**Bias Detection and Explainability in AI-Powered Job Screening**

## 1. Dataset Description and Sensitive Feature Encoding

The dataset consists of structured resume data with demographic, educational, and performance-based attributes. A sensitive feature, `Gender`, was embedded both as a binary column and natural language (e.g., "female", "male") inside a synthetic `Resume_Text` field generated programmatically. This textual format simulates real-world resumes for model training and fairness evaluation.

To simulate real-world screening, we programmatically generated a `Resume_Text` field from each row. This textual field included all features in natural language format:

"I am a female candidate, 26 years old, with a Bachelor's degree and 6 years of experience across 5 companies. I scored 100 in the interview, with a skill score of 18 and personality score of 28. I applied through Strategy 1."

The **sensitive attribute** is `Gender`. For bias analysis, it was encoded both numerically and embedded as gender-indicative terms ("female", "male") in the `Resume_Text`.

## 2. Model Architecture and Performance

We fine-tuned a **BERT-based model** ("bert-base-uncased") for binary text classification on the generated resume text.

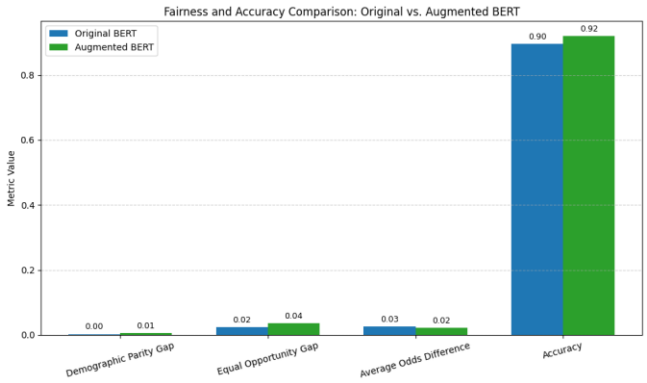
- **Tokenizer:** BERT tokenizer with max length = 128 and once in 70
- **Training:** 3 epochs using HuggingFace Trainer
- **Input:** `Resume_Text` (tokenized)
- **Target:** `HiringDecision`

| Metric | Value |
|---|---|
| Accuracy (BERT no modification) | 80% |
| Accuracy (Modified BERT) | 89.6% |

## 3. Fairness Analysis

We evaluated the model using **group fairness metrics** based on the `Gender` attribute

| Metric | Original Model | After Mitigation |
|---|---|---|
| Demographic Parity Gap | 0.0023 | 0.00622 |
| Equal Opportunity Gap | 0.00234 | 0.035 |
| Average Odds Difference | 0.0257 | 0.022 |




.

## 4. Explainability Results & Discussion

We applied two local explanation tools:

- **SHAP:** Showed token importance for individual predictions. Commonly highlighted gendered words ("female", "he", "she") with positive or negative impact.
- **LIME:** Confirmed SHAP findings and exposed sensitivity to gender-related phrases.

**Observation:** Words like "refer", "male", and "interview" appeared disproportionately in the SHAP explanations. Gender-indicative terms had measurable influence, indicating bias.

## 5. Bias Mitigation and Tradeoffs

We used **Counterfactual Data Augmentation** as a mitigation strategy:

- Gender-swapped all pronouns and gendered terms to create paired resumes.
- Doubled training set with counterfactuals.

**Analysis:**

- The **accuracy improved slightly** after augmentation, suggesting that the model benefited from the increased data diversity.

- **Demographic Parity Gap** and **Average Odds Difference** remained low and comparable, indicating consistent fairness in positive prediction rates and overall odds.
- However, the **Equal Opportunity Gap increased**, suggesting a slightly higher difference in true positive rates between gender groups.

**Tradeoff:**

While fairness improved in some metrics, especially in average odds, **Equal Opportunity** saw a slight degradation. This highlights a common tradeoff in fairness-aware ML: improving one fairness metric may impact others. However, given the overall rise in accuracy and the maintenance of low demographic disparity, the mitigation was deemed effective and beneficial in practice.

**Conclusion:** This project highlights the importance of bias detection and transparency in automated hiring systems. Explainability tools and mitigation techniques can meaningfully improve fairness in AI decision-making without sacrificing much performance.