Uncovering Bias and Explaining Decisions in a Text-Based Job Screening Model

Yara Allam

July 2025

1 Dataset

The dataset is a synthetic tabular set of 1500 records of individuals and their resumes. We generated a free-text resume from these tabular features following the template "I am a [GENDER] candidate, [AGE] years old, with [EDUCATION LEVEL] and [EXPERIENCE] years of experience across [NUM COMPANIES] companies. I live [DISTANCE] km from the company. I score [INTERVIEW SCORE] in the interview, with a skill score of [SKILL SCORE] and a personality score of [PERSONALITY SCORE]. I applied through strategy [STRATEGY NUMBER]." In this case, the sensitive attribute is gender (0 for male, 1 for female) and the label is a binary hiring decision (0 for not hired, 1 for hired).

We split the data on an 80/20 train-test split stratified based on the label (Hiring Decision) to have similar label proportions in both the train and test sets. Within the training data, we keep 100% of the male resumes, and 30% of the female resumes, thus creating an imbalanced training set.

2 Model Architecture and Performance

For fine-tuning, we use a base BERT model (110M parameters) over three epochs.

Preliminary results showed an accuracy of 87%, with an F1-score of 91% for the no-hire label and 79% for the hire label. After applying counterfactual data augmentation (seen later), accuracy drops to 82%, with F1-scores dropping to 87% and 72% for no-hire and hire respectively.

3 Fairness Analysis

To analyze any existing biases, we calculate several metrics for the gender group: demographic parity, equal opportunity, and average odds difference.

Demographic parity is defined as the difference in positive predictions in demographic groups.

$$\Delta_{DP} = \left|\Pr(\hat{Y} = 1 \mid A = 1) - \Pr(\hat{Y} = 1 \mid A = 0)\right|$$

Where $\hat{Y}$ is the predicted label (e.g., hire) and A is the sensitive attribute (e.g., gender: 0 = male, 1 = female). Before mitigation, the difference was extremely small at about 0.037. After mitigation, that number rose slightly to 0.095.

Equal opportunity compares true positive rates across groups.

$$\Delta_{EO} = \left|\Pr(\hat{Y} = 1 \mid Y = 1, A = 0) - \Pr(\hat{Y} = 1 \mid Y = 1, A = 1)\right|$$

It measures whether qualified individuals (i.e., $Y = 1$) are equally likely to be selected. Before mitigation, it was at about 0.136, rising to 0.211 after mitigation.

Average odds difference averages the differences in true positive and false positive rates between groups.

$$\Delta_{AO} = \frac{1}{2}\left(\left|\Pr(\hat{Y} = 1 \mid Y = 1, A = 0) - \Pr(\hat{Y} = 1 \mid Y = 1, A = 1)\right| + \left|\Pr(\hat{Y} = 1 \mid Y = 0, A = 0) - \Pr(\hat{Y} = 1 \mid Y = 0, A = 1)\right|\right)$$

Before mitigation, its value was at 0.101, and did not change significantly afterwards, only increasing to 0.109.

From our findings, we can see that the model does not exhibit significant bias towards women during a hiring process. To further visualize this, we plot the predicted hiring decisions by gender:

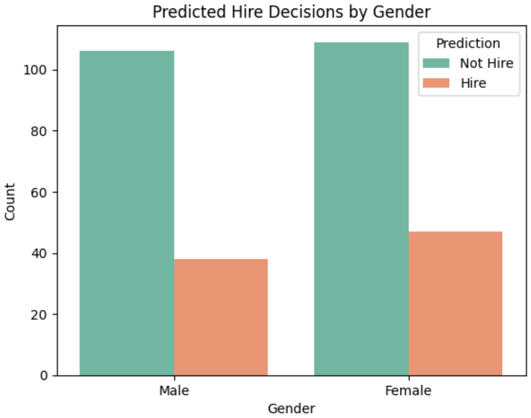

Figure 1: Hire Rate by Gender (Before Mitigation)

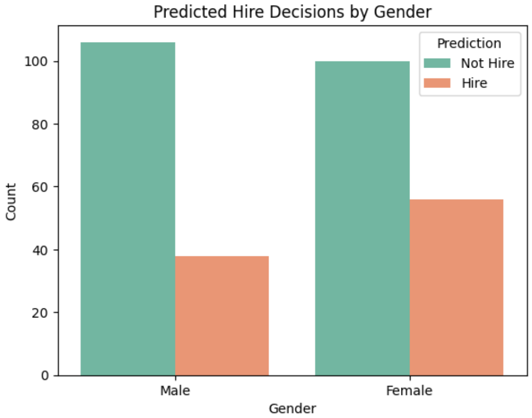

Figure 2: Hire Rate by Gender (After Mitigation)

4 Explainability Results

We use SHAP (**SH**apley **A**dditive ex**P**lanations), an explainability method based on game-theory. It assigns each feature a value that indicates its contribution to the prediction. In our case, it helps us quantify how individual words influence the model's decision to hire/not hire.

From our findings, gender does not appear to influence the model's decisions, but rather the strategy that the candidate applied through, as well as the scores on the technical tests. This persists before and after mitigation.

From results, strategy 1 appears to be the most effective, as opposed to strategies two and three.

5 Mitigation Results and Tradeoffs

As discussed above, the model does not appear to exhibit a bias towards women even before mitigation. In fact, it seems to slightly favor women in the hiring process. Regardless, we continue with a mitigation technique known as counter-factual data augmentation. It works by creating alternate versions of training examples by modifying sensitive attributes. In our case, we swap 'male' and 'female' in every resume text.

Results show that the model now favors women even more in the hiring process (as can be seen in the fairness analysis conducted above).

A trade-off of this technique is a lower accuracy This might be due to the synthetic variations in the training data that may or may not reflect real patterns, especially when we modify only a single attribute.