Gender Bias in Job Predictions

Om Gor (669965504)
Amit Bhatt (660978630)
Namith Chandrashekar (673544255)
Mario Tabares (672655341)
Joshua John (655421958)

University of Illinois Chicago

Abstract

Automated tools are becoming increasingly prevalent in the recruitment process, where language models are used to analyze a professional's applications and biographies to help match the candidate to an occupation. While these systems improve efficiency for recruiters, they raise concerns about whether language models may inadvertently reproduce patterns of bias that already exist in society. In this project, we investigate whether gender bias exists in language models that analyze biographies. We compare a TF-IDF logistic regression baseline with a pretrained BERT classifier to examine if language models introduce additional bias beyond the frequency of words. The model performance is evaluated using accuracy and macro F1-score. The fairness evaluation is conducted using the Chi-squared test, the true positive rate gap between genders, and the accuracy gaps. Our results show that although the BERT classifier has a higher accuracy and improved fairness, it does not remove bias. The BERT model can amplify or retain gender bias in highly stereotyped occupations. We conduct multiple bias mitigation techniques, including reweighting, counter factual augmentation, and threshold adjustment. Counterfactual augmentation was found to be the best bias mitigator as it slightly reduced the gender disparities while retaining high accuracy.

1 Introduction

In today's age, automated decision making systems play an influential role in our lives, especially in the recruitment industry. Recruiters are becoming more dependent on automated tools to manage high volumes of applications, and the predictions generated by these systems have the potential to influence real job opportunities. Companies rely on automated systems to analyze an applicant's background and information to match qualified applicants to job roles. As these systems take on more responsibilities in the recruitment process, a question arises about whether language models may inadvertently reproduce patterns of bias that already exist in society. If a pre-trained language model contains unnoticed demographic biases, it may favor certain individuals over others even when qualifications are similar. The widespread use of professional networking platforms makes this a key source of information for recruiters to find potential candidates. Language models have been given the task of analyzing individuals' texts to infer job occupations to help recruiters promote positions. Our goal is to investigate whether demographic bias, specifically gender bias, exists in the model, measure how strongly demographic signals affect the occupations predicted, and apply bias mitigation techniques to improve the fairness and reliability of the model.

Preprint. Under review.

2 Methods

2.1 Dataset

To investigate whether gender bias exists in the classification, we used the Bias on Bios dataset found from HuggingFace which consists of short biographies scraped from online professional networking platforms. The dataset contains 257K rows, each row representing an individual's occupation biography, gender, and occupation. Each biography describes an individual's background, education, and work experience in the third person. The dataset contains 28 occupations, chosen because they appeared most frequently in the original set of biographies, encoded numerically. For exploratory data analysis, we mapped the gender occupation labels to their corresponding categories to better visualize the distributions. The proportions of each gender in the dataset were relatively similar, with male representing 53.9% while the female represented 46.1%. Looking closer at the frequencies of occupations present in the dataset, professors dominate it with more than twice the next category, which is physicians. Additionally, several occupations exhibited strong gender skew, one being nursing which had 11,000 females compared to 1000 male biographies. These imbalances are important because the models may learn to associate occupation with gender based on frequency, which motivates our fairness evaluation.

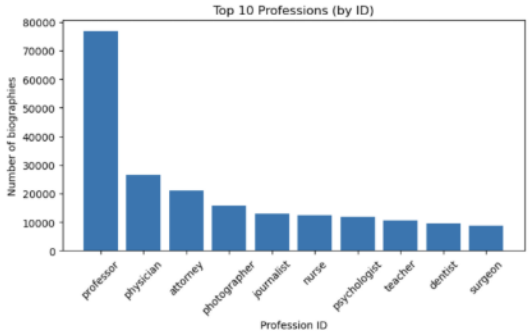

Figure 1: Top 10 Frequency of occupations in the dataset. It is seen that the professor occupation is much greater than the rest of the occupations.

Biography
She received her Ph.D. in Economics from the University of California, Irvine in 2013. Her research focuses on monetary economics, search theory, and international economics, with a particular emphasis on the effects of monetary policy on payment systems and credit markets.
Occupation Label
21
Gender Label
1

Table 1: Example row from the Bias in Bios dataset. The biography is the text given. The occupation is a numeric value ranging from 0-27. The gender is a numeric value, 0 representing male, and 2 representing female.

