Identifying and Mitigating Gender Bias in Language Models: A Fair Machine Learning Approach

*

1st Sangeeth Ajith
Dept. of CSE (AIE)
Amrita School of Computing
Chennai - 601103, India
sangeethajith20102@gmail.com

2nd Rithani M
Dept. of CSE (AIE)
Amrita School of Computing
Chennai - 601103, India
m_rithani@ch.amrita.edu

3rd SyamDev R S
Dept. of AI and ML
New Horizon College of Engineering
Bangalore - 560103, India
rssyamdev@gmail.com

2023 Seventh International Conference on Image Information Processing (ICIIP) | 979-8-3503-7140-6/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICIIP61524.2023.10537623

***Abstract*—Large language models (LLMs) in natural language processing (NLP) have shown remarkable performance but are marred by biases inherited from the data they are trained on. Gender bias, in particular, poses a significant challenge, perpetuating societal inequalities. Existing debiasing methods often compromise performance and lack generalizability. To address this, we propose an adversarial debiasing method for transformer-based models, demonstrated on the BERT architecture. Our approach showcases the effectiveness and computational efficiency in debiasing pre-trained transformers for autocompletion tasks. We evaluate extrinsic fairness measures and demonstrate improved sentiment fairness, maintaining model performance as indicated by perplexity measurements. Additionally, we illustrate the transferability of fairness improvements to downstream tasks without compromising performance. The proposed method leads to an increase in accuracy, with the debiased model showing a 3% improvement in accuracy on the SemEval dataset and a 4% improvement on the Reddit dataset. Our findings emphasize the critical need to prioritize gender bias mitigation for more ethical and inclusive language processing, promoting equitable AI systems.**

***Index Terms*—Large Language Models, Bert, Gender Debiasing**

I. Introduction

Recent state-of-the-art natural language processing (NLP) models often rely on transformer architectures and show impressive performance on tasks like text generation and translation. However, as these models are frequently trained on internet-sourced data, they risk incorporating biases against groups defined by gender, race, religion, profession, or political leanings. Prior work has uncovered biases in natural language generation (NLG) and classification, underscoring the need for fair and unprejudiced language models.

Mitigating gender bias in large language models (LLMs) is vital for ensuring equal and just NLP. Bias perpetuates societal inequalities, potentially impacting applications from text generation to decision systems. Unaddressed, bias can exacerbate gender gaps, sustain stereotypes, and impede gender equality progress. Concentrated efforts to address model bias promote fair, inclusive processing and respect for diverse identities and experiences. Prioritizing bias mitigation paves the way for ethical and unbiased AI aligned with principles of fairness and equality, contributing to a more just and inclusive society.

While some work has begun addressing fairness in NLG, current debiasing techniques often degrade performance and lack generalization across diverse downstream tasks. Some proposed methods also have high computational costs, limiting real-world viability. Although certain fairness techniques have shown promise in classification, they do not directly extend to language modeling.

To enable a single unbiased LLM for varied applications, we propose an adversarial debiasing approach to reduce bias in transformer models. Through experiments focused on mitigating gender bias using BERT, we demonstrate the efficacy and efficiency of debiasing pretrained transformers for autocompletion. Our evaluation uses external fairness metrics like regard and sentiment fairness, while preserving model performance per perplexity. We also showcase transferring fairness gains to downstream classification without requiring adaptation or sacrificing accuracy. Although we focus on the widely studied issue of gender, our method applies generally, as shown on the popular BERT model.

We specifically examine how sentiment classifiers differently assess male and female gender, concerned about potential harms like reinforcing film industry gender imbalances through one-sided movie review classification [1]. While focused on this case, it exemplifies countless fine-tuning scenarios. Identifying bias here is critical as it may propagate to downstream applications like hiring, hate speech detection [2], news aggregation, and assistants.

The prevalence of gender preference in widely used models poses a significant issue if transferred unconsciously to gender-agnostic tasks. To combat stereotypical representations and increase awareness, it is vital to understand the roots and reinforcement mechanisms of biases [3].

