SPEECH EMOTION RECOGNITION

BIANNA GAS

BACKGROUND

- Speech Emotion Recognition (SER) is the task of speech processing that aims to recognize and categorize the emotions expressed in spoken language.

- The goal is to determine the emotional state of a speaker, such as happiness, anger, or sadness from their speech patterns such as loudness, pitch, rhythm, and tone.

SER USE CASES

Use Case n° 1

The technology can detect a customers emotional state from the automated menu and automatically transfer customers with negative emotional state to a live agent or prioritize those with such emotions in the live queue

Use Case n° 2

SER can be used to detect acoustic cues to recognize hate speech and discrimination and automatically end the call to protect employees from verbal abuse

Use Case n° 3

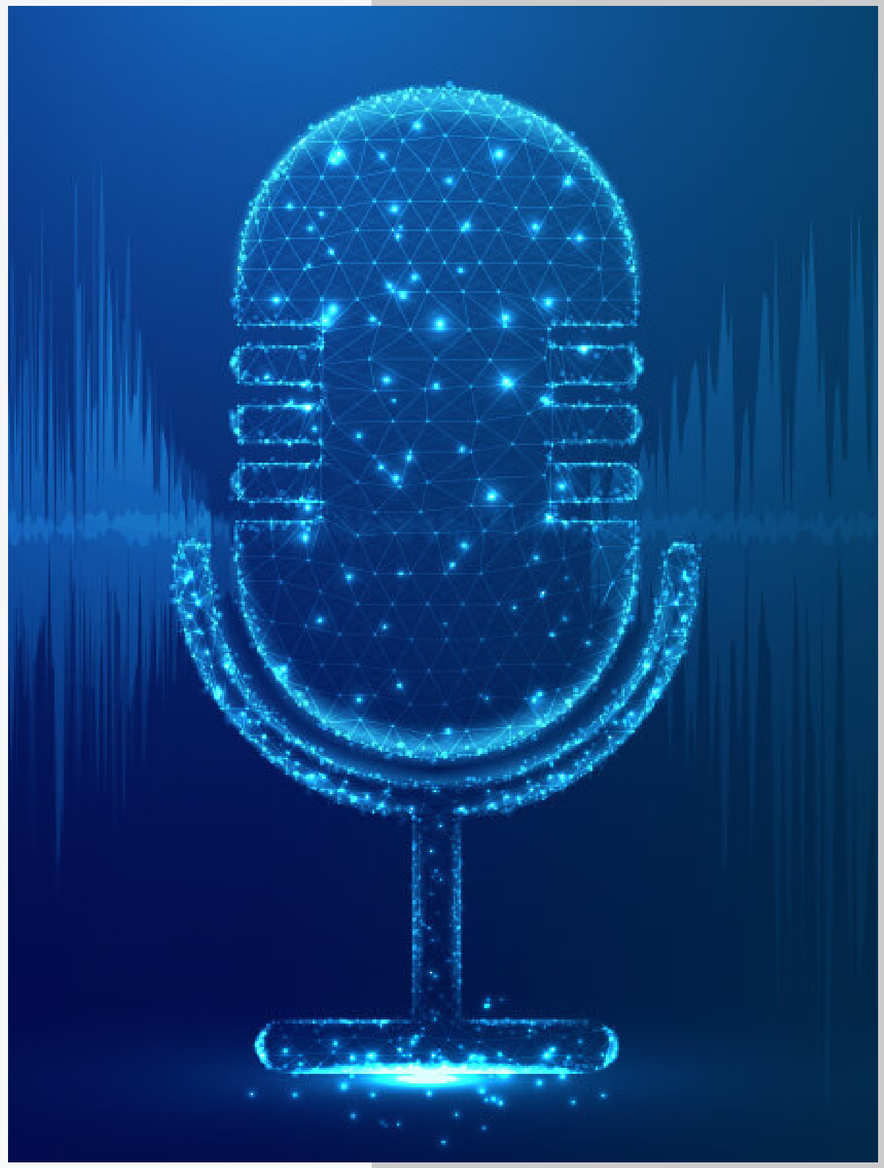
SER can help expand on existing databases collected from recording customer calls for employee training to better train employees on how to respond to customers experiencing dissatisfaction