2.2 Data Pre-processing and Feature Engineering

First, we cleaned the biographies by converting all the text to lowercase, removing punctuation, numbers, and irrelevant symbols, and collapsing repeated whitespace. Before training the models, we applied several preprocessing methods to ensure the biographies were formatted correctly for classification. For the logistic regression baseline, which required the inputs to be numeric values, we used TF-IDF (Term Frequency-Inverse Document Frequency) as the feature representation. TF-IDF transforms each of the biographies based on the frequency each word appears in the document

relative to the rest of the biographies. This helps us identify which words are most informative for an occupation and have smaller weights for words with higher frequencies across biographies. Moving on, the BERT model expects the biographies to be in a tokenized format to be compatible with the architecture. Using the BERT tokenizer, each of the biographies was transformed into BERT-compatible input features. Compared to the transformed data for the baseline, the BERT tokenized text allows the model to understand the semantic and contextual meaning from the biographies. After pre-processing the data, we move on to building our classification models.

2.3 Baseline Model

To properly quantify gender bias in occupation classification, we established a baseline logistic regression model. The baseline model uses a traditional machine learning pipeline of TF-IDF features with a Logistic Regression model. This was used for its simplicity and ease of interpretability when it comes to understanding which words impacted the classification of occupation. As explained in the above section, we used TF-IDF as it measures the importance of the word relative to all the biographies. We trained a multinomial logistic regression model to classify biographies into one of the 28 occupations. To further increase model accuracy, we finetuned the logistic regression model by applying Ridge Regularization, tuned the inverse regularization strength, and the TF-IDF parameters using the validation set. The baseline is important in this analysis because it captures bias purely from word frequency. By comparing this to the BERT model, we can determine if the disparity is from word frequency imbalance or if the model further perpetuates gender bias.

2.4 Pre-trained Language Model

Next, we fine-tuned a pretrained BERT model to classify occupations. The BERT language model can understand the context of words based on the words surrounding them. If additional bias is identified in this model, it would imply that bias may not have originated from the dataset but directly from the pretrained language model. For fine-tuning, we applied the same train, validation, and test split used in the baseline model for a fair comparison. Each biography was tokenized using the BERT tokenizer, which converts it to tokens and constructs input IDs and attention masks needed for the model. We optimized the model for accuracy using AdamW optimization, running a total of three epochs. After building both the baseline and BERT model it is important to evaluate them using Bias Evaluation techniques.

2.5 Accuracy and Bias Evaluation Techniques

To evaluate the model's performance, we used evaluation methods such as accuracy, precision, recall, and macro F1-score. Macro F1-score was specifically used to evaluate all occupations fairly, giving all the classes equal weights. Beyond accuracy metrics, we evaluated model fairness using methods such as accuracy gap, true positive rate (TPR), TPR Gap, and Chi-squared Test. The accuracy gap is used to determine the difference between the accuracy of the male minus female gender, with a positive value indicating higher accuracy for males and a negative value indicating higher accuracy for females. The TPR gap measures the disparity between correctly predicted positive outcomes in each occupation across both genders. Lastly, the Chi-squared test evaluates whether the prediction outcomes are independent of gender. A value close to zero would indicate that gender and predictions are dependent, while a larger value closer to one would indicate that both are independent.

2.6 Bias Mitigation Strategy

After evaluating the accuracy and bias of the models, we implemented bias mitigation techniques such as pre-processing, in-processing, and post-processing, to discover the best way to reduce gender bias in occupation classification. For pre-processing, the male gender is weighted slightly higher than the female gender in the dataset, so we applied reweighting to balance the influence of the underrepresented gender class. Another technique used was Counterfactual Augmentation (CFA), which swaps the gendered terms in the biographies to create alternate biographies. These new biographies were used to test the models built to see if gender bias exists within them. For the post-processing method, we used threshold adjustment, which modifies the model's final prediction without retraining the model. It alters the lower confidence prediction when predicting to reduce the disparity between genders for the specific occupation without changing the model.

