Bias Detection and Mitigation in Pre-Trained Natural Language Processing Models

Domain: *AI Safety - Fairness and Trustworthiness in Machine Learning*

Problem Statement

Pre-trained natural language processing models often show systematic bias that can lead to unfair treatment of demographic groups even though they have a widespread adoption in the real world. These ML models are trained on large datasets from internet and learn social bias present in training datasets. Sentiment Analysis model can associate some profession with a gender or a toxicity detection system can unfairly flag content related to an ethnic group so these bias pose significant risks when they are deployed in critical applications like job hiring systems.

Objective

- Detect bias in pre-trained models by testing performance with multiple demographic groups and with sensitive content.
- Measure unfairness with the help of existing fairness metrics and evaluate the extent of bias in the model.
- Implement bias mitigation techniques to reduce identified bias and maintain the performance of the model at the same time.
- Evaluate effectiveness of the mitigation techniques through analysis of model behavior before and after the implementation of mitigation techniques.
- Provide insights on identifying and addressing bias in similar NLP systems.

Solution Approach

Phase 1: Select a pre-trained model specifically sentiment analysis ones because of their interpretability and real world relevance. Set baseline performance metrics on the evaluation dataset.
Phase 2: Try to develop test suites that have sentences that probe for gender, racial and occupational biases and will measure bias using established fairness metrics.
Phase3: Implement and compare the mitigation strategies including post-processing threshold adjustment, prompt based debiased techniques and each of these will be evaluated for their effectiveness in reducing bias.
Phase 4: Conduct comparative analysis of the performance of the model before and after mitigation strategies were applied and measure both fairness improvement and any potential degradation overall. Try to visualize results with the help of charts and statistical analysis.