DLINE JOURNALS

Journal of Digital Information Management

ISSN: 0972-7272

JDIM 2025: 23 (4)

Mitigating Intersectional Gender and Racial Bias in Sentiment Analysis: A T5-Based Data Augmentation Approach for English and Low-Resource Bengali

Md Saiful Islam
Zhongyuan University of Technology
450007. China
saifulzut@hotmmail.com

ABSTRACT

Sentiment analysis models are increasingly deployed in real world applications, but their fairness across demographic groups remains under examined, particularly in low resource languages. In this paper, we examined intersectional bias in sentiment classification across gender and racial subgroups in English and Bengali datasets, using race labels such as White+Black to investigate bias more holistically. To utilize low resource languages, classical machine learning models, such as Logistic Regression (LR), Naïve Bayes (NB), Support Vector Machines (SVM), and Random Forest (RF), were trained using Word2Vec and FastText embeddings. Across the experiments, we observe significant disparities in F1 performance across various gender and race combinations, particularly affecting the Female+Black subgroup. To mitigate performance issues, a T5-based data augmentation strategy was investigated to target underrepresented demographic intersections. Our findings show that augmentation substantially reduces performance gaps and improves overall accuracy by 9.5% absolute, or roughly 9.5% relative. Therefore, this paper demonstrates the critical need for intersectional evaluation in sentiment classification and validates the viability of targeted data augmentation to reduce demographic bias.

Subject Categories and Descriptors:
[**H.3.1 Content Analysis and Indexing];** Linguistic processing: **[H.5.2 User Interfaces];** Natural language: **[I.2.7 Natural Language Processing]** Language models

General Terms: Large Language Models, Sentiment Analysis, Machine-Learning Language, Natural Language Processing

Keywords: Bias and Fairness, Sentiment Analysis, Intersectionality, Large Language Models, Gender And Racial Disparities

Received: 13 March 2025, Revised 12 July 2025, Accepted 26 July 2025

Review Metrics: Review Scale: 1-6, Review Score: 5.2, Inter-reviewer consistency: 84.3%

DOI: https://doi.org/10.6025/jdim/2025/23/4/203-219

1. Introduction

Sentiment analysis assesses the intensity of sentiment in text, which is pivotal in Natural Language Processing (NLP) for interpreting consumer feedback, including monitoring political discourse [1], mental health analysis [2], and social media content [3]. The principal objective of sentiment analysis is to identify the emotional polarization of text into generally positive, negative, or neutral categories according to linguistic patterns. These deployment systems expand into real world, high stakes scenarios, and concerns surrounding their fairness and bias have come to the forefront of academic and industry discourse [4]. Recently, several studies have identified various forms of algorithmic bias in sentiment analysis, driven by factors such as imbalanced data. [5], historical biases, and sociolinguistic variations in training data [4]. These biases lead to inconsistent model performance, with marginalized groups, such as gender and racially minoritized individuals, often experiencing lower accuracy [6]. In this work, we consider a transformer based data augmentation method using a T5 model to improve accuracy.

In NLP, gender and racial bias are increasingly prominent, especially in tasks like coreference resolution. [7], sentiment classification, and occupation association [8]. The study in [9], revealed that sentiment classifiers perform better on texts in the men sub group, reflecting a gender based performance disparity. On the other hand, Racial bias has been identified, as models often overlook understated linguistic, cultural, and stylistic differences among racial groups. [10], resulting in reduced accuracy for racially minoritized individuals. These biases are not just technical limitations and have significant real world implications; ignoring inter sectional factors such as gender and race can cause models to reinforce societal stereotypes instead of mitigating them. Addressing demographic bias in sentiment analysis is essential for creating equitable AI systems [11], and it requires data diversity, fair evaluation of models across subgroups, and an intersectional perspective that recognizes the complexity of identity [7, 8]. Without such efforts, NLP systems may worsen social inequalities rather than foster inclusivity and fairness.

This study examines intersectional gender and race bias in sentiment classification models trained on English and Bengali datasets, including a low-resource Bengali dataset [11], introduces added complexity due to the scarcity of annotated demographic data and the limited availability of pre-trained embeddings [14]. To complete a fair evaluation, we generate demographic subgroups for gender and race labels and provide a controlled setting to measure and mitigate demographic bias. In particular, we evaluate the four classifiers, like Logistic Regression (LR), Naïve Bayes (NB), Support Vector Machine (SVM), and Random Forest (RF), with Word2Vec and FastText embeddings, and also evaluate each model on four intersectional subgroups: Male+White, Male+Black, Female+White, and Female+Black.

To mitigate bias, we proposed a transformer based data augmentation method using a T5 model to generate synthetic training data for underrepresented groups, such as Black women. This method improved model fairness and overall performance, as demonstrated by significant F1-score improvements in post augmentation evaluations. This study contributes to the growing literature on ethical Natural Language Processing (NLP) by rigorously analysing intersectional demographic bias in sentiment analysis, extending fairness research to low resource, multilingual contexts such as Bengali. The proposed augmentation strategy offers a practical solution for reducing demographic disparities in sentiment models while enhancing accuracy and inclusivity.

2. Background

2.1 Gender and racial bias in NLP models

Gender and racial bias in NLP models significantly impact the fairness and accuracy of sentiment analysis applications. Bias occurs when models disproportionately associate emotional tone or occupation with a specific gender [15], for example, linking words like 'compassionate' or 'supportive' more frequently to women. Although racial bias in NLP models has received less attention than gender bias, particularly when models consistently misclassify or underperform on language varieties associated with marginalised racial or ethnic groups [16, 17], such as African American Vernacular English (AAVE) used by Black speakers. These disparities often arise from differences between the language patterns in the training data and those used by underrepresented groups, leading to systematic performance gaps [15], For instance, while a model might correctly classify standard English phrases like 'I'm thrilled', it may misinterpret culturally specific expressions like 'that's sick' or 'that's bad' (meaning something positive in context) from Black users as negative sentiment.

This unruly becomes particularly grave when biases are at intersectional levels, the model may perform moderately well for women or black individuals separately, but fail for Black women due to compounded biases. Unmitigated biases produce exclusionary user experiences in real world applications and reinforce negative stereotypes. [17, 19]. To address these issues, comprehensive solutions are required, including creating bias sensitive algorithms, incorporating fairness aware loss functions and adversarial debiasing, curating balanced datasets that represent a variety of dialects and demographics, and utilising controlled data.

2.2 Intersectionality

Intersectionality emphasizes the compounded biases that arise when multiple identities, such as gender, race, age, and others [20], create compounded biases in NLP systems. In sentiment analysis [21], a model might perform well for white men and moderately well for white women or Black men, but significantly worse for Black women. For example, studies show that [12], a sentiment model trained mainly on texts written by white users. It might accurately classify positive sentiments expressed by white men and white women, and even moderately well for Black men. However, it may struggle with texts from Black women, who often use culturally specific phrases, slang, or dialects such as African American Vernacular English (AAVE). Words like 'bad' or 'sck' might carry positive connotations in specific contexts but be misclassified as negative by models not exposed to these patterns.

