Evaluation of Gender Bias in Depression Detection Models

Kalyani Jaware Erin Richardson Dayn Reoh Dhanavikram Sekar

University of Colorado Boulder

Introduction

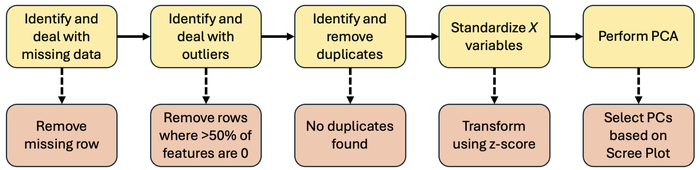
Digital healthcare is increasingly leveraging speech-based ML technologies due to the ease of unobtrusive data collection through smartphones and wearables. Speech carries valuable insights into human behavior and mental state, stemming from the complex interplay of cognitive planning and articulation. Acoustic measures from speech, like prosody, provide critical information for mental healthcare, but may also be confounded by demographic factors, potentially leading to biases in ML algorithms detecting depression.

In this project we **build predictive models of depression and gender**. We also **explore gender bias** in these models. We use the Distress Analysis Interview Corpus Wizard of Oz (DAIC-WoZ) dataset [1]. Further information about the dataset is displayed below.

Category	Male	Female	Total	Male to Female ratio
Train	63	44	87	1.43
Test	14	17	20	0.82
Total	77	61	107	1.26

Data Preparation

To prepare the data for **traditional machine learning models**, all participant data were consolidated into a single dataset, incorporating Participant ID, Gender, and Depression as additional features.



For **Convolutional Neural Networks** and **LSTM based RNNs**, each participant's data was padded with 0s to ensure that the data matrices are of the same size.

Depression Classification

Various models were used to classify depression, such as Logistic Regression, Random Forests, fully connected neural networks (FCNNs), convolutional neural networks (CNNs) and Long Short Term Memory (LSTM) recurrent neural networks. Of these, the FCNNs and Balanced Random Forests gave the best results in terms of balanced accuracy, precision, recall and Equality of Opportunity (EO). The confusion matrix of these models are displayed in Figures 1 and 2 respectively.

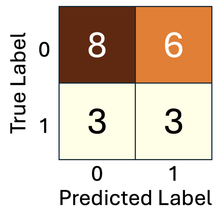


Figure 1. Balanced Random Forests

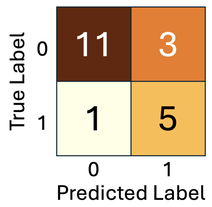


Figure 2. Fully Connected Neural Networks

Gender Classification

Three different models were used to classify gender: CNNs, FCNNs, and Random Forests along with the balanced implementation of Random Forests. Similar to the depression classification, FCNNs and Random Forests gave the best results. Their confusion matrices are displayed in 3 and 5 respectively.

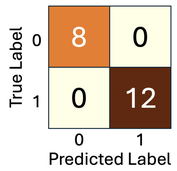


Figure 3. Random Forests

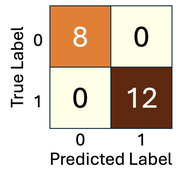


Figure 4. Fully Connected Neural Networks

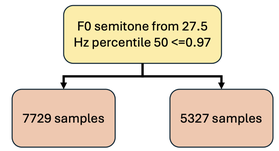


Figure 5. Decision tree of depth 1 with 95% accuracy.

The feature importance obtained from the random forest model was investigated in order to interpret the 100% accuracy. One interesting finding from this investigation was 95% accuracy can be obtained by just using one feature which is 'F0semitoneFrom27.5Hz_sma3nz_percentile50.0'. This feature represents median pitch of the person's voice in that specific time period.

Results before Feature Selection

The monitored metrics for each model is displayed in 1

Model	Target	Acc.	Bal. Acc.	EO
Fully Connected Neural Nets	Depression	0.80	0.81	0.80
Balanced Random Forests	Depression	0.55	0.53	0.40
Fully Connected Neural Nets	Gender	1.00	1.00	–
Random Forests	Gender	1.00	1.00	–

Feature Selection based on Feature Importance

For the first filter approach, random forest model's feature importance was used to select the most informative for gender and depression classification. On varying n from 10 to 80, it was found that by using 60 most important features, the random forests model was able to build 300 trees of depth 200, which were able to classify a person's depression with 60 % balanced accuracy and an EO score of 0.6. Similarly, the model was able to classify the gender of a person with 95 percent accuracy with just one feature. But to classify with 100 percent accuracy 7 features were required. The progression of metrics on increasing the number of features used is displayed in 8 and 9

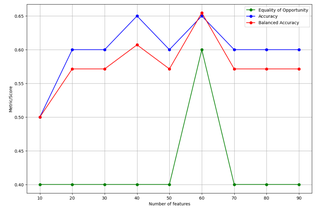


Figure 6. Feature Selection for Depression using feature importance

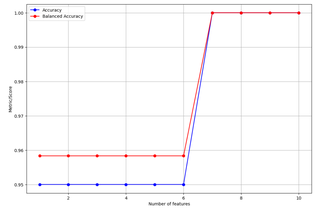


Figure 7. Feature Selection for Depression using gender importance

Feature Selection based on Fisher's Criterion

For selecting the n most informative features for gender and depression classification using a CNN, the Fisher Criterion was used. Varying n features from 5 to 30, it was found that using the 15 most important features led to a 90% accuracy and 89% balanced accuracy for gender classification. Similarly, using the 15 most important features led to a 70% accuracy, 60% balanced accuracy, and 60% EO score for depression classification.

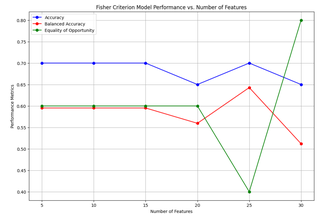


Figure 8. Feature Selection for Depression using Fisher's Criterion

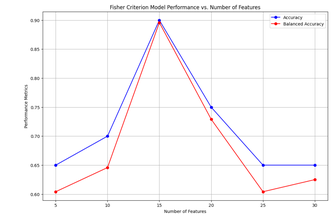


Figure 9. Feature Selection for Gender using Fisher's Criterion

Mitigating bias via removing gender-dependent features

The FCNN was used after removing 8 gender dependent features from the dataset. The accuracy was impacted negatively in case of class 1 that is, in case of participants with depression.

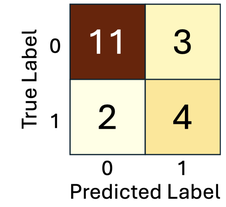


Figure 10. Gender-independent FCNN.

Future Work

- Implement **internal validation** to improve reliability of performance metrics (e.g., cross-validation or bootstrapping)
- Implement **ensemble learning** to leverage the strengths of different models and create stronger final predictions
- Investigate **bias based on other demographic metrics** (e.g., race, first language, etc) and develop strategies to mitigate it

References

[1] J. Gratch, R. Artstein, G. Lucas, G. Stratou, S. Scherer, A. Nazarian, R. Wood, J. Boberg, D. DeVault, S. Marsella, D. Traum, S. Rizzo, and L.-P. Morency, "The Distress Analysis Interview Corpus of human and computer interviews," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, May 2014, pp. 3123–3128.