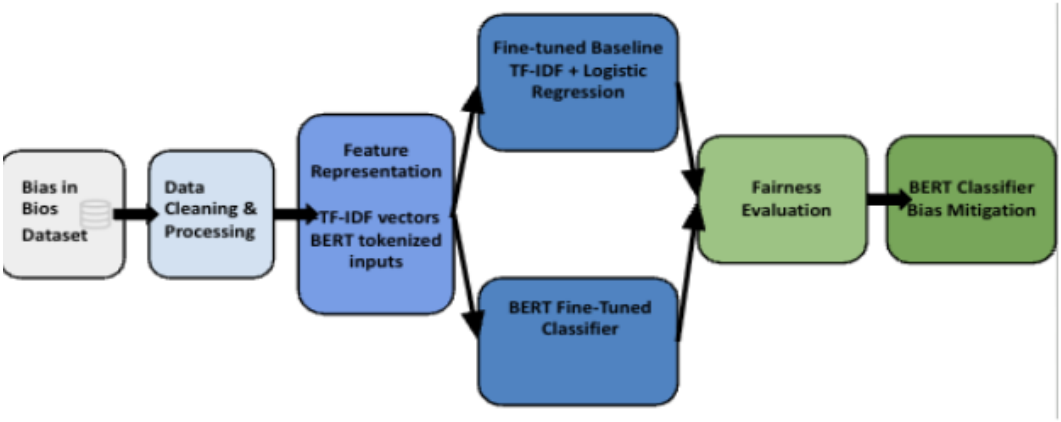

Figure 2: This is a visual depiction of the pipeline used in this project for Fairness Evaluation and Bias Mitigation.

3 Experimental Results

Figure 3 summarizes the accuracy results comparing the baseline logistic regression model to the BERT classifier. Each of the models also includes a male and female accuracy to see any disparities that may occur for gender classes. The BERT classifier achieves a higher accuracy and F1-score compared to the logistic regression model by around 3.8%.

Model	Accuracy	Macro F1 Score
Overall Logistic Regression	0.8220	0.7645
Logistic Regression (Male)	0.82	0.72
Logistic Regression (Female)	0.83	0.75
Overall BERT classifier	0.8599	0.8108
BERT classifier (Male)	0.86	0.78
BERT classifier (Female)	0.86	0.80

Figure 3: Accuracy Evaluation results comparing Baseline Logistic Regression model to BERT classifier with overall and genders separated.

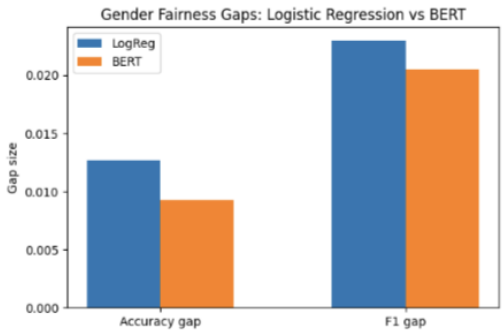

Figure 4: Visualization between the accuracy gap and F1 gap between genders.

3.1 Baseline Model Results

For the baseline classifier of Logistic Regression, the baseline accuracy was 80.79% on the validation set without hypertuning. After fine-tuning the logistic TF-IDF Logistic Regression model, we increased the model accuracy to 82.2% which was conducted on the test data. The macro F1-score on the finetuned model was 76.45%. To conduct a fairness evaluation, we first performed a chi-squared test to measure whether the accuracy of the model's prediction is independent of gender. The result was a value of 0.00, which indicates that the predictions are not independent of the gender. Next, we evaluated the counterfactual augmented dataset on the model trained from the original dataset and found the accuracy to be 82.14%. Counterfactual augmented data did not harm the performance of the classes. Lastly, we conducted the equality of opportunity by calculating the TPR Gap by subtracting the male true positive rate by female. The highest three gaps we found in the baseline model is model with gap of -0.42, rapper with 0.23, and dietitian with -0.22. We can see that the model is highly biased towards females for models and dietitians, as it accurately predicted them compared to males. The opposite is seen in the occupation of rappers, as it is more male-dominated, which shows in the high positive gap.