This phenomenon has received relatively little attention from NLP research, and only a few NLP studies have examined it systematically [22], particularly in multilingual or low resource. Addressing intersectionality requires building datasets that capture real world language diversity, evaluating performance across combined demographic groups such as black women, Asian women, and designing models that recognize how linguistic features differ at intersections. This approach helps sentiment analysis systems more accurately and fairly reflect the voices of all users. By recognizing intersectionality, sentiment analysis can move beyond superficial fairness and create systems that more accurately and equitably capture the diverse ways people express sentiment across different identities.

2.3 Word Embeddings

Word embeddings are essential to modern NLP activities that transform textual data into numerical

representations that capture semantic meaning. *Word2Vec* and *FastText* are the two most popular wordembedding models [23]. Both represent words as dense vectors, with semantically similar words mapped closer together in the vector space.Word2Vec is a two layer neural network model that learns word embeddings by predicting a target word depending on its context [25, 26], CBOW (Continuous Bag of Words) or the context from a given word, Skip gram (see Fig.1). Through its computational efficiency and strong capability to capture semantic relationships, the CBOW variant is used for sentiment classification. We utilized pre-trained 300-dimensional *Word2Vec* embeddings derived from the Google News dataset for English texts, and custom-trained *Word2Vec* embeddings on Bengali Wikipedia, newswire, and social media data to address low-resource challenges. For sentence representation, we averaged the embeddings of each sentence's keywords, balancing computational efficiency with semantic richness. This embedding technique remains highly beneficial for modeling short to medium length texts in tasks such as sentiment analysis, offering a balance between computational efficiency and representational power.

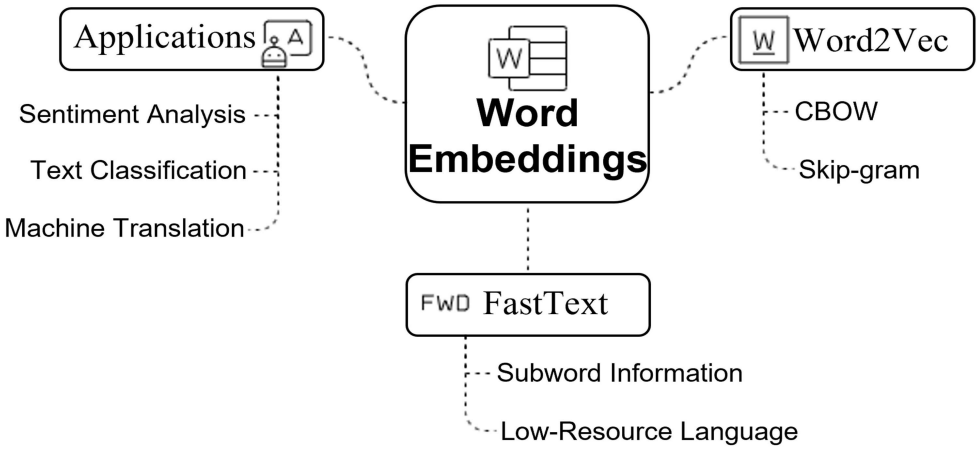

Figure 1. Overview of Word2Vec and FastText embeddings used for sentiment classification

FastText extends Word2Vec by incorporating subword (see Fig. 1) information through character n-grams, which helps capture morphological variations, spelling changes, and compound words common in user-generated text. [27, 28]. In this study, English *FastText* embeddings were sourced from large-scale Common Crawl and Wikipedia corpora (600B tokens). In contrast, Bengali *FastText* embeddings were trained on Indic Corp and OSCAR datasets with 5-character n-grams. *FastText's* subword modelling significantly improved coverage of misspelt and informal terms, making it especially suitable for analyzing sentiment in low-resource and linguistically diverse contexts.

2.4 Bias Mitigation

Bias mitigation in NLP ensures that machine learning systems are fair and equal [28], allowing models to perform consistently well across all demographic groups. Models used for sentiment analysis and other NLP tasks can exhibit biases from imbalanced training data or an insufficient representation of underrepresented groups. [29]. Mitigating bias can be made at one of the three steps of the general machine learning pipeline: the training data, the learning procedure, and the output predictions, with three corresponding categories of bias mitigation algorithms: *pre-processing* approaches. [30], *in-processing* approaches [31], and *post-processing* approaches [32]. *Preprocessing* approaches modify the representation of input data to ensure fair predictive outcomes. For instance, the authors of [33] adopted adversarial fine tuning of word embedding based sentiment analysis by re-training the model using adversarial examples dynamically generated by *BiasFinder*

[34], a tool for detecting biased predictions in sentiment analysis systems. Using various Natural Language Processing (NLP) techniques, *BiasFinder* identifies words associated with a class of a characteristic, like gender-specific words such as female names, 'she', and 'her', and generates new texts from these templates by mutating words.

In-processing (also called in training) approaches revise the training of the state of the art models to achieve fairness. More specifically, they either apply fairness constraints or design an objective function that considers fairness in predictions [35]. Such approaches assume that sensitive attribute information is available in the training samples and enforce fairness during the training process, either by directly imposing fairness constraints and solving constrained optimization problems or by adding penalization terms to the learning objective [30]. *Post-processing* approaches [32], mitigate bias by adjusting model predictions after training, without changing the data or model itself. Standard techniques include modifying decision thresholds for different demographic groups to equalized performance metrics, recalibrating predicted probabilities, or relabeling outputs to satisfy fairness criteria like demographic parity or equalized odds. These methods are practical when retraining is not feasible, though they may slightly reduce overall accuracy to improve fairness across subgroups.

Moreover, bias mitigation in sentiment analysis is significant because it ensures that the sentiments expressed by minority groups are accurately interpreted rather than misclassified or ignored [22, 4]. Demographic aware data augmentation, fairness constraints, and adversarial debiasing are frequently used to enhance model fairness. [35], making models fairer without losing overall accuracy. Through implementing these mitigation strategies, NLP models can more accurately represent diverse populations and reduce the risk of perpetuating harmful stereotypes or reinforcing systemic inequalities.

3. Methodology

The process begins with analyzing demographic bias arising from dataset characteristics, pre-processing, word embeddings, and model choices, leading to unequal performance across subgroups such as *Male+White* vs. *Female+Black*. To address this, synthetic samples are generated for underrepresented subgroups using T5 fine tuning and prompt based data augmentations (see Fig. 2). These synthetic samples are then combined with original training data, resulting in reduced demographic bias and improved performance for underrepresented groups, as validated by balanced test sets and fairness aware evaluation.