verizon

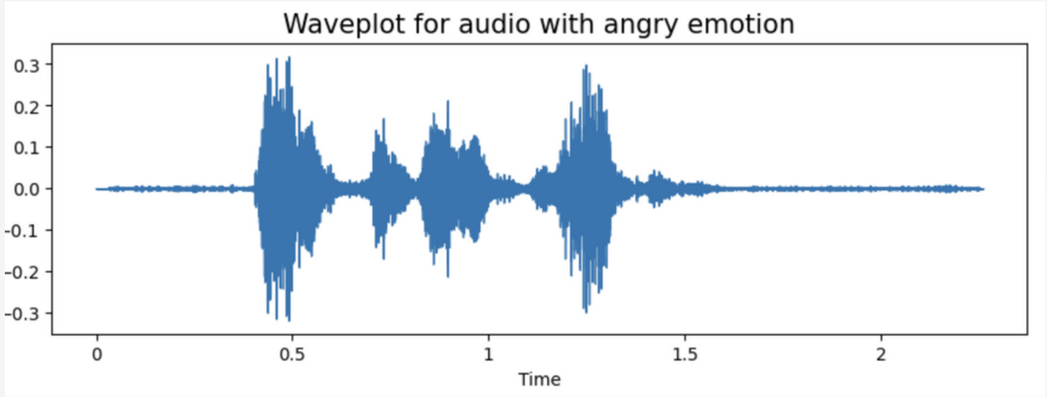
DATASET

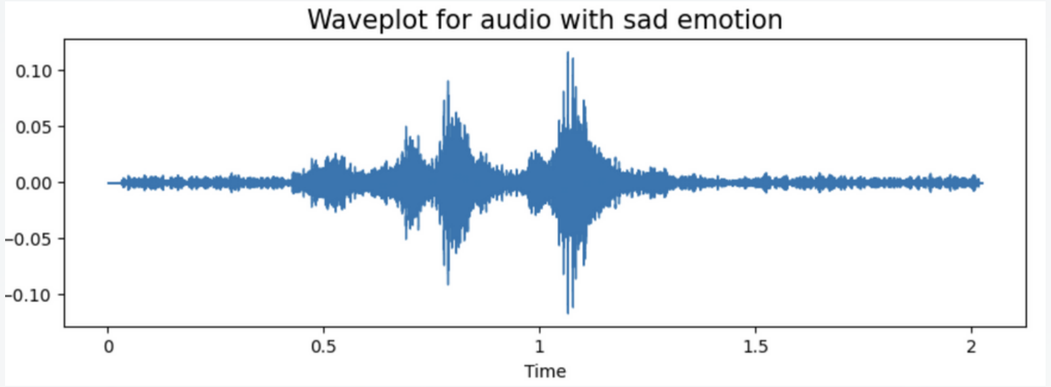
CREMA-D (Crowd-sourced Emotional Multimodal Actors Dataset) is a data set consisting of:

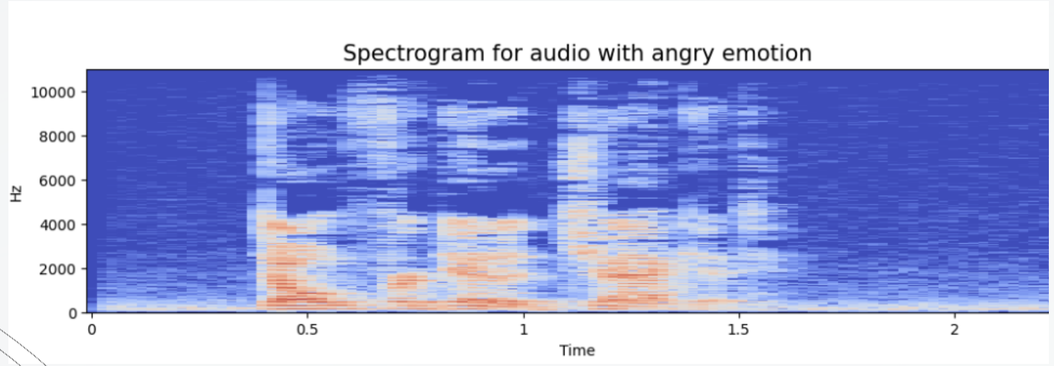
- 7,442 original clips from 91 actors. These clips were from 48 male and 43 female actors between the ages of 20 and 74 coming from a variety of races and ethnicities.
- Actors spoke from a selection of 12 sentences.
- The sentences were presented using one of six different emotions (Anger, Disgust, Fear, Happy, Neutral, Sad) and four different emotion levels (Low, Medium, High, and Unspecified).