3.2 BERT Model Results

The BERT classifier significantly outperformed the logistic regression baseline model, achieving an accuracy of 85.99%. This demonstrates the advantage of contextual representation for occupation prediction. The accuracy gap and the F1-score gap between the male and female genders have reduced compared to the baseline model. To conduct a fairness evaluation, we performed a chi-squared test to measure whether the accuracy of the model's prediction is independent of gender. We found that both the baseline and BERT model predictions were not independent of the gender. Next, we evaluated the counterfactual augmented dataset on the BERT model, which was fine-tuned from the original training dataset. Counterfactual augmented data did not harm the performance of the classes, as the accuracy of the data was 85.31%. The classes that were highly biased slightly became more symmetrical, but that came with the cost of accuracy. Lastly, we conducted the equality of opportunity by calculating the TPR Gap by subtracting the male true positive rate by female. The highest three gaps we found in the BERT model is model with -0.38, dietitian with -0.27, and surgeon with 0.16. Comparing these values to the baseline value, the overall gaps are slightly smaller; however, the model failed to eliminate bias in heavily stereotyped professions. Nursing is a heavily stereotyped profession as it is heavily female-dominated. Compared to the logistic regression TPR Gap, the gap in the BERT model is significantly higher, which means it increases bias in some of the occupations. Even with contextual understanding, the model increases or retains stereotypes in occupations.

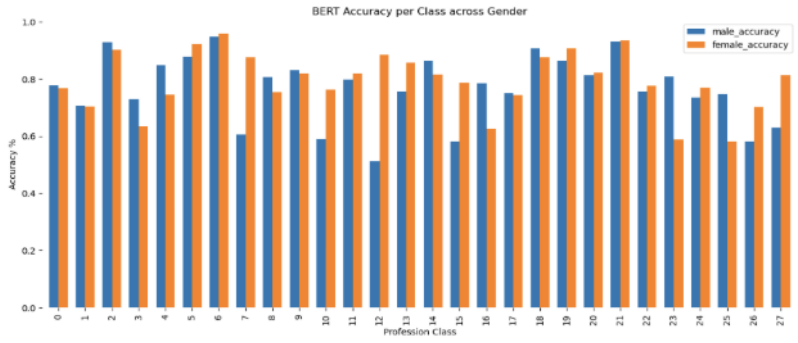

Figure 5: BERT model accuracy per class across genders. The average accuracy for the BERT model is 0.86, while classes such as (2, attorney), (6, dentist), and (21, professor) exceed the overall accuracy. Classes such as (12, model), (15, paralegal), and (10, interior designer) are far below the average, with one gender being classified far more accurately.

3.3 Bias Mitigation Results

To address the observed gender disparities present in the BERT model, we applied both preprocessing and post-processing bias mitigation techniques. One of the preprocessing methods involved fine-tuning the BERT model with the counterfactual augmented dataset. The CFA BERT model achieved an accuracy of 0.8532, which is a decrease in performance compared to the original model. Similarly, the F1-score also decreased, which indicated that CFA did not harm predictive accuracy. To assess the fairness of the CFA model, we looked at the accuracy difference between the original BERT for male and female biographies. Across the predictions for occupations, we noticed it to be slightly more symmetrical than compared to the original model. The post-processing method was implemented by using the group specific confidence threshold adjustment. Similar to pre-processing, the model accuracy decreased to 0.8469, but the gaps between genders were unaffected, making Counter Factual Augmentation the best way we found to reduce bias in occupation classification in BERT models.

4 Conclusion

All in all, comparing the BERT model with the baseline TF-IDF logistic regression model, we found that while contextual language models had higher accuracy, they failed to eliminate demographic bias. In certain occupations, the model may also amplify bias in highly stereotyped occupations. Our fairness evaluation showed consistent gender disparities between both of the two models. We explored multiple bias mitigation techniques and found counter factual augmentation to be the most effective compared to others, as it balanced accuracy and reduced bias. As language models continue to be prevalent in hiring systems, we need to be cautious about how these models are deployed, as they may unintentionally introduce bias within certain occupations. This highlights a stronger need for bias mitigation techniques in these systems to ensure all genders can get an equal opportunity in a job. These results emphasize that while language models can improve accuracy in occupation classification, applying bias mitigation techniques reduces inequalities.