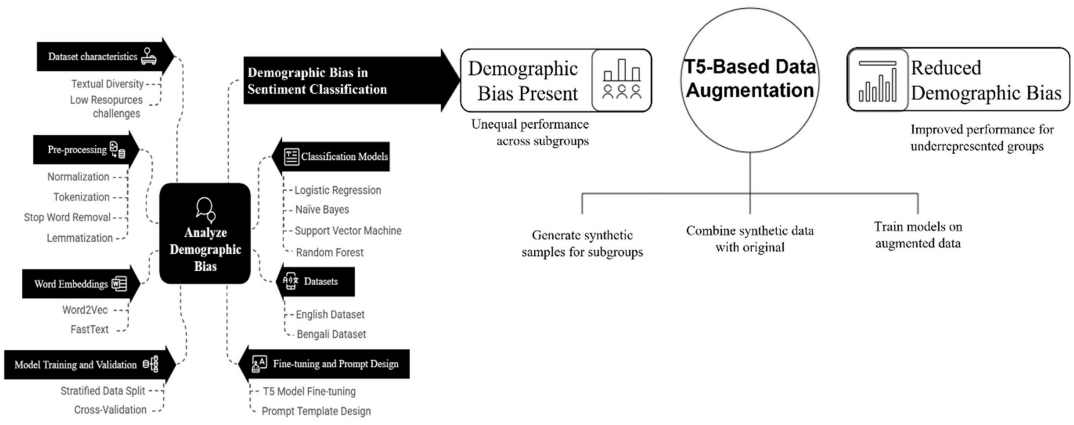

Figure 2. Framework for detecting and mitigating demographic bias in sentiment classification using T5-based data augmentation

3.1 Dataset Description and Pre-processing

To investigate demographic bias in sentiment classification, we used two textual datasets, English and Bengali. The English dataset, comprising 50,000 user generated reviews labelled as positive, negative, or neutral, was sourced from IMDb and Amazon. Although explicit demographic metadata, such as gender and race, are available, the lexical and syntactic diversity within the text allowed for demographic inference. The Bengali dataset consists of 20,000 manually annotated entries collected from local e-commerce platforms, news comment forums, and product discussion forums. As a low-resource language, Bengali presented challenges such as non standard spellings, colloquial grammar, and code mixed Bengali English text. The same three class sentiment labelling scheme was used to ensure comparability with the English dataset.

Both Bengali and English datasets underwent a multi-step pre-processing pipeline to ensure data consistency and quality, which began with the typical normalization steps of lowercasing, removal of punctuation, non-Unicode characters, HTML tags, and emojis. Tokenization was language specific; *spaCy* was used for English, while the Indic NLP library was used to process Bengali since it is optimal for Indic scripts. Stop words were filtered out according to standard lists, and lemmatization was performed wherever required. In the Bengali corpus, significant effort was made to mitigate vocabulary sparsity by correcting non standard spellings and normalizing script variation. All entries were filtered to include at least 5 words, thereby ensuring semantic depth. The datasets were divided into training (70%), validation (15%), and testing (15%) sets using stratified sampling techniques. The test sets were balanced across gender and racial groups for fair bias evaluation. In contrast, the training sets were intentionally biased to reflect real-world demographic disparities and assess performance differences.

3.2 Classification Models

Classification models in NLP are machine learning or deep learning models that categorize text into predefined classes or labels, and these models are widely used in sentiment analysis [37], spam detection, topic labelling, intent recognition, and hate speech detection tasks [39]. In this research, we employed four popular classification models to assess sentiment analysis performance and demographic bias, including Logistic Regression (LR), Naïve Bayes (NB), Support Vector Machine (SVM), and Random Forest (RF), as shown in Fig. 3.

Logistic Regression is a linear classifier that predicts the sentiment class probabilities by computing a sigmoid or *softmax* of a linear function of features [39]. Its interpretability and simplicity make it worthwhile for bias analysis, since feature coefficients can be inspected directly. Naïve Bayes, a generative Bayes' Theorem based model, is based on the assumption of feature independence and performs well in high dimensional, sparse environments [40]. For consistency, we applied the Multinomial NB variant to dense embeddings. As a margin-based classifier, SVM constructs optimal separating hyperplanes in high dimensional space. A linear SVM (LinearSVC) with hinge loss and grid optimized regularization parameter C [33, 34, and hinge loss was utilized in conjunction with a one vs rest multiclass strategy with class weighting to address the problem of label imbalance. Lastly, Random Forest is an ensemble of decision trees that learns nonlinear relations and mitigates overfitting through majority voting. RF models were configured with 100 trees and a maximum depth of 20 [35, 36], and enabled a hierarchical decision making process that was beneficial for identifying bias beyond linear separability.

3.3 Model Training and Validation Protocol

All classifiers were realized with the Scikit learn machine learning library (version 1.x). The models were

trained on static embedding representations (*Word2Vec* or *FastText*), and any changes to the embeddings were disallowed throughout the training. All classifiers were trained on a stratified dataset split into training, validation, and test sets in proportions of 70%, 15%, and 15%, respectively. The hyper parameters of the models were selected using 5 fold cross validation on the validation set, with the macro averaged F1 score as the performance metric. Statistical robustness was ensured by running each experimental configuration three times with varying random seeds, and the final result was presented as the mean F1 score across runs.

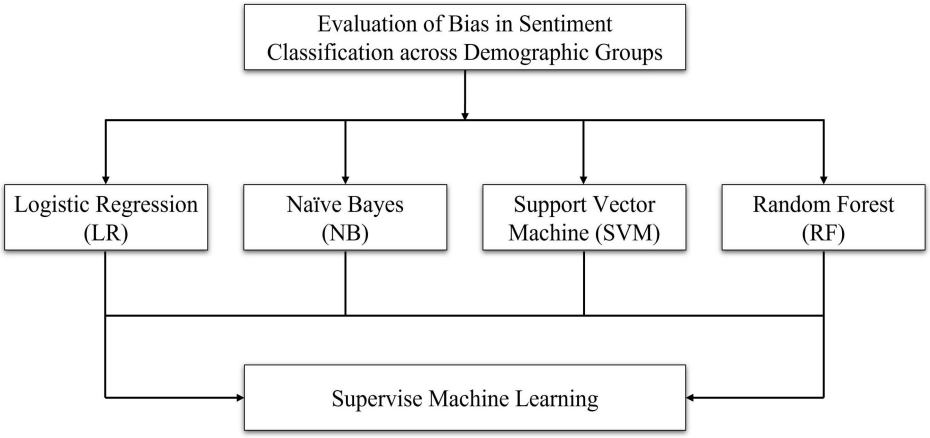

Figure 3. Evaluation framework using supervised learning models to assess bias in sentiment classification across demographic subgroups

3.3 Model Training and Validation Protocol

All classifiers were realized with the Scikit learn machine learning library (version 1.x). The models were trained on static embedding representations (*Word2Vec* or *FastText*), and any changes to the embeddings were disallowed throughout the training. All classifiers were trained on a stratified dataset split into training, validation, and test sets in proportions of 70%, 15%, and 15%, respectively. The hyperparameters of the models were selected using 5-fold cross validation on the validation set, with the macro averaged F1 score as the performance metric. Statistical robustness was ensured by running each experimental configuration three times with varying random seeds, and the final result was presented as the mean F1 score across runs.

