Abdelrahman Radwan

48 Hours Challenge: Bias Detection And Explainability In Ai Models

1. Dataset Description and Encoding of Sensitive Features

The dataset is read from a CSV file

A new column features_text was created by combining all important features including the sensitive feature Gender, as well as other attributes like:

- EducationLevel
- ExperienceYears
- PreviousCompanies
- InterviewScore
- SkillScore
- PersonalityScore
- RecruitmentStrategy

The features were formatted into text using:

df['features_text'] = df.apply(lambda row: f"Gender: {row['Gender']} EducationLevel: {row['EducationLevel']} ...", axis=1)

This approach embeds sensitive attributes like Gender directly into the input text, which is then passed to the BERT model.

---

2. Model Architecture and Performance

The model used is a BERT (Bidirectional Encoder Representations from Transformers) classifier:

- Model: bert-base-uncased
- Tokenizer: Hugging Face's BertTokenizer

- Input: features_text
- Output: Binary classification (HiringDecision)

Steps:

- X_train and X_test were tokenized with truncation and padding.
- Model trained for 3 epochs with a batch size of 8.
- Optimizer: Adam
- Loss: SparseCategoricalCrossentropy
- Metric: SparseCategoricalAccuracy

Evaluation Result:

loss, accuracy = bert_model.evaluate(...)

The final loss and accuracy (loss: 0.4183 - accuracy: 0.8400) provide the baseline performance before fairness analysis and mitigation.

3. Fairness Analysis (with Plots and Metrics)

Fairness metrics were calculated using different methods:

Demographic Parity:

The goal of demographic parity is to ensure that the model's predictions are independent of sensitive attributes (like Gender). The fairness is analyzed by comparing the rates at which different groups are selected (hired) by the model.

Equal Opportunity Difference:

This metric measures fairness in terms of true positive rates. It compares how often the model correctly identifies candidates from different gender groups.

Average Odds Difference:

This metric looks at the average of the true positive rate and false positive rate differences across groups. It evaluates how the model performs across different groups

Plot Disparities in Prediction Rates Across Gender Groups:

The disparities were plotted to visualize how different gender groups were impacted by the model's predictions.

---

4. Explainability Results and Discussion

Explainability was approached using **model interpretability techniques**, such as **bias attribution**.

The notebook provides code for extracting and displaying model predictions, focusing on identifying whether sensitive features (like Gender) influenced the model's decision

Using these techniques, the discussion elaborates on how the model's decisions are related to sensitive attributes and evaluates how interpretable the results are for both technical and non-technical stakeholders.

---

5. Mitigation Results and Tradeoffs

Bias mitigation methods were applied to adjust for any detected disparities:

Mitigation Results:

The notebook explored using various techniques to mitigate bias, such as re-weighting training samples or modifying the decision thresholds for different groups.

Tradeoffs:

The tradeoffs associated with bias mitigation were discussed. While fairness was improved, model performance (in terms of accuracy) was slightly reduced. The mitigation methods were evaluated by comparing pre- and post-mitigation accuracy scores.