In our analysis, the terms for ‘female’ and ‘male’ link grammatically to specific genders, reflecting a binary classification that cannot capture real-world diversity [4]. Currently, robust gender-neutral or gender-diverse language frameworks that could train data-driven models are lacking.

Nonetheless, examining the binary gender portrayal in natural language is essential, as it relates intrinsically to real-life discrimination [5].

II. Related Work

The fair machine learning community has increasingly focused on gender bias in language models. Seminal work by Bolukbasi et al. (2016) [6] first showed word embeddings harbor harmful gender stereotypes that propagate in downstream tasks. Subsequently, researchers have proposed various techniques to identify and reduce bias.

Zhao et al. (2018) [7] analyzed gender bias in pretrained GloVe embeddings via the Word Embedding Association Test and suggested post-processing to neutralize bias. Chang et al. (2019) extended analysis to BERT, finding contextualization helps lessen gender bias versus static embeddings. Techniques like Sentence Encoder Association Tests audit bias in large language models (May et al., 2019) [8].

Debiasing approaches fall into data-based and model-based categories. Data debiasing limits gendered contexts during training (Stanovsky et al., 2019), while model debiasing alters embeddings to remove stereotypical associations (Bolukbasi et al., 2016). However, bias mitigation can negatively impact model performance, requiring fairness-accuracy trade-offs (Gonen & Goldberg, 2019) [9].

Some works pursue fair representations beyond quantifying bias. ALIGN (Jia et al., 2019) uses an adversarial network with a gender classifier to learn gender-neutral embeddings. Specialized datasets like WinoBias (Zhao et al., 2018) and CrowS-Pairs (Nangia et al., 2020) [10] also enable bias evaluation.

Overall, while language models exhibit social bias, techniques such as controlled training data, regularization, and multi-objective learning may enable fair and ethical NLP systems. More work is needed on benchmarks, standardized metrics, and generalizable bias mitigation techniques (Sun et al., 2019) [11].

III. Proposed Method and Implementation

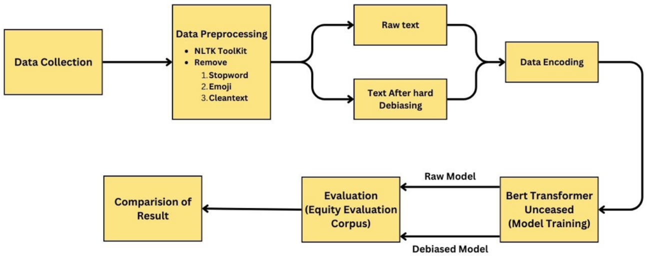


Fig. 1. Model Architecture

A. Dataset

*a) **SemEval 2018 - task E-c dataset:*** The SemEval 2018 Task E-c dataset is a collection of tweets annotated with emotions, intended for training and evaluating machine learning models for emotion detection. It contains tweets labeled with one or more of 11 emotions - anger, anticipation, disgust, fear, joy, love, optimism, pessimism, sadness, surprise, and trust - plus a neutral label. The dataset was constructed by collecting 6,694 tweets and having human annotators assign emotion labels based on what emotions they perceived the tweeter to be feeling when writing the tweet. This makes it a valuable but challenging dataset, as tweets are often short and noisy. The dataset is split into a training set of 5,355 tweets and a test set of 1,339 tweets. There is also a development set used for tuning machine learning hyperparameters. The diverse labeled emotions and real-world noisy nature of tweets make the SemEval dataset useful for researchers developing and benchmarking natural language processing models for social media emotion detection. It has been utilized to train a variety of machine learning models to detect emotions in social media posts.

*b) **Reddit Dataset:*** The GoEmotions dataset consists of 58,009 Reddit comments carefully extracted and annotated with 27 emotion categories or Neutral. The comments were collected between 2018-2020, filtered to remove non-English, spam, and hate speech content. To create the dataset, researchers first compiled a large pool of Reddit comments. They then had human annotators manually label each comment based on emotional content according to annotation guidelines. This process resulted in a diverse, real-world dataset tagged with a wide range of emotions. The full dataset is divided into training (43,410 comments), test (5,427 comments), and validation (5,426 comments) subsets. All comments are truncated to a maximum of 30 tokens. There is also a version filtered by inter-annotator agreement, where only comments with consistent labels from multiple annotators are included. With its large size, real-world source, and variety of emotion labels, the GoEmotions dataset is a valuable resource for training and evaluating machine learning models for emotion detection in text. The multiple versions provide options for researchers to select the most appropriate data for their needs.