3.4 Fine-tuning and Prompt Design

To address demographic bias, we fine-tune a pre-trained T5-base model for English and an *mT5*-base model for Bengali using sentiment labeled datasets to enhance prompt design. Fine-tuning was performed using maximum likelihood estimation with early stopping based on validation loss. The model was trained to generate review texts from structured prompts such as:

"Generate a [positive/negative] review written by a [Black/White] [man/woman] about a [product/topic]". To maintain both demographic and semantic coherence, we incorporated the following controls:

- Sentiment control: Explicitly specifying whether the review should convey a positive or negative sentiment.

- Demographic control: The prompt template includes race and gender identities.

- Topic anchoring: A product, movie, or service that fits into a category is included for domain alignment. Approximately 600 T5-generated samples were generated for the most underrepresented subgroup (*Female*

+ *Black*) to balance the training distribution with the majority subgroup (Male + White, 800 samples). The decision was guided by demographic balancing and validated through improvements in the F1 score in model evaluations.

Moreover, an initial automated filter using sentiment classifiers and keyword heuristics was followed by manual validation of 30% of the generated data. Criteria included grammatical correctness, demographic alignment, and sentiment accuracy. Human validators were NLP trained research assistants with expertise in sentiment analysis and fairness evaluation. Potential biases among human validators were low, and the Inter annotator agreement was 91%. This approach ensured both the linguistic integrity and demographic relevance of synthetic samples, supporting targeted bias mitigation in downstream classification models.

A primary methodological concern is that demographic labels (gender, race), which are not explicitly available in the datasets, are inferred from text, likely using name based or linguistic proxies. This eliminates the potential for misclassification, especially in multilingual contexts such as Bengali, where naming conventions may not align with Western racial or gender categories.

4. Experiment Setup

The process begins with demographically imbalanced training data, which is pre-processed and augmented with demographic attributes to obtain textual features using *Word2Vec* or *FastText* embeddings. We train and evaluate four classification models on the dataset: *LR, NB, SVM,* and *RF*. To assess bias, a balanced test dataset with equal representation across intersectional groups (e.g., *Male+White, Male+Black, Female+White*, and *Female+Black*) is used; performance disparities across these subgroups indicate bias. To categorize significant bias, we implement a T5-based data augmentation to generate additional synthetic samples for underrepresented groups, such as *Female+Black*. When the classification models are retrained, these synthetic samples are combined with the original training sets. Finally, re evaluate the retrained model on the same test set to assess the improvements in fairness and overall accuracy.

The demographic composition of the training and test sets used in the experimental setup was evaluated to evaluate model bias. The training set shows a significant demographic imbalance in the real world dataset; it contains a larger proportion of *Male+White* samples, approximately 42% (800 samples), while *Female+Black*

Demographic Subgroup	Training Set	Test Set
Male+White	800	500
Male+Black	600	500
Female+White	400	500
Female+Black	200	500

Table 1. Simulated demographic distribution showing training set imbalance and balanced test set across intersectional subgroups

samples account for only 6% (200 samples). In contrast, the test set is designed to have an equal number of samples from each of the four groups, 25% (500 samples) across all demographic subgroups (see Table 1), allowing us to assess any bias in the model fairly. Training and test datasets serve two critical purposes:(1) they preserve the natural biases present in real world data during model training, and (2) they provide an unbiased benchmark for detecting learned algorithmic biases through controlled evaluation. This design's effectiveness is demonstrated by its ability to reveal significant performance disparities among groups that are equally represented in testing, with the most pronounced gaps typically appearing for inter sectionally underrepresented groups, such as *Female+Black* individuals.

5. Results

5.1 Baseline Bias (Pre-Augmentation)

Our initial set of experiments for the English dataset focused on identifying bias in sentiment classification performance across demographic subgroups using unaugmented training data, reflecting the natural demographic imbalances in real world datasets. Fig. 4 (a-b) illustrates F1 score distributions across demographic subgroups for sentiment classification models using Word2Vec (see Fig. 4a) and *FastText* (see Fig. 4b) embeddings before augmentation. Four distinct models, such as RF, SVM, NB, and LR, demonstrate a consistent performance hierarchy, with *Word2Vec Male+White* samples achieving the highest scores (0.86) in RF, and Female + Black samples achieving the lowest scores (0.72) in NB. For *FastText Male+White* samples achieve the highest scores (0.87) in RF, and Female + Black samples achieve the lowest scores (0.70) in NB. The lowest scores were typically recorded for the Female + Black subgroup, which showed an intersectional disadvantage.

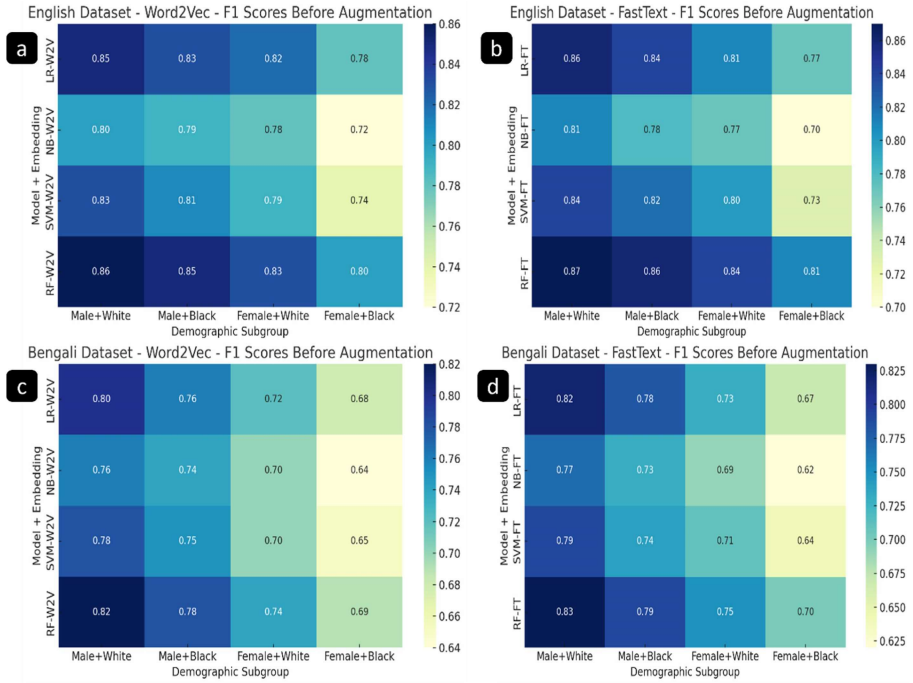

Figure 4. Pre-augmentation F1-score heatmaps across demographic subgroups for English (a–b) and Bengali (c–d) using Word2Vec (a, c) and *FastText* (b, d) embeddings

The Bengali dataset demonstrated significant disparities in model performance across demographic subgroups, particularly when using *Word2Vec* and *FastText* embeddings with RF, SVM, NB, and LR classifiers, as shown in Fig. 4(c-d). For **Word2Vec,** the F1 scores range from 0.64 to 0.82 (see Fig. 4c), indicating a clear bias

toward accommodating the gender and race subgroups. In Fig. 6d, the *Male + White* subgroup consistently achieves the highest scores (0.82), while the Female + Black subgroup shows the lowest performance, with scores as low as 0.64. For *FastText*, the F1 scores range from 0.62 to 0.83, demonstrating a clear bias in favour of the gender and race subgroups, with the Male + White subgroup consistently achieving the highest scores (0.83) and the *Female + Black* subgroup performing the worst, with scores as low as 0.63.

