Depression Classification and Mitigating Gender Bias

Team 4
Leo Beck
Eric Fithian
Naveena Ganesan
Indu Varshini Jayapal

Introduction

Recent advancements in digital healthcare have leveraged speech-based machine learning (ML) technologies to unobtrusively monitor mental health states via devices like smartphones and wearables. However, these technologies might inadvertently amplify demographic biases, affecting their fairness and effectiveness. This project explores gender bias in ML algorithms using speech data to detect depression, based on the Distress Analysis Interview Corpus Wizard of Oz (DAIC-WoZ) dataset.

Methods

We applied simple feed-forward neural networks and random forests for separate depression and gender classifications. Feed Forward Neural Networks yielded higher accuracy results, while Random Forests performed similarly with a slight advantage in interpretability.

For depression, we evaluated accuracy, balanced accuracy, and equality of opportunity across genders. For gender, we assessed simple and balanced classification accuracies.

We employed filter feature selection methods, including feature importance from decision tree classifiers and mutual information, to identify key features influencing both depression and gender predictions. Various numbers of these features were removed from the models.

To mitigate potential bias, we utilized techniques such as gender-important feature removal and sampling weights for features.

Discussion

Our findings underscore the challenge and necessity of designing bias-aware ML models in healthcare settings. While our models performed effectively in identifying depression and gender, the disparity in accuracy between different demographic groups suggests that more nuanced approaches are needed to mitigate bias. Future work will focus on refining models to enhance fairness without compromising diagnostic performance.

> **Removing gender features while modeling for Depression can help mitigate bias, but there is always a trade-off between bias mitigation and model performance in terms of accuracy.**

What we recommend from our experiments

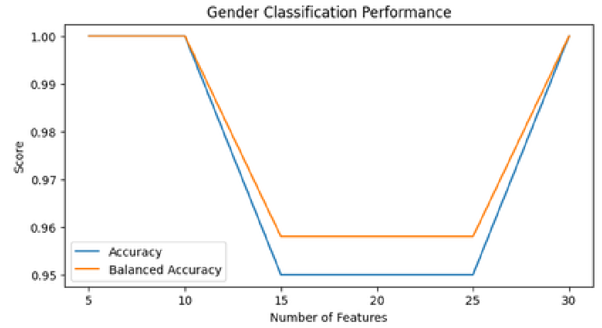
Use classification algorithms that are inherently capable of handling imbalanced datasets, explore gender-dependent features, and experiment with adding and removing a varying number of features until the desired accuracy is attained.

Check out our codebase here:
https://github.com/InduVarshini/ML-HW5-Identify-Depression-and-Gender-From-Audio-Features

Results

Gender Classification

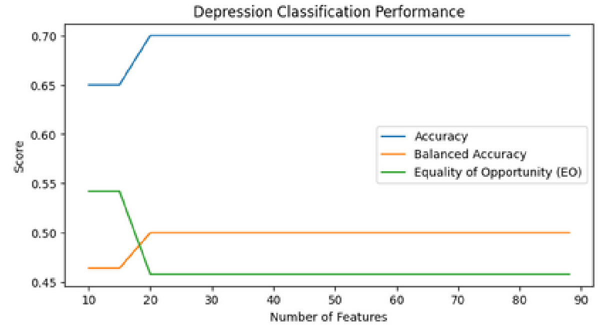
Feature selection experiments for Gender showed that our models performed the best with 5 features. Our analysis of the top features indicated that 'F0semitoneFrom27.5Hz_sma3nz_percentile-50.0' was the most indicative feature of gender at a weight of around 0.69.



Before feature filtering and bias mitigation we found that gender classification received an accuracy and a balanced accuracy of 100% on the test set. The results remained the same after feature selection.

Depression Classification

Feature selection experiments for Depression showed that our models performed the best with 15 features.



Before feature selection, Depression classification model received test accuracy of 70%, a balanced test accuracy of 50%, and a test equality of opportunity (EO) of 45.8%. These were calculated between men and women, highlighting potential bias. After feature selection, accuracy dropped to 65% and balanced accuracy to 46.4% and EO increased to 54.2%.

Bias Mitigation

Removing gender-dependent features improved fairness (Equality of Opportunity: 45.8% → 54.2%) but slightly decreased overall accuracy (70% → 65%) and balanced accuracy (50% → 46.4%), with male accuracy decreasing (91.7% → 83.3%) and female accuracy remaining unchanged (37.5%). Sample re-weighting had minimal impact on performance and fairness. Our study highlights the trade-off between mitigating bias and maintaining performance in depression classification models.