CIS-2025-19 Research Internship Challenge

Bias Detection and Explainability in AI Models

Prepared by
Azza Hassan Said

azzahassan1118051@gmail.com

Introduction

This challenge explores bias detection and mitigation in AI-powered resume screening. We build a binary classifier to predict hiring decisions from synthetic resumes, with a focus on gender fairness. Our pipeline includes converting structured data to text, training TF-IDF and DistilBERT models, evaluating fairness with group metrics, interpreting predictions using SHAP, and applying counterfactual augmentation for bias mitigation. We summarize our key methods, results, and trade-offs between accuracy and fairness.

1.Dataset Description and Sensitive Feature Encoding

The dataset is a synthetic resume screening dataset with 10 structured features per applicant (e.g., Age, Gender, EducationLevel, ExperienceYears). A custom function was used to convert these into natural-language resumes for NLP model input.

- Gender (sensitive attribute) was encoded as binary (0 = female, 1 = male) and directly reflected in the resume text.
- EducationLevel was mapped to degrees (e.g., 2 → Bachelor's).
- RecruitmentStrategy: Encoded using a custom map

Each resume was phrased like: "I am a male candidate, 30 years old, with a Master's degree... I applied through campus recruitment."

2. Model Architecture and Performance

We experimented with two models:

- TF-IDF + Logistic Regression: Preprocessing included text cleaning (lowercasing, lemmatization, stopword removal) and exploratory data analysis (EDA). This model was evaluated under a deliberately imbalanced split (90% male, 10% female), as per challenge requirements.

TF-IDF Performance:

Accuracy: 82% **F1-score (Hire): 0.80%**

- DistilBERT Fine-tuning: Used on synthetic resume text without heavy preprocessing. Tokenization and context handling are inherent to the model.

DistilBERT Performance (imbalanced split):

Accuracy: 82.3% **F1-score (Hire): 0.79**

Further fairness evaluation was performed using BERT due to its superior contextual understanding.

3. Fairness Analysis

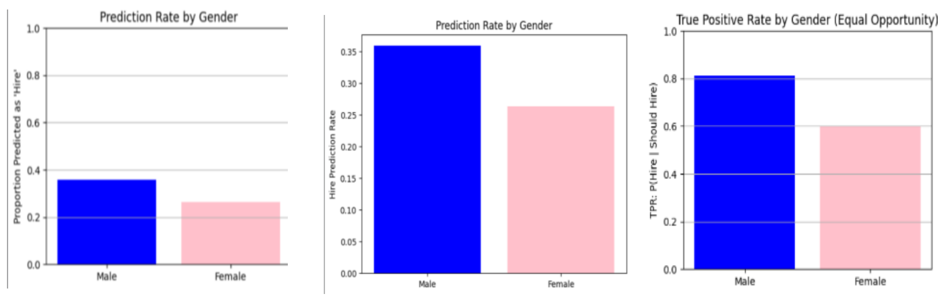
Using Gender as the sensitive attribute, we evaluated fairness on the test set:

-Demographic Parity Difference: 0.0951

-Equal Opportunity Difference: 0.2113

-Average Odds Difference: 0.1099

Observation: The model was slightly more favorable toward males based on higher opportunity metrics



4. Explainability Results

We applied **SHAP** to explain predictions for 5 samples (3 "Hire", 2 "No-Hire"):

- Top contributing terms: "campus", "recruitment", "Master'", "Bachelor'", "interview".
- **Gendered terms** (male, female) had negligible SHAP values and were not among top contributors.

This suggests **no direct influence of gender terms** in the top model decisions.

5. Bias Mitigation and Trade-offs

We implemented **counterfactual data augmentation**:

- Created flipped-gender resumes by swapping gender terms (e.g., "male" → "female").
- Augmented the training set with these counterfactual samples.

Metric	Before	After (Augmented)
Accuracy	82.3%	88.0%
F1-score (Hire class)	0.72	0.80
Demographic Parity Diff.	0.0951	0.0908
Equal Opportunity Diff.	0.2113	0.1991
Average Odds Difference	0.1099	0.1089

Result: The counterfactual augmentation strategy improved both performance and fairness metrics with minimal trade-offs. It helped reduce gender bias while maintaining strong predictive power.