5.2 Bias Mitigation: T5-Based Post-Augmentation

Our experiments determine that targeted T5-based augmentation effectively reduces demographic performance disparities across English and Bengali datasets, as shown in Fig. 5. In the English dataset with *Word2Vec* embedding, the F1 score for the demographic subgroup Female + Black increased from 0.80 to 0.84 (see Fig. 5a), which is closing the gap with the Male + White demographic subgroup (0.86) in just 0.02. Using *FastText* embedding, the Female + Black demographic subgroup improved from 0.68 to 0.79, reducing the bias gap from 0.12 to 0.03 (see Fig. 5b). The improvement in performance for *Female + Black* and the narrowing gaps across all demographic subgroups.

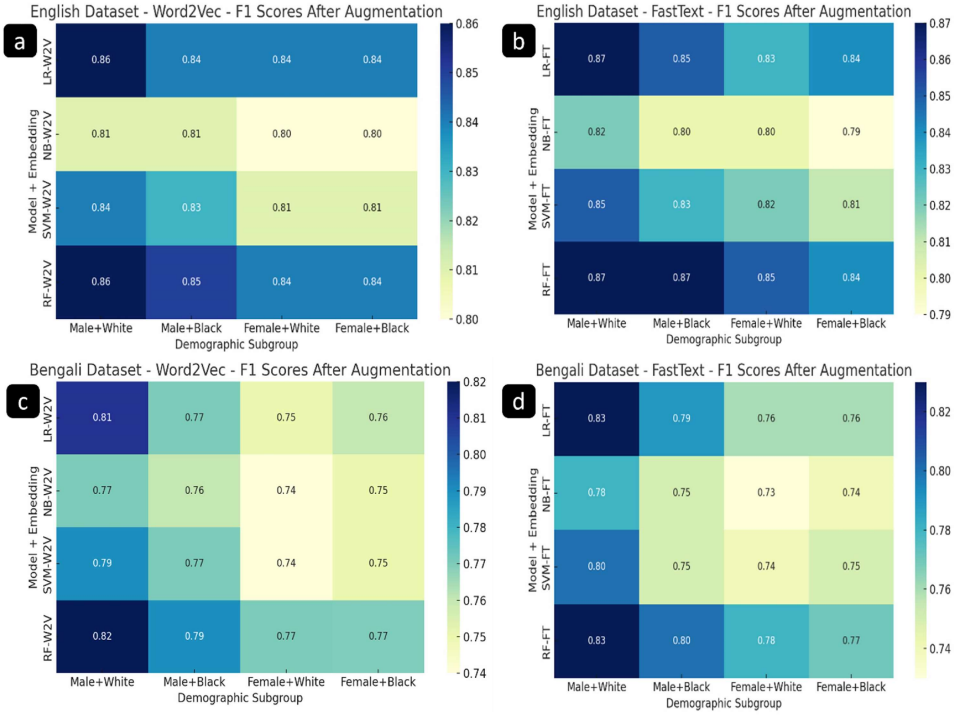

Figure 5. Post-augmentation F1-score heatmaps across demographic subgroups for English (a–b) and Bengali (c–d) using Word2Vec (a, c) and *FastText* (b, d) embeddings

In the Bengali dataset with Word2Vec embedding, the F1 score for the demographic subgroup Female + Black increased from 0.75 to 0.77, which is closing the gap with the *Male + White* demographic subgroup (0.82) in just 0.05 (see Fig. 5c). Using *FastText* embedding, the Female + Black demographic subgroup also improved from 0.74 to 0.77, reducing the bias gap from 0.09 to 0.06 (see Fig. 5d). The improvement in performance for *Female + Black* and the narrowing gaps across all demographic subgroups.

5.3 Bias Gap Pre and Post Augmentation

In both the English and Bengali datasets, a comprehensive analysis of identifying bias reduction through T5-based augmentation was conducted by comparing pre and post augmentation gaps between the highest (*Male+White*)

and lowest (*Female+Black*) F1 scores across models. Pre-augmentation phase, bias gaps range from 0.06 to 0.13, with the most significant disparities occurring in models that use Naïve Bayes or SVM with *FastText* (see Fig. 6). These results strengthen the earlier observation that simpler models like NB or overly lexical embeddings such as *FastText* may exacerbate demographic disparities due to overfitting to dominant group patterns. Bias is generally more pronounced in the Bengali dataset, accentuating the compounding effect of under representation in low-resource settings.

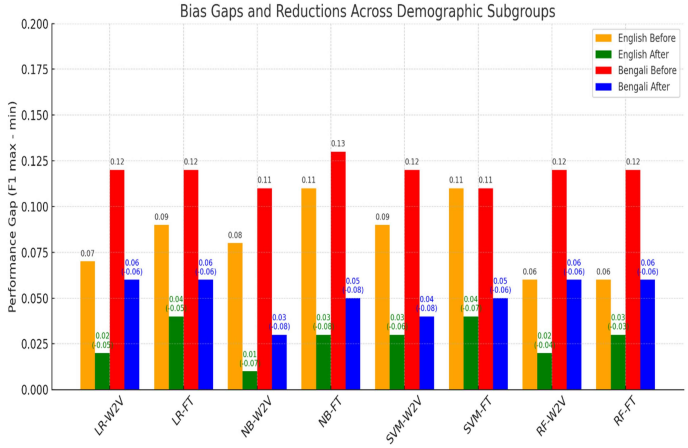

Figure 6. Impact of T5-based augmentation on reducing F1-score bias gaps across models and languages for intersectional demographic subgroups

Applying targeted T5-based data post-augmentation results demonstrates theatrical improvements. Nearly all model configurations show a substantial reduction in the bias gap, with several dropping below the 0.05 threshold often used to indicate acceptably fair performance. The most skewed models, such as NB-FT and SVM-W2V, saw a bias gap drop by up to 0.09, demonstrating the augmentation's impact on fairness. The figure also clearly shows that gap reductions were more pronounced in Bengali, where the added synthetic samples likely provided the most outstanding linguistic and demographic balance. The inclusion of different labels highlights that the most significant improvements occurred for the most disadvantaged subgroups, like *Female+Black*, who benefited the most from the intervention. This ensures that T5-based augmentation delivers consistent, targeted benefits to underserved groups without destabilizing overall model performance. The visual contrast between the pre-and post-augmentation bars in Fig. 7 powerfully reinforces that synthetic data can most effectively mitigate intersectional bias where it is most severe, particularly in resource constrained settings.

