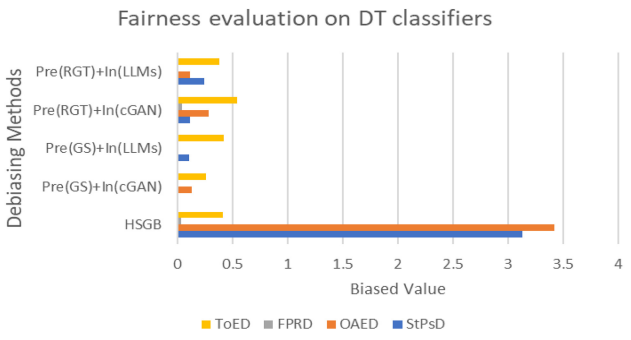


**Fig. 6** The graph representation of fairness matrices of DT classifier

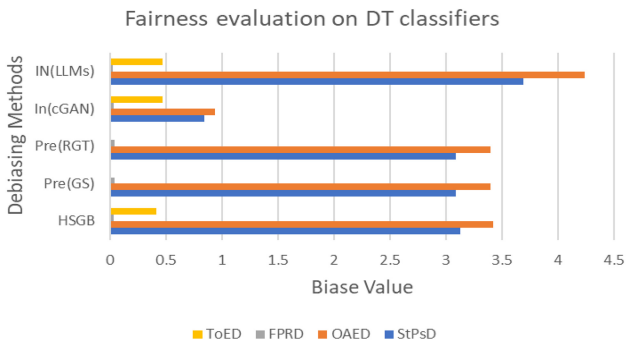


**Fig. 7** The graph representation of fairness matrices of DT classifier

balance the dataset and debias the models during training. The first technique uses cGAN, and the second utilizes prompt engineering with the pre-trained T5 model. The LLM (T5) model was pre-trained on large datasets and did not generate abusive comments targeting specific genders; therefore, the intuitive prompt engineering in ChatGPT for

zero-shot performance could not benefit from additional scrutiny of prompt design to achieve improved results. Given the proprietary nature of ChatGPT's data and model, we used few-shot learning prompt engineering to generate text. The main aim was to mitigate the existing bias in AI/ML-based hate speech classification models. We found that the pre-processing debiasing methods effectively reduced bias in the proposed model during the initial phase. At the same time, the in-processing debiasing methods also decreased bias during training and increased the model's fairness independently. When we simultaneously applied both debiasing methods, the fairness of the HSC model improved compared to the individual application of each method. The proposed methodology contributes to the text-based classification models, and the social media administrator does their practical implication to identify and mitigate the bias in their text/hate speech classification model. The proposed model must be generalized with multilingual and codemix languages to generate the text during training time and reduce bias using in-processing methods. The work is only focused on gender bias, not targeting the other biases, i.e. racist bias. The model should be generalized with the other text classification tasks to reduce bias. Our research can encourage other researchers to dedicate more effort toward minimizing bias in AI-based hate speech classification and identification models.

**Author contributions** All the author has equally contributed.

**Funding** No funding.

**Availability of data and materials** The data that support the findings of this study are available at: https://github.com/gunjankumar20/Gender_Bias-Hatespeech/blob/main/data_sexist.csv.

## Declarations

**Conflict of interest** There is no conflict of interest.

## References

Akter, S., McCarthy, G., Sajib, S., Michael, K., Dwivedi, Y.K., D'Ambra, J., Shen, K.N.: Algorithmic bias in data-driven innovation in the age of AI. Int. J. Inf. Manag. **60**, 102387 (2021)

Angwin, J., Larson, J., Mattu, S., Kirchner, L.: Machine bias. In: Ethics of Data and Analytics, Ed, Kirsten Martin, pp. 254–264. Auerbach Publications, London (2022)

Calmon, F., Wei, D., Ramamurthy, K.N., Varshney, K.R.: Optimized pre-processing for discrimination prevention. CoRR arXiv:1704.03354 (2017)

Chen, L., Dai, S., Tao, C., Zhang, H., Gan, Z., Shen, D., Zhang, Y., Wang, G., Zhang, R., Carin, L.: Adversarial text generation via feature-mover's distance. Adv. Neural Inform. Process. Syst. **31**, 1 (2018)

Chen, T., Li, Z., Wu, J., Ma, H., Su, B.: Improving image captioning with Pyramid attention and SC-GAN. Image Vis. Comput. **117**, 104340 (2022)

Chung, J.J.Y., Kamar, E., Amershi, S.: Increasing diversity while maintaining accuracy: text data generation with large language models and human interventions. Preprint arXiv:2306.04140 (2023)

Dale, R.: GPT-3: What's it good for? Nat. Lang. Eng. **27**(1), 113–118 (2021)

Diao, S., Shen, X., Shum, K., Song, Y., Zhang, T.: TILGAN: transformer-based implicit latent GAN for diverse and coherent text generation. In: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pp. 4844–4858 (2021)

Ding, B., Qin, C., Liu, L., Chia, Y.K., Joty, S., Li, B., Bing, L.: Is GPT-3 a good data annotator? Preprint arXiv:2212.10450 (2022)

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, pp. 67–73 (2018)

Drozdowski, P., Rathgeb, C., Dantcheva, A., Damer, N., Busch, C.: Demographic bias in biometrics: a survey on an emerging challenge. IEEE Trans. Technol. Soc. **1**(2), 89–103 (2020)