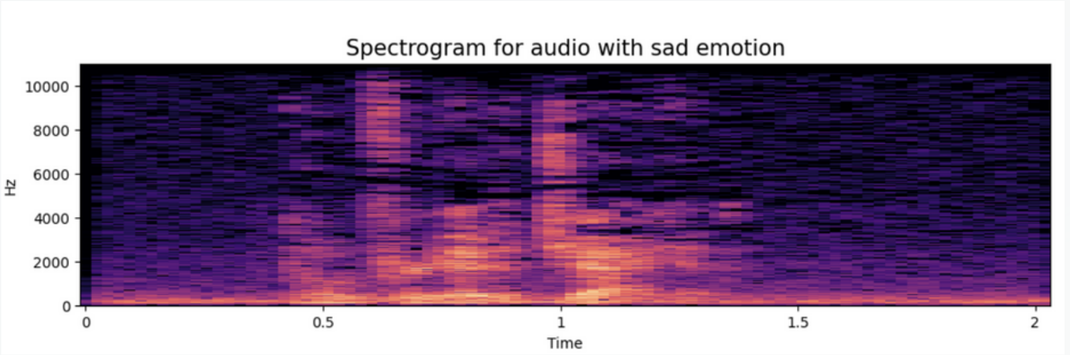


SOUND FEATURES





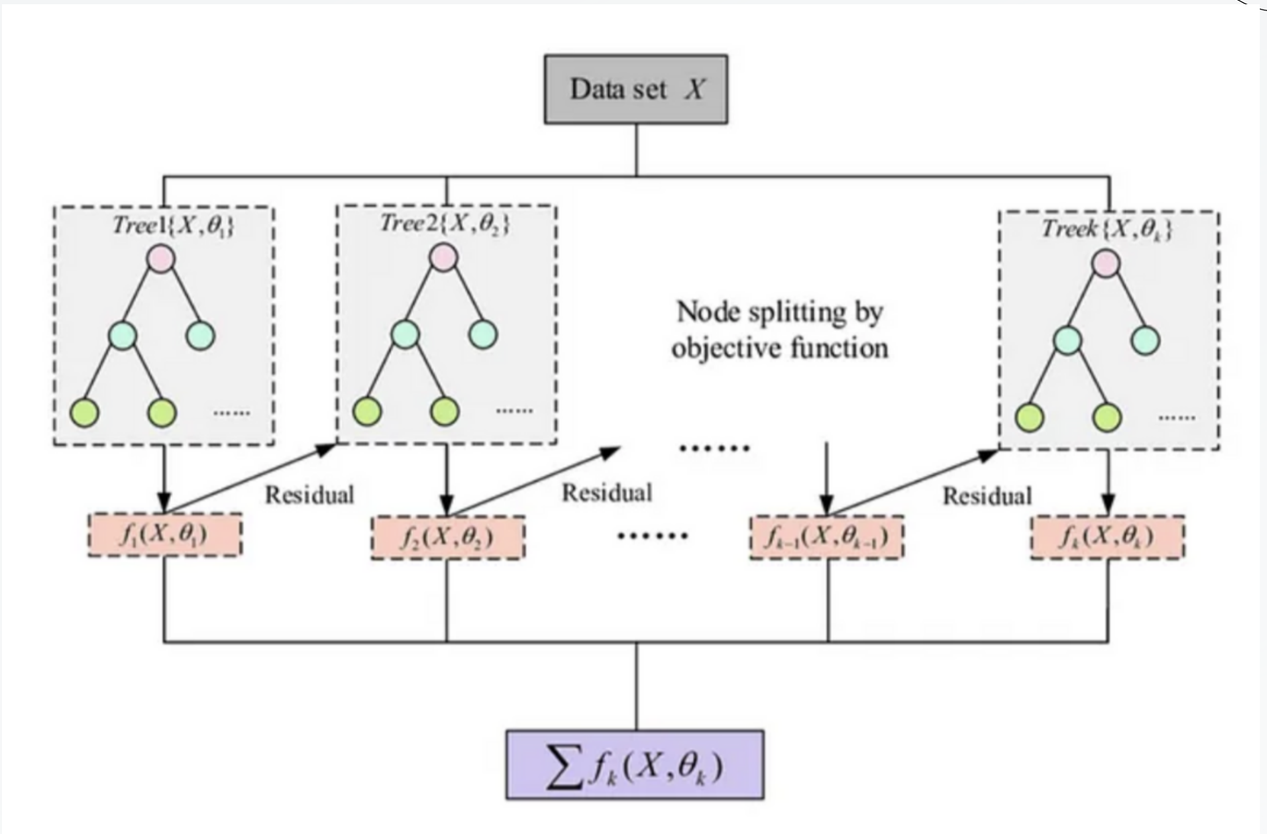