Unveiling and Mitigating Occupational Gender Stereotypes by Advancing Sentiment Analysis Models for Fairer Outcomes

Mariam Abdullah, Kina Huang, Duyi Liu, Fanqi Cheng
Center for Data Science and ASH, New York University
{ma3259, yh5266, dl5334, fc2456}@nyu.edu

1 Introduction

Social Role Theory[1] highlights how societal gender roles create occupational stereotypes, which persist in algorithms, such as Amazon's AI hiring tool favoring men due to male-dominant training data[2]. This study closely examines this problem, replicating and extending Bhaskaran and Bhallamudi's[3] study to evaluate gender bias in advanced models (e.g., BERT, ALBERT, RoBERTa, GPT-4o-mini). We also test debiasing methods like GN-GloVe embeddings and adversarial training efficacy on gender bias.

2 Related Work

Bhatia and Bhatia's Changes in Gender Stereotypes Over Time[4] finds that biases related to feminine traits have weakened, while masculine biases remain stable. Similarly, Good Secretaries, Bad Truck Drivers? by Bhaskaran and Bhallamudi[3] demonstrates gender bias in sentiment models, showing that male-associated terms in professions like "pilot" or "truck driver" receive more positive sentiment than female-associated terms in the same context.

3 Approach and Methods

We evaluated both traditional and contextual models for gender bias via structured experiments, focusing on performance metrics and debiasing methods to reduce stereotypes in predictions.

Baseline models consist of a Bag-of-Words approach using TF-IDF combined with logistic regression, in addition to a BiLSTM architecture with GloVe embeddings. Our primary contributions then focus on transformer-based models, exploring retention across varying scales using RoBERTa, ALBERT, and GPT-4o-mini to dimension bias in small to large-scale models.

First, all models are trained on the *Stanford Sentiment Treebank 2 (SST-2)* dataset[5], with evaluations comparing predicted positive sentiment probabilities for sentences with male versus female nouns.

Debiasing methods include GN-GloVe embeddings[6] to reduce bias via gender-neutral word representations, and a Generative Adversarial Networks (GAN)[7] with a discriminator to remove gender bias from model representations.

We specifically used GN-GloVe 1b-vectors300-0.8-0.8 embeddings, which were trained on 1 billion tokens as described in the paper *Learning Gender-Neutral Word Embeddings*[8] . These embeddings aim to isolate gender information in certain dimensions while preserving other functionalities of the model.

For the GAN setup, we implemented a discriminator on BERT's CLS token representations with fully connected layers and BCE-based adversarial loss. The training combines classification loss from BERT's logits and adversarial loss targeting neutral embeddings, promoting debiasing while preserving task performance.

4 Experiments

4.1 Data

Good Secretaries, Bad Truck Drivers Dataset: This dataset contains the 800 gender-specific sentences structured as "[NOUN] is a profession", for example, "This woman is a nurse." [3]. This dataset provides a balanced, controlled setup to replicate the original sentiment analysis portion of the study.

SST-2 (Stanford Sentiment Treebank 2)[5]: While not specifically focused on gender bias, SST-2 is included as a baseline dataset to establish general sentiment analysis performance across models. Training models on SST-2 before fine-tuning on gender-specific datasets helps preserve core sentiment classification capabilities.

4.2 Evaluation method

First, we replicated the study Good Secretaries, Bad Truck Drivers? Occupational Gender Stereotypes in Sentiment Analysis[3] to assess gender bias in sentiment predictions.

Six models (Bag-of-Words + Logistic Regression, BiLSTM, BERT, ALBERT, GPT-4o-mini, and ROBERTa) were trained on the SST-2 dataset, which includes movie review phrases with binary sentiment labels (positive/negative). Each model was tested on a custom dataset to measure differences in mean positive sentiment predictions between sentences containing male versus female nouns, helping identify stereotypical biases in sentiment output.

The evaluation methodology followed the approach of Kiritchenko and Mohammad[9], using paired male-female sentence pairs and applying a paired t-test with a Bonferroni correction to control for multiple comparisons. We set the significance threshold to 0.01, adjusting it to 0.01/3 for multiple hypotheses.

4.3 Experimental details

BoW + LogReg: TF-IDF representations with logistic regression were used as a baseline, trained on SST-2 with L2 regularization and liblinear solver, ensuring convergence in 1000 iterations.

BiLSTM: A BiLSTM model with 100-dimensional GloVe embeddings was used to capture contextual word representations. It used 128 LSTM units, a 0.25 dropout rate, and an Adam optimizer with a learning rate of 0.001, trained for 3 epochs with a batch size of 32.