Fedus, W., Goodfellow, I., Dai, A.M.: MaskGAN: better text generation via filling in the_. Preprint arXiv:1801.07736 (2018)

Feldman, T., Peake, A.: End-to-end bias mitigation: removing gender bias in deep learning. Preprint arXiv:2104.02532 (2021)

Fortuna, P., Nunes, S.: A survey on automatic detection of hate speech in text. ACM Comput. Surv. (CSUR) **51**(4), 1–30 (2018)

Gomez, R., Gibert, J., Gomez, L., Karatzas, D.: Exploring hate speech detection in multimodal publications. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 1470–1478 (2020)

Kamiran, F., Calders, T.: Data preprocessing techniques for classification without discrimination. Knowl. Inf. Syst. **33**(1), 1–33 (2012)

Karnouskos, S.: Artificial intelligence in digital media: the era of deepfakes. IEEE Trans. Technol. Soc. **1**(3), 138–147 (2020)

Khakurel, U., Abdelmoumin, G., Bajracharya, A., Rawat, D.B.: Exploring bias and fairness in artificial intelligence and machine learning algorithms. In: Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications IV, vol. 12113, pp. 629–638. SPIE (2022)

Kusner, M.J., Paige, B., Hernández-Lobato, J.M.: Grammar variational autoencoder. In: International Conference on Machine Learning, pp. 1945–1954. PMLR (2017)

Kwiatkowska, M.Z.: Survey of fairness notions. Inf. Softw. Technol. **31**(7), 371–386 (1989)

Li, J., Tang, T., Zhao, W.X., Nie, J.Y., Wen, J.R.: Pre-trained language models for text generation: a survey. ACM Comput. Surv. **56**(9), 1–39 (2022)

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. ACM Comput. Surv. (CSUR) **54**(6), 1–35 (2021)

Mozafari, M., Farahbakhsh, R., Crespi, N.: Hate speech detection and racial bias mitigation in social media based on BERT model. PLoS ONE **15**(8), e0237861 (2020)

Nabi, R., Malinsky, D., Shpitser, I.: Learning optimal fair policies. In: International Conference on Machine Learning, pp. 4674–4682. PMLR (2019)

Park, J.H., Shin, J., Fung, P.: Reducing gender bias in abusive language detection. Preprint arXiv:1808.07231 (2018)

Parra, C.M., Gupta, M., Dennehy, D.: Likelihood of questioning AI-based recommendations due to perceived racial/gender bias. IEEE Trans. Technol. Soc. **3**(1), 41–45 (2021)

Pennington, J., Socher, R., Manning, C.D.: GloVe: global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543 (2014)

Peters, D., Vold, K., Robinson, D., Calvo, R.A.: Responsible AI-two frameworks for ethical design practice. IEEE Trans. Technol. Soc. **1**(1), 34–47 (2020)

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. J. Mach. Learn. Res. **21**(1), 5485–5551 (2020)

Reed, S.E., Akata, Z., Mohan, S., Tenka, S., Schiele, B., Lee, H.: Learning what and where to draw. Adv. Neural Inform. Process. Syst. **29**, 1 (2016)

Rezaei, M., Uemura, T., Näppi, J., Yoshida, H., Lippert, C., Meinel, C.: Generative synthetic adversarial network for internal bias correction and handling class imbalance problem in medical image diagnosis. In: Medical Imaging 2020: Computer-Aided Diagnosis, vol. 11314, pp. 82–89. SPIE (2020)

Ryu, H.J., Mitchell, M., Adam, H.: Improving smiling detection with race and gender diversity. arXiv:1712.00193 (2017)

Wang, H., Zhang, S., Li, Y., et al.: Automated diagnosis of bipolar depression through Welch periodogram and machine learning techniques. Proc. Indian Natl. Sci. Acad. **89**, 858–868 (2023). https://doi.org/10.1007/s43538-023-00201-w

Yoo, K.M., Park, D., Kang, J., Lee, S.W., Park, W.: GPT3Mix: leveraging large-scale language models for text augmentation. Assoc. Comput. Ling. **2024**, 2225–2239 (2021)

Zafar, M.B., Valera, I., Gomez Rodriguez, M., Gummadi, K.P.: Fairness beyond disparate treatment & disparate impact: learning classification without disparate mistreatment. In: Proceedings of the 26th International Conference on World Wide Web, pp. 1171–1180 (2017)

Zhao, S., Li, M., Huajin, et al.: Presenting a novel approach based on deep learning neural network and using brain images to diagnose Alzheimer's disease. Proc. Indian Natl. Sci. Acad. **89**, 884–890 (2023). https://doi.org/10.1007/s43538-023-00198-2

Zhuo, T.Y., Huang, Y., Chen, C., Xing, Z.: Red teaming ChatGPT via jailbreaking: bias, robustness, reliability and toxicity. arXiv:2301.12867 (2023)

Zicari, R.V., Brodersen, J., Brusseau, J., Düdder, B., Eichhorn, T., Ivanov, T., Kararigas, G., Kringen, P., McCullough, M., Möslein, F., et al.: Z-inspection®: a process to assess trustworthy AI. IEEE Trans. Technol. Soc. **2**(2), 83–97 (2021)