5.4 Bias Gap Between Majority and Minority Subgroups

In the bias mitigation experiment that utilized a T5-based model, the majority subgroup (*Male+White*) performed significantly better than the minority subgroup (*Female+Black*) in both pre-and post-augmentation F1 scores, as shown in Fig. 7. Before augmentation, the *Male+White* subgroup substantial performance to the *Female+Black* subgroup, achieving F1 scores of 0.80 compared to 0.68, resulting in a gap of 0.12. However, after augmentation, the minority group's performance improved significantly to 0.79, narrowing the gap to just 0.03, while the majority group's score increased slightly to 0.82. This 75% reduction in the performance disparity (from 0.12 to 0.03) demonstrates that targeted synthetic data generation can effectively mitigate representation bias while maintaining the majority group's performance. Therefore, demographic aware augmentation can achieve near parity in model outcomes across subgroups without compromising overall accuracy, marking a critical step forward in building more equitable NLP systems.

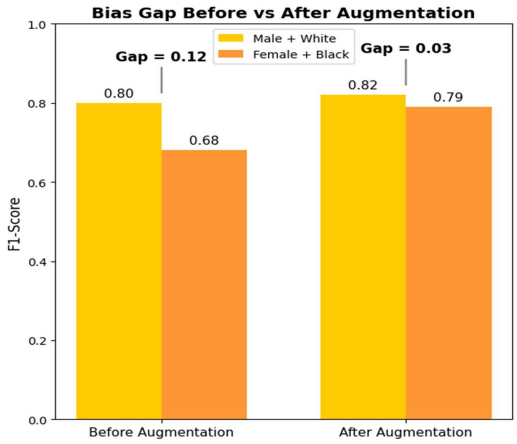

Figure 7. F1-score bias gap reduction between majority (*Male+White*) and minority (*Female+Black*) subgroups following T5-based data augmentation

5.5 Overall Performance Comparison

Figure 8 indicates an apparent positive effect of augmentation on model performance. The macro averaged F1 score on the English dataset increased moderately from 0.81 to 0.84 (+3 points), confirming that synthetic data improved model utility without introducing noise. In the Bengali dataset, which started at a lower baseline (0.74), it showed additional gains, rising from 0.74 to 0.81 (+7 points), highlighting the substantial impact of augmentation in low-resource settings where data scarcity overwhelmingly affects baseline performance. These performance gains were achieved while minimizing demographic bias, as discussed in the previous figures and analysis. The simultaneous improvement in fairness and accuracy refutes the often-cited trade-off in debiasing literature. It suggests that carefully targeted interventions, such as T5-generated samples for underrepresented groups, can enhance equity and utility in NLP models. The findings confirm that T5-generated samples for underrepresented groups can improve equity (by closing subgroup gaps) and efficacy (by boosting overall scores), providing a scalable solution for inter sectionally fair NLP across diverse linguistic settings.

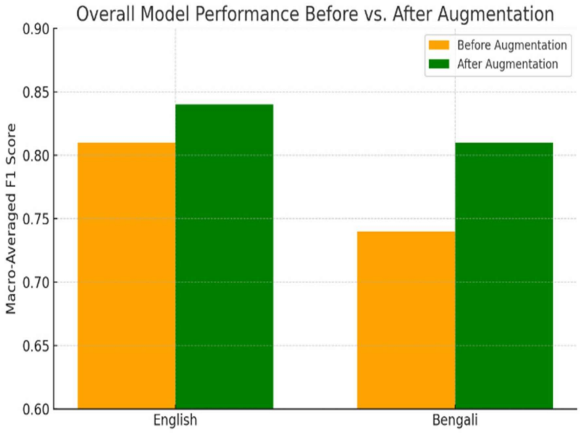

Figure 8. Macro-averaged F1 score improvement after T5-based augmentation on English and Bengali datasets

6. Discussion

6.1 Effects of T5-Based Augmentation on Bias

Our results demonstrate that targeted T5-based data augmentation significantly mitigates intersectional bias,

especially for the most disadvantaged *Female+Black* subgroup (see Fig. 4-5). This approach proved particularly impactful in Bengali, where constraints of low-resource datasets and demographic underrepresentation led to significant initial disparities. For instance, *LR+FastText* embeddings improved F1 score (0.68'™0.79) for *Female+Black* in Bengali, reducing the performance gap by 67% while maintaining the accuracy of the majority group. Similarly, the English dataset saw consistent gains, with RF + *Word2Vec* reducing the *Female+Black* gap from 6% to just 2%. These improvements occurred without compromising performance on the majority subgroups (*Male+White* scores remained stable within ±1%), challenging the common assumption of a fairness-accuracy trade off. The greater relative improvements in Bengali underscore how data augmentation disproportionately benefits low-resource settings where manual data collection is most constrained. These findings indicate that transformer based synthetic data generation can effectively address representation imbalances while enhancing overall model robustness. This method's success across both languages and various classifier architectures suggests it is a practical, scalable solution for mitigating intersectional bias in diverse NLP applications.

6.2 Impact on Overall Performance and Fairness

T5-based augmentation achieved dual benefits, simultaneously reducing intersectional bias while improving overall model performance. As shown in Figure 8, the macro-averaged F1 score across all subgroups increased post-augmentation for both languages. The average F1 score in English increased from 0.81 pre-augmentation to 0.84 post-augmentation, while in Bengali, it improved from 0.74 to 0.81. This augmentation strategy successfully reduced bias gaps without sacrificing performance, potentially enhancing the model's overall effectiveness and achieving balanced improvement across both areas. Figure 8 quantitatively summarizes the bias gap reduction, indicating that augmentation decreased the F1 score gaps by roughly 45% on average across all configurations, with the most significant improvements (up to 70% gap reduction) for the most disadvantaged (*Female+Black*) group. These results emphasize that targeted data augmentation can substantially improve equity in model performance. In summary, the discussion of Figures 4–5 presents a consistent narrative. Pre-augmentation, models exhibit significant intersectional bias, particularly in low-resource languages, while post-augmentation, these biases are notably reduced, and the overall model efficiency is slightly improved. This demonstrates the effectiveness of the proposed T5-based augmentation pipeline in addressing demographic bias in sentiment analysis without sacrificing accuracy.

6.3 Future Work

Future research should focus on incorporating real world demographic labels through ethical data collection methods, expanding beyond binary categories to include non binary identities and additional dimensions such as age and socio economic status, and exploring hybrid approaches that combine data augmentation with algorithmic debiasing techniques like adversarial training or fairness constrained optimization. Addi tionally, while this study focused on classical models, extending this work to transformer based architectures such as BERT and RoBERTa in multilingual contexts, and conducting longitudinal studies of bias evolution in continuously updated models represent critical next steps. This underscores the urgent need for standardized evaluation protocols and community developed best practices for assessing and mitigating demographic bias in NLP systems, particularly for low resource languages where the risks of inequitable performance are most acute, yet the resources for addressing them are limited.