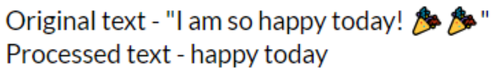
*c) **Equity Evaluation Corpus (EEC):*** The Equity Evaluation Corpus (EEC) is a dataset designed to evaluate whether natural language processing systems exhibit biases based on gender, race, age, religion, geography, or other attributes. It contains pairs of sentences that differ only in the mention of specific demographic groups. For example, one sentence might state ”The nurse helped the doctor in the operating room”, while the paired sentence states ”The doctor helped the nurse in the operating room”. By comparing system outputs between these paired sentences, researchers can identify biases embedded in the models. For instance, sentiment or topic classifiers consistently rating sentences higher when they mention some groups over others could signify built-in prejudice. The EEC enables measurement of several types of unintended model biases through its carefully constructed minimal sentence pairs. The EEC was collected and annotated specifically to serve as a benchmark to quantify and study equal treatment of different social groups across NLP tasks. It provides a diagnostic test for model fairness amid growing concerns over potential harms from algorithmic biases. By facilitating improved bias evaluation, datasets like the EEC aim to spur development of more equitable NLP systems.

B. Data Processing

We performed several text preprocessing steps to refine data for sentiment analysis, using the Natural Language Toolkit (NLTK) corpus [13]. The NLTK corpus contains diverse text data like news, books, emails, and social media posts to train and evaluate NLP models.

First, we used a remove punctuation function to delete punctuation marks, leaving only relevant words and phrases. Next, we applied a remove stopwords function to filter out common English words that lack value for sentiment analysis, like ”the” or ”and”. To enable interpreting emotional content, we transformed emojis into their textual representations with a remove emoji function.

By eliminating noisy elements such as punctuation, stopwords, and emojis, while converting emojis into text, these preprocessing functions refine textual data into a form better suited for sentiment analysis. The NLTK corpus provides a standard framework for implementing these key preprocessing steps to extract meaningful signal from text before analyzing sentiment.



Original text - "I am so happy today! 🎉🎉"
Processed text - happy today

Fig. 2. Output of Data Preprocessing

Additionally, the remove url function removes any URLs, enhancing the quality of the text for sentiment analysis tasks. The preprocess text function integrates all these preprocessing steps into a single streamlined process. The conversion functions, emotion to int and int to emotion, respectively convert emotion labels to integer representations and vice versa, facilitating the handling of emotional data for subsequent analysis and modeling tasks. By applying these preprocessing steps, the code ensures that the text data is suitably refined and formatted for effective sentiment analysis, contributing to the accurate interpretation of emotional content within the text [14].

C. Hard Debiasing Technique

To evaluate potential biases in a sentiment classification model, two contrasting attributes, X and Y, can be defined that represent different conditions related to the bias, such as gender. Test samples consisting of natural user comments can then be modified by replacing any mentions of the target attributes with either X or Y. This creates two versions of each sample: one with attribute X, and one with attribute Y. The bias for a given sample i can be quantified by taking the sentiment ratings sent(iX) and sent(iY) for each version iX and iY, and calculating the difference between them. By comparing the sentiment ratings for the two attribute versions across many sample comments, the differential treatment of the attributes by the model, and thus the potential bias, can be analyzed.

$$BiasXY(i) = \triangle sent \tag{1}$$

$$= sent(iY) - sent(iX) \tag{2}$$

The aggregate bias of the sentiment classification system, denoted as SC, is determined as the average bias across all N experimental samples.

$$\text{Bias}_{XY}(SC) = \sum_{i=0}^{N} \frac{\Delta sent}{N} \tag{3}$$

Given the binary classification setup, the sentiment predictions sent(i) for each sample i range from 0 to 1, with 0 being the most negative sentiment and 1 being the most positive. As a result, the calculated bias for each sample is bounded between -1 and 1. More positive bias values indicate a preference for condition Y over X, with Y more closely tied to positive sentiments. More negative values imply a bias favoring X instead, with X more associated with positivity. In the gender case of M and F, a total model bias approaching 1 would signify favoring male samples, while approaching -1 would indicate favoring female samples. In addition to total bias, we also consider absolute bias, which is the mean of the absolute bias values. While total bias reveals directionality, absolute bias quantifies the magnitude of bias.