BERT[10]: We fine-tuned BERT-Base (uncased) for binary sentiment classification using BertForSequenceClassification. Training settings included a batch size of 8, AdamW optimizer, 3 epochs, 500 warm-up steps, and 0.01 weight decay for regularization.

ALBERT[11]: ALBERT-Base v2 was trained on SST-2 for binary classification with the following settings: a batch size of 8, a learning rate of 2e-5, weight decay of 0.01, and 3 training epochs. Predictions were generated for the evaluation set and custom control sentences.

RoBERTa[12]: RoBERTa-Base is a variant of BERT optimized for robust training, and it was fine-tuned for binary sentiment classification with the following settings: a batch size of 8, a learning rate of 2e-4, 3 epochs with 500 warming-up steps, and 0.1 weight decay.

GPT-4o-mini[13] : We fine-tuned GPT-4o-mini for sentiment classification using OpenAI's API. Training settings included 3 epochs, cl100k_base tokenizer with 4096 token limit, zero temperature for inference, and a system prompt for sentiment analysis rules.

4.4 Results

4.4.1 Baseline Models

Takeaway: *Traditional models (M.1 and M.2) show lower accuracies (0.827, 0.829) and significant gender bias (0.035, 0.101).*

Model	Development Accuracy	Female - Male
M.1 (BoW+LogReg)	0.827	0.035**
M.2 (BiLSTM)	0.829	0.101**
M.3 (BERT)	0.931	-0.013
M.4 (ALBERT)	0.874	-0.041
M.5 (RoBERTa)	0.905	-0.042
M.6 (GPT-4o-mini)	0.969	-0.011

** denotes statistical significance with $p < 0.01$.

Table 1: Development accuracy and gender difference (Female - Male) for different models.

Main experimental results are summarized in Table 1, evaluated using development accuracy and the Female-Male bias score, with significance assessed at $p < 0.01$. H_0 assumes identical mean predicted positive probabilities for female and male sentences.

The BoW + Logistic Regression model (M.1) achieved a development accuracy of 0.827 and a statistically significant Female-Male bias score of 0.035, reflecting higher positive sentiment predictions for female nouns. Similarly, the BiLSTM model (M.2) attained a slightly higher accuracy of 0.829 but exhibited a larger bias score of 0.101, likely due to its reliance on pre-trained embeddings encoding gender bias.

4.4.2 Transformer Models

Takeaway: *Transformer-based models outperform traditional methods, achieving higher accuracies (for e.g., GPT-4o-mini at 0.969) with minimal bias (-0.011), compared to significant biases in models like BiLSTM (0.101**). However, residual stereotypes exist.*

The GPT-4o-mini model (M.6)[13] surprisingly demonstrated the highest accuracy at 0.969 with a bias score of -0.011, reflecting both top-tier sentiment performance and minimal gender bias.

Meanwhile, BERT(M.3) achieved a development accuracy of 0.931 with a non-significant Female-Male bias score of -0.013, suggesting minimal gender bias while maintaining high classification performance. Similarly, ALBERT (M.4)[11] and RoBERTa (M.5)[12] achieved development accuracies of 0.874 and 0.905, respectively, with bias scores of -0.041 and -0.042, indicating strong bias mitigation while still preserving predictive capabilities.

These results demonstrate transformers' ability to outperform traditional models in both metrics while balancing bias and sentiment capability.

4.4.3 Occupational Stereotypes

Model	Top 3 professions	Bottom 3 professions
M.1 (BoW+LogReg)	Secretary, Teacher, Writer	Truck Dr., Fl. Att., (many)
M.2 (BiLSTM)	Dancer, Secretary, Scientist	Truck Dr., Gym Tr., Nurse
M.3 (BERT)	Professor, Bartender, Secretary	Teacher, **Pilot (-0.3)**, Truck Dr.
M.4 (ALBERT)	Bartender, Baker, **Scientist (0.13)**	**Pilot (-0.17)**, **Nurse (-0.14)**, Mechanic
M.5 (RoBERTa)	Secretary, (many)	Bartender, Doctor, Teacher
M.6 (GPT-4o-mini)	Soldier, (many)	Secretary, Clerk, Bartender

Table 2: Top 3 and bottom 3 professions per model, based on predicted positive class probability.

We now look at mean distributions of the positive class probability (between genders) for each profession, as shown in Table 2.

Traditional models like BoW + Logistic Regression (M.1) strongly aligned with gender roles, with gender-associated professions at the forefront—"Secretary," "Teacher," "Truck Driver," and "Flight Attendant."

Transformer-based models, meanwhile, demonstrated progressive reductions in bias. BERT (M.3) showed moderate improvements but assigned significantly lower positive probabilities for female pilots compared to male pilots. Similarly, ALBERT (M.4) shows negative bias for female nurses. The GPT-4o-mini model, while showing the least bias among the evaluated models, continued to reflect traces of occupational stereotypes—the bottom 3 professions including "Secretary" and "Clerk" .