We also plan to conduct a sensitivity analysis to show how performance varies across different demographic assignment assumptions and to validate naming conventions.

7. Conclusion

This study thoroughly investigates intersectional demographic bias in sentiment analysis across English and Bengali datasets. The thorough evaluation of model configurations and subgroup analyses exposes consistent performance disparities, especially within the *Female+Black* subgroup, which shows the lowest accuracy and emphasizes the critical need for intersectional approaches beyond single axis fairness assessments. Our results show that traditional ML models, Strained on biased data, exacerbate demographic imbalances despite high accuracy.

Specifically, Naïve Bayes and SVM models display significant gaps in F1-scores across subgroups, while *FastText* embeddings, though effective for rare words, sometimes amplify biases, especially in Bengali. To address these challenges, we propose a T5-based data augmentation framework that reduces bias gaps by up to 70% without sacrificing accuracy. Bengali datasets are experiencing the most significant improvement due to their low-resource nature. By demonstrating that synthetic data generation can enhance both model performance and equitable representation, this work advances algorithmic fairness in NLP. It also expands research into multilingual and intersectional domains, offering a practical approach for developing more inclusive sentiment analysis systems.

References

[1] Albladi, A., Islam, M., Seals, C. (2025). Sentiment Analysis of Twitter Data Using NLP Models: A Comprehensive Review, *IEEE Access*, vol. 13, p. 30444–30468. doi:10.1109/ACCESS.2025.3541494.

[2] He, M. (2024). Enhancing E-Health with Natural Language Processing: The Role of Sentiment Analysis in Modern Healthcare, *Trans. Comput. Sci. Intell. Syst. Res.*, vol. 7, p. 285–290, Nov. doi:10.62051/1b 4mdm36.

[3] Lak, A. J., Boostani, R., Alenizi, F. A., Mohammed, A. S., Fakhrahmad, S. M. (2024). RoBERTa, Res Next and Bilstm with self attention: The ultimate trio for customer sentiment analysis, *Appl. Soft Comput.*, vol. 164, p. 112018, Oct. doi:10.1016/j.asoc.2024.112018.

[4] Venugopal, J. P., Subramanian, A. A. V., Sundaram, G., Rivera, M., Wheeler, P. (2024). A Comprehensive Approach to Bias Mitigation for Sentiment Analysis of Social Media Data, *Appl. Sci.*, 14 (23). p. 11471, Dec. doi:10.3390/app142311471.

[5] Corizzo, R., Hafner, F. S. (2024). Mitigating social bias in sentiment classification via ethnicity aware algorithmic design, *Soc. Netw. Anal. Min.*, 14 (1) p. 208, Oct. doi:10.1007/s13278-024-01369-9.

[6] Iqbal, M., Karim, A., Kamiran, F. (2019). Balancing Prediction Errors for Robust Sentiment Classification, *ACM Trans. Knowl. Discov. Data,* 13 (3). p. 1–21. Jun. doi:10.1145/3328795.

[7] Zhao, J., Wang, T., Yatskar, M., Ordonez, V., Chang, K. W. (2018). Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods, *In:* Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human LanguageTechnologies, Volume 2 (Short Papers), New Orleans, Louisiana: *Association for Computational Linguistics*, doi:10.18653/v1/n18-2003.

[8] Mandis, I. S. (2021). Reducing Racial and Gender Bias in Machine Learning and Natural Language Processing Tasks Using a GAN Approach, *Int. J. High Sch. Res.*, 3 (6) p. 17–24, Dec. doi:10.36838/v3i6.5.

[9] Gao, L., Zhan, H., Sheng, V. S. (2023). Mitigate Gender Bias Using Negative Multi-task Learning, *Neural Process. Lett.*, 55 (8) p. 11131–11146, Dec. doi:10.1007/s11063-023-11368-0.

[10] Garb, H. N. (2021). Race bias and gender bias in the diagnosis of psychological disorders, *Clin. Psychol. Rev.*, 90 p. 102087, Dec. doi:10.1016/j.cpr.2021.102087

[11] Ding, Y., You, J., Machulla, T. K., Jacobs, J., Sen, P., Höllerer, T. (2022). Impact of Annotator Demographics on Sentiment Dataset Labeling, *Proc. ACM Hum-Comput. Interact.*, 6, no. CSCW2, p. 1–22, Nov. doi: 10.1145/3555632.

[12] Madaio, M., Egede, L., Subramonyam, H., Wortman Vaughan, J., Wallach, H. (2022). Assessing the Fairness of AI Systems: AI Practitioners' Processes, Challenges, and Needs for Support, *Proc. ACM Hum.Comput. Interact.*, 6, no. CSCW1, p. 1–26, Mar. doi:10.1145/3512899.

[13] Wang, A., Ramaswamy, V. V., Russakovsky, O. (2022). Towards Intersectionality in Machine Learning: Including More Identities, Handling Underrepresentation, and Performing Evaluation, in *2022 ACM Conference on Fairness Accountability and Transparency*, Seoul Republic of Korea: *ACM,* Jun. p. 336–349) doi:10.1145/3531146.3533101.

[14] Garcia, N., Hirota, Y., Wu, Y., Nakashima, Y. (2023). Uncurated Image Text Datasets: Shedding Light on Demographic Bias, *In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Vancouver, BC, Canada: *IEEE,* Jun. 2023, p. 6957–6966. doi:10.1109/cvpr52729.2023.00672.

[15] Bhardwaj, R., Majumder, N., Poria, S. (2021). Investigating Gender Bias in BERT, *Cogn. Comput.*, 13 (4) p. 1008–1018). Jul. doi:10.1007/s12559-021-09881-2.

[16] Hamel, E., Kani, N. (2024). Factors That Influence Automatic Recognition of African American Vernacular English in Machine Learning Models, *IEEEACM Trans. Audio Speech Lang. Process.*, vol. 32, p. 509–516, doi:10.1109/TASLP.2023.3331139.

[17] Czarnowska, P., Vyas, Y., Shah, K. (2021). Quantifying Social Biases in NLP: A Generalization and Empirical Comparison of Extrinsic Fairness Metrics, *Trans. Assoc. Comput. Linguist.*, vol. 9, p. 1249–1267, Nov. doi:10.1162/tacl_a_00425.

[18] A Field *et al.* (2023). Examining risks of racial biases in NLP tools for child protective services, in *2023 ACM Conference on Fairness, Accountability, and Transparency*, Chicago IL USA: *ACM,* Jun., p. 1479–1492. doi:10.1145/3593013.3594094.

[19] Turner Lee, N. (2018). Detecting racial bias in algorithms and machine learning, *J. Inf. Commun. Ethics Soc.*, 16 (3) p. 252–260, Aug. doi:10.1108/JICES-06-2018-0056.

[20] Lalor, J., Yang, Y., Smith, K., Forsgren, N., Abbasi, A. (2022). Benchmarking Intersectional Biases in NLP," *In:* Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Seattle, United States: Association for Computational Linguistics, doi:10.18653/v1/2022.naacl-main.263.