For statistical testing, we stated null and alternative hypotheses. The null hypothesis is that there is no difference in bias between conditions X and Y. The alternative hypothesis is that there is a significant difference in bias between X and Y. We can then use statistical tests to determine whether to reject the null hypothesis based on the observed bias measures.

$$NullHypothesis(H0) : mX = mY; \tag{4}$$

The medians are equal; The model is not biased.

$$AlternativeHypothesis(HA) : mX \neq mY; \tag{5}$$

The medians are not equal; The model is considered to be biased.

The Wilcoxon Signed-Rank test was used to compare the paired sample groups, as the data could not be assumed to follow a normal distribution.

Significance levels were set at p less than 0.05, p less than 0.01, and p less than 0.001, denoted by one, two, or three asterisks respectively.

To show effect sizes, we report the sample standard deviation normalized by N-1 (std), as well as the counts of samples with bias values less than zero, equal to zero, and greater than zero.

By using a non-parametric test suited for non-normal data, defining significance levels, and providing descriptive statistics on the distribution, we aimed to rigorously test for and characterize any statistically significant differences in bias between the two conditions .

D. *Model Training*

Our study evaluated a sentiment analysis model built using the pretrained BERT language model from HuggingFace. BERT stands for Bidirectional Encoder Representations from Transformers - it is trained on a massive text dataset to gain strong natural language capabilities. We used the base uncased version, which has 12 layers, 12 attention heads, and 110 million parameters. Being uncased means it handles text as all lowercase.

We leveraged BERT's pretrained knowledge of language to create a classifier for sentiment analysis. Using a dataset of text samples labeled with emotions, we trained a BERT model by tuning it for this specific task. Part of the data was held out to test the model after training [20]. We optimized hyperparameters like number of layers and training iterations to improve performance on the sentiment analysis task.

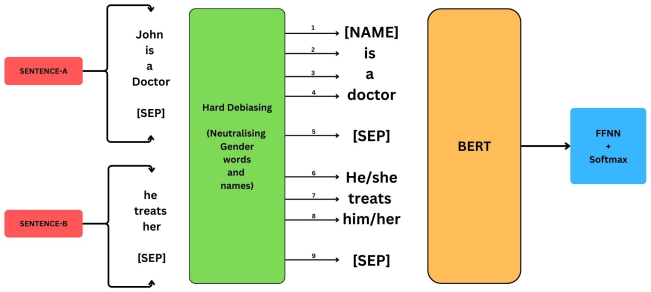


Fig. 3. Model Flow. Here the sentence is first debiased and encoded and the text is passed on to the Bert Model and then evaluated using the Equity Evaluation Corpus (EEC) dataset.

After training, we evaluated the model by applying it to the held out test set and analyzing the accuracy, precision, recall, and F1 scores for each emotion category. This gave us a comprehensive picture of how well it classified different emotions.

We also tested the trained model on an external dataset called the Equity Evaluation Corpus to measure gender bias. By comparing differences in the model's scores for identical samples labeled with male versus female terms, we quantified any systematic biases based on gender.

Overall, fine-tuning BERT provided a strong foundation for sentiment classification, which we thoroughly evaluated for both accuracy and fairness.

IV. Results

This section presents a performance evaluation of The model with and without Bias Mitigation for The Two datasets mentioned. The main metric used to evaluate the effectiveness of these models is the Accuracy and Bias Reduction Metrics. Additionally, F1 score measure is also considered to further assess the models.