These findings show the persistence of deeply ingrained societal biases within sentiment analysis models, even as more advanced architectures demonstrate incremental improvements.

4.4.4 Debiasing Techniques

Model	Development Accuracy	Female - Male
BiLSTM	0.829	0.101**
BiLSTM + GN-GloVe	0.846	0.046**
BERT	0.931	-0.013
BERT + Adversarial training	0.909	-0.013
ALBERT	0.874	-0.041
ALBERT + Adversarial training	0.875	-0.007

** denotes statistical significance with $p < 0.01$.

Table 3: Debiasing Comparison

Debiasing methods were applied to the BiLSTM and BERT models to evaluate their efficacy in mitigating gender bias, as shown in Table 3. For the BiLSTM model, replacing GloVe embeddings with GN-GloVe embeddings improved the development accuracy from 0.829 to 0.846 and reduced the Female-Male bias score from 0.101 to 0.046. Despite this reduction, the bias score of 0.046 remained statistically significant, indicating residual bias even after embedding-based mitigation.

In contrast, adversarial training on the BERT model preserved the Female-Male bias score at -0.013 (not statistically significant) with a slight drop in development accuracy from 0.931 to 0.909, indicating BERT's inherent low gender bias and no significant fairness improvements.

Adversarial training yielded further gains for Albert with accuracy increasing slightly to 0.915, and reducing the bias score to -0.007. This outcome demonstrates that ALBERT not only retains the bias-mitigation strengths of BERT, but also achieves modest improvements, suggesting that it may be more responsive to debiasing strategies.

5 Discussion

Our study highlights persistent gender bias in sentiment analysis models. Traditional models, such as BoW + Logistic Regression and BiLSTM, exhibited significant biases, aligning female terms with roles like "Secretary" and "Teacher" and male terms with "Truck Driver" and "Pilot," underscoring their reliance on biased training data. Transformer-based models, including BERT, ALBERT, RoBERTa, and GPT-4o-mini, showed substantial improvements, with non-significant Female-Male bias scores. However, residual biases persisted, particularly in professions like "Pilot" and "Nurse," reflecting underlying societal stereotypes in pre-trained embeddings.

Our debiasing methods, such as GN-GloVe embeddings and adversarial training, demonstrated varying success. GN-GloVe reduced bias in BiLSTM models, but left some residual biases, while adversarial training effectively mitigated bias in BERT without compromising accuracy. However, models like BERT, which initially showed minimal bias, saw limited improvements post-debiasing, highlighting the limitations of these methods.

Study limitations include the use of the SST-2 dataset, which lacks domain-specific contexts for occupational bias, and a focus on binary gender differences, neglecting intersectional and non-binary dimensions. Future research should incorporate domain-specific datasets, explore advanced debiasing techniques like counterfactual augmentation, and evaluate biases across multiple dimensions (e.g., race, age). Expanding to real-world applications, such as hiring systems, could provide a more comprehensive assessment of model fairness.

References

[1] Alice H. Eagly and Wendy Wood. Social Role Theory of Sex Differences. In *The Wiley Blackwell Encyclopedia of Gender and Sexuality Studies*, pages 1–3. John Wiley & Sons, Ltd, 2016.

[2] Jeffrey Dastin. Amazon scraps secret AI recruiting tool that showed bias against women, 2018.

[3] Jayadev Bhaskaran and Isha Bhallamudi. Good secretaries, bad truck drivers? occupational gender stereotypes in sentiment analysis, Jul 2019.

[4] Nazlı Bhatia and Sudeep Bhatia. Changes in gender stereotypes over time: A computational analysis. *Psychology of Women Quarterly*, 45(1):106–125, Mar 2021.

[5] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. In David Yarowsky, Timothy Baldwin, Anna Korhonen, Karen Livescu, and Steven Bethard, editors, *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA, October 2013. Association for Computational Linguistics.

[6] Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings, Jul 2016.

[7] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27, pages 2672–2680, 2014.

[8] Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. Learning gender-neutral word embeddings. *ArXiv*, 2018. Retrieved from `https://arxiv.org/abs/1809.01496`.

[9] Svetlana Kiritchenko and Saif M. Mohammad. Examining gender and race bias in two hundred sentiment analysis systems. In *Proceedings of the 7th Joint Conference on Lexical and Computational Semantics (*SEM)*, New Orleans, USA, 2018. Association for Computational Linguistics.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1805.04508*, 2018.

[11] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*, 2019.

[12] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[13] OpenAI. Gpt-4o-mini: A compact variant of gpt models, 2023.