[21] Luitel, S., Liu, Y., Anwar, M. (2025). Investigating fairness in machine learning based audio sentiment analysis, *AI Ethics*, 5 (2) p. 1099–1108, Apr. doi:10.1007/s43681-024-00453-2.

[22] Chi, Z., Huang, H., Liu, L., Bai, Y., Gao, X., Mao, X. L. (2024). Can Pretrained English Language Models Benefit Non English NLP Systems in Low-Resource Scenarios *IEEEACM Trans. Audio Speech Lang. Process.*, vol. 32, p. 1061–1074, doi:10.1109/TASLP.2023.3267618.

[23] Shukla, M., Kumar, A. (2023). An Experimental Analysis of Deep Neural Network Based Classifiers for Sentiment Analysis Task, *IEEE Access*, vol. 11, p. 36929–36944, doi:10.1109/ACCESS.2023.3266640.

[24] Jang, B., Kim, I., Kim, J. W. (2019). Word2vec convolutional neural networks for classification of news articles and tweets, *PLOS ONE*, vol. 14, no. 8, p. e0220976, Aug. doi:10.1371/journal.pone.0220976.

[25] Iatsenko, D. V. (2023). Texts Classification with the usage of Neural Network based on the Word2vec's Words Representation, *Int. J. Soft Comput.*, 14 (2) p. 1–13, May doi:10.5121/ijsc.2023.14201.

[26] Choi, J., Lee, S. W. (2020). Improving FastText with inverse document frequency of subwords, *Pattern Recognit. Lett.*, vol. 133, p. 165–172, May doi:10.1016/j.patrec.2020.03.003.

[27] Dundar, E. B., Alpaydýn, E. Learning Word Representations with Deep Neural Networks for Turkish, in *2019 27th Signal Processing and Communications Applications Conference (SIU)*, Sivas, Turkey: *IEEE,* Apr. 2019, p. 1–4. doi:10.1109/SIU.2019.8806491.

[28] Tripathi., Agarwal, R. (2025). Bias Mitigation in NLP: Automated Detection and Correction, *Int. J. Res. Mod. Eng. Emerg. Technol.*, 13 (5) p. 45–60, May. doi:10.63345/ijrmeet.org.v13.i5.130503.

[29] Zhou, H., Inkpen, D., Kantarci, B. (2024). Evaluating and Mitigating Gender Bias in Generative Large Language Models, *Int. J. Comput. Commun. CONTROL*, 19 (6) Nov. doi:10.15837/ijccc.2024.6.6853.

[30] Sattigeri, P., Hoffman, S. C., Chenthamarakshan, V., Varshney, K. R. (2019). Fairness GAN: Generating datasets with fairness properties using a generative adversarial network, *IBM J. Res. Dev.*, 63 (4/5), p. 3:1-3:9, Jul. doi:10.1147/jrd.2019.2945519.

[31] Zhang, B. H., Lemoine, B., Mitchell, M. (2018). Mitigating Unwanted Biases with Adversarial Learning, *In:* Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, New Orleans LA USA: ACM, Dec. p. 335–340. doi:10.1145/3278721.3278779.

[32] Lohia, P. K., Ramamurthy, Natesan, K., Bhide, Saha, M. D., Varshney, K. R., Puri, R. (2019). Bias MitigationPost-processing for Individual and Group Fairness, *In*: ICASSP 2019 - 2019 IEEE International

Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, United Kingdom: *IEEE,* May, p. 2847–2851. doi:10.1109/icassp.2019.8682620.

[33] Sze Khoo L., *et al.*, (2023). Exploring and Repairing Gender Fairness Violations in Word Embedding-based Sentiment Analysis Model through Adversarial Patches, in *2023 IEEE International Conference on Software Analysis, Evolution and Reengineering (SANER)*, Taipa, Macao: *IEEE,* Mar., p. 651–662. doi:10.1109/saner56733.2023.00066.

[34] Asyrofi, M. H., Yang, Z., Yusuf, I. N. B., Kang, H. J., Thung, F., Lo, D. (2021). BiasFinder: Metamorphic Test Generation to Uncover Bias for Sentiment Analysis Systems, *IEEE Trans. Softw. Eng.*, p. 1–1, doi:10.1109/tse.2021.3136169.

[35] Almuzaini, A. A., Singh, V. K. (2020). Balancing Fairness and Accuracy in Sentiment Detection using Multiple Black Box Models, *In:* Proceedings of the 2nd International Workshop on Fairness, Accountability, Transparency and Ethics in Multimedia, Seattle WA USA: *ACM,* Oct., p. 13–19. doi:10.1145/3422 841.3423536.

[36] Chen Xinying, V., Hooker, J. N. (2023). A guide to formulating fairness in an optimization model, *Ann. Oper. Res.*, 326 (1) p. 581–619, Jul. doi:10.1007/s10479-023-05264-y.

[37] Dogra, V., *et al.*, (2022). A Complete Process of Text Classification System Using State of the Art NLP Models, *Comput. Intell. Neurosci.*, vol. 2022, p. 1–26, Jun., doi:10.1155/2022/1883698.

[38] Xu, S., Zhang, C., Hong, D. (2022). BERT based NLP techniques for classification and severity modeling in basic warranty data study, *Insur. Math. Econ.*, vol. 107, p. 57–67, Nov. doi:10.1016/j.insmatheco.2022.07.013.

[39] Gosho, M., Ohigashi, T., Nagashima, K., Ito, Y., Maruo, K. (2023). Bias in Odds Ratios From Logistic Regression Methods With Sparse Data Sets, *J. Epidemiol.*, 33 (6) p. 265–275, Jun. doi:10.2188/jea.JE20210089.

[40] Tang, Y., Martin, R. (2024). Empirical Bayes inference in sparse high dimensional generalized linear models, *Electron. J. Stat.*, 18 (2) Jan. doi:10.1214/24-EJS2274.

[41] Du, K. L., Jiang, B., Lu, J., Hua, J., Swamy, M. N. S. (2024). Exploring Kernel Machines and Support Vector Machines: *Principles, Techniques, and Future Directions, Mathematics,* 12 (24) p. 3935, Dec., doi:10. 3390/math12243935.

[42] Perros, H. G. (2021). Support Vector Machines, in *An Introduction to IoT Analytics*, 1st ed., First edition. Boca Raton/:CRC Press, 2021: *Chapman and Hall/CRC*, p. 279–302. doi:10.1201/9781003139041-11.

[43] Schonlau, M., Zou, R. Y. (2020). The random forest algorithm for statistical learning, *Stata J. Promot. Commun. Stat. Stata*, 20 (1) p. 3–29, Mar., doi:10.1177/1536867X20909688.

[44] Hong, S., Lynn, H. S. (2020). Accuracy of random forest based imputation of missing data in the presence of non-normality, non-linearity, and interaction, *BMC Med. Res. Methodol.*, 20 (1) p. 199, Dec. doi: 10.1186/s12874-020-01080-1.