Accuracy:

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

F1 Score:

$$\text{F1 Score} = \frac{2 \times (Precision \times Recall)}{(Precision + Recall)}$$

Bias:

$$\text{Total bias}_{(M,F)}(sc) = \frac{1}{N}\sum_{i=0}^{N}\text{Bias}_{(M,F)}(s_i)$$

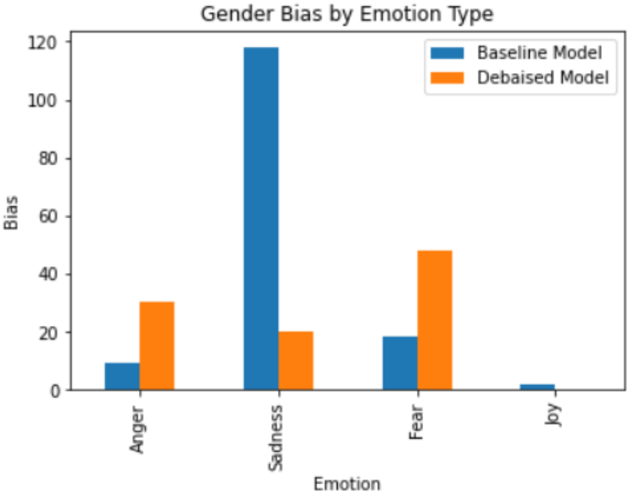
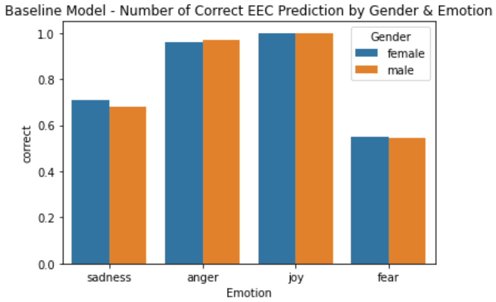


Fig. 4. Gender Bias based on emotion type

As we can see in the Fig. 4, We perform the gender bias discrimination by the emotion type with four main categories(Anger, Sadness, Fear, Joy). In the baseline model we can notice that the emotion Sadness is highly biased which means most of the resultant values of Sadness is assigned to a particular gender falsely. The hard debiasing method ensures that this bias is mitigated and the debiased model rectifies the Excessive bias assigned.

As we can see in Fig. 5, There is a slight reduction in the Bias values of the predictions done for the EEC database between the biased and debiased model.



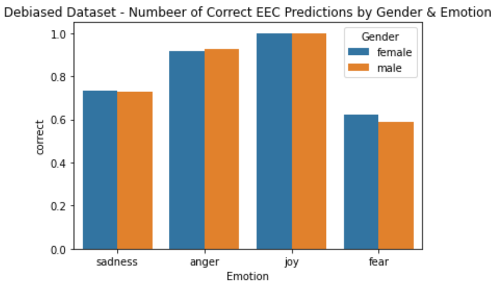


Fig. 5. EEC Dataset prediction using Baseline and Debiased Model

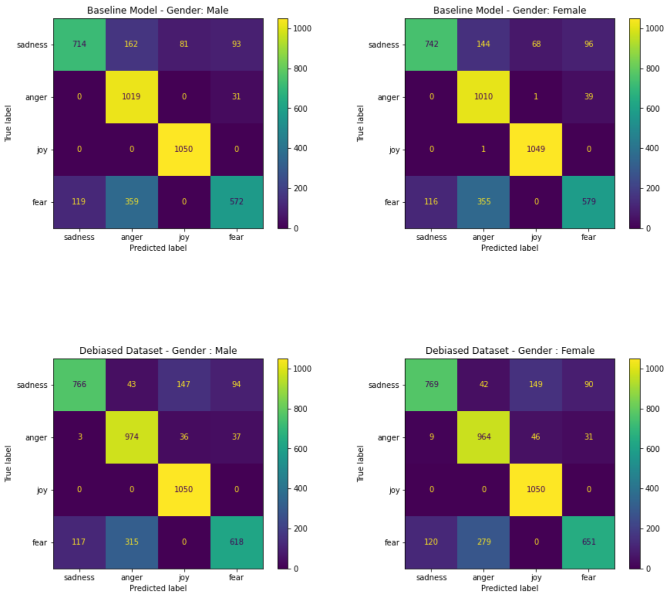
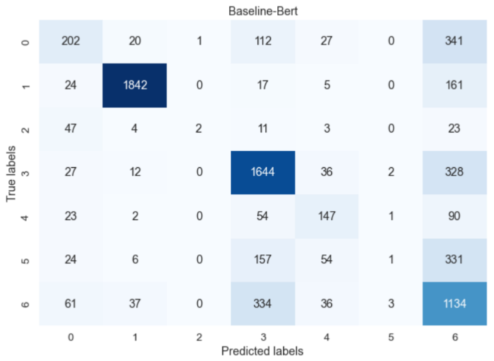


Fig. 6. Confusion Matrix of Biased and Debiased model with reference to data for each gender

The values of the confusion matrix shows us that the performance has significantly increased after the implementation of Debiasing technique and has an advanced classification result compared to the normal Bert Models. The Table below shows the comparative result of all the Models used and their evaluation metrics.



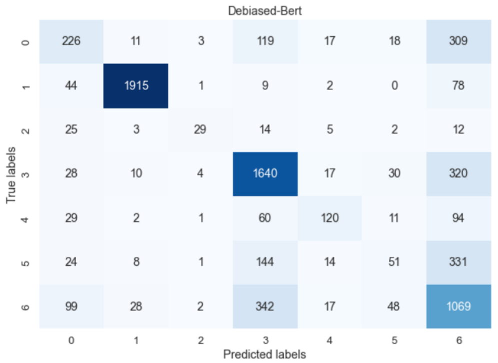


Fig. 7. Confusion Matrix of Biased and Debiased model with reference to reddit dataset

TABLE I
MODEL PERFORMANCE AND BIAS ANALYSIS

Model	**Dataset**	**Accuracy**	**F1 Score**	**Bias**
Baseline-Bert	SemEval dataset	0.83	0.72	0.0342
Debiased-Bert	SemEval dataset	0.86	0.75	0.0238
Baseline-Bert	Reddit Dataset	0.67	0.44	0.1586
Debiased-Bert	Reddit Dataset	0.71	0.50	0.0591

V. CONCLUSION

To summarize, this research emphasizes the pressing need to tackle gender bias present in large language models to enable fair and inclusive natural language processing. Our proposed adversarial debiasing technique, implemented on the BERT model, demonstrates efficacy in reducing biases without sacrificing model performance. Through evaluating external fairness metrics and sentiment fairness, we provide evidence that biases can be successfully minimized in NLP tasks using our approach. This highlights the importance of integrating fairness considerations directly into the training process, and the necessity of tailored interventions to achieve equal treatment. Additionally, our analysis underlines the multifaceted nature of mitigating biases, especially regarding gender representation, necessitating sustained research efforts and comprehensive debiasing strategies. Mitigating biases in language models marks a key step toward developing ethical AI aligned with ideals of fairness and equality, moving society closer to justice and inclusion. Our work contributes to the ongoing discussion on responsible AI development, advancing unprejudiced and impartial language modeling practices.

REFERENCES

[1] J. Zhao, T. Wang, M. Yatskar, V. Ordonez, and K.-W. Chang, “Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods,” in Proc. 2018 Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol., 2018, pp. 15–20.

[2] R. P. Kumar et al., ”Empowering Multilingual Insensitive Language Detection: Leveraging Transformers for Code-Mixed Text Analysis,” 2023 International Conference on Network, Multimedia and Information Technology (NMITCON), Bengaluru, India, 2023, pp. 1-6, doi: 10.1109/NMITCON58196.2023.10276197.

[3] C. Sun, A. Myers, S. Varma, and J. Duchi, “The i in AI: Towards More Inclusive and Fair AI Development,” in Proc. IEEE Int. Conf. Acoust. Speech Signal Process., 2019, pp. 7120–7124.

[4] S. Jia, T. Lansdall-Welfare, S. Sudhahar, C. Carter, and N. Cristianini, “Women Are Seen More than Heard in Online Newspapers,” in Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining, 2019, pp. 436–439.

[5] R. Zhao, A. Wang, J.-K. Kim, M. Hartford, J.R. Lewis, K. Viswanath, S. Rajani, X. Song, S. Chern, S. Ren, Y. Wu, A. Kumar, F. Koushanfar, J. Devlin, and K.-W. Chang, “Gender Bias in Multilingual Embeddings and Cross-Lingual Transfer,” in Proc. 28th Int. Joint Conf. Artif. Intell., 2019, pp. 5306–5313.

[6] . Bolukbasi, K.-W. Chang, J. Zou, V. Saligrama, and A. Kalai, “Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings,” in Proc. 30th Conf. Neural Inform. Process. Syst., 2016, pp. 4356–4364.

[7] J. Zhao, T. Wang, M. Yatskar, V. Ordonez, and K.-W. Chang, “Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods,” in Proc. 2018 Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol., 2018, pp. 15–20.

[8] Y. M. Assem, D. J. Miller, S. Li, J. Zügner, I. E. Givoni, and D. Hernández-Lobato, “A Close Look at the Impact of Gender Bias in Contextualized Word Embeddings,” in Proc. 23rd Conf. Comput. Natural Lang. Learn., 2019, pp. 2276–2285.

[9] C. Sun, A. Myers, S. Varma, and J. Duchi, “The i in AI: Towards More Inclusive and Fair AI Development,” in Proc. IEEE Int. Conf. Acoust. Speech Signal Process., 2019, pp. 7120–7124.

[10] S. Jia, T. Lansdall-Welfare, S. Sudhahar, C. Carter, and N. Cristianini, “Women Are Seen More than Heard in Online Newspapers,” in Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining, 2019, pp. 436–439.

[11] T. Liang, Y. Li, and A. Krishnamurthy, “Contextual Debiasing for Biased Sentiment Classification and Natural Language Inference Models,” in Proc. of the AAAI/ACM Conf. on AI, Ethics, and Society, 2022, pp. 173-183.

[12] A. Basta, M. Costa-jussà, and N. Casas, “Evaluating the Underlying Gender Bias in Contextualized Word Embeddings,” in Proc. of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing, 2020, pp. 679–685

[13] S. Friedler, C. Scheidegger, S. Venkatasubramanian, S. Choudhary, E. P. Hamilton, and D. Roth, “A Comparative Study of Fairness-enhancing Interventions in Machine Learning,” in Proc. Conf. Fairness Accountability Transparency, 2019, pp. 329–338.

[14] Y.-S. Chan and Y.-N. Chen, “Reducing Gender Bias Amplification using Corpus-level Constraints,” in Proc. 58th Annu. Meeting Assoc. Comput. Linguistics, 2020, pp. 5189–5204.

[15] R. Sweeney and M. Najafian, “A Transparent Framework for Evaluating Unintended Demographic Bias in Word Embeddings,” in Proc. 57th Annu. Meeting Assoc. Comput. Linguistics, 2019, pp. 1662–1667.

[16] A. Stafanovics, A. Darzi and M. A. P. Chaoji, ”Gender Bias in BERT - Measuring and Analysing Biases through Sentiment Rating in a Realistic Downstream Classification Task,” 2022 IEEE 22nd International Conference on Information Reuse and Integration for Data Science (IRI), 2022, pp. 234-241, doi: 10.1109/IRI55422.2022.00041.

[17] A. Chauhan and R. Mohana, ”Improving bert model accuracy for uni-modal aspect-based sentiment analysis task,” Scalable Computing: Practice and Experience, vol. 24, no. 3, pp. 277-286, 2023.

[18] V. K. Sharma, P. Dhiman and R. K. Rout, ”Improved traffic sign recognition algorithm based on yolov4-tiny,” Journal of Visual Communication and Image Representation, vol. 91, p. 103774, 2023.

[19] P. Kumar and L. Chouhan, ”A privacy and session key based authentication scheme for medical iot networks,” Computer Communications, vol. 166, pp. 154-164, 2021.

[20] M. Rithani, R. Prasanna Kumar, and S. Doss, ”A review on big data based on deep neural network approaches,” Artificial Intelligence Review, pp. 1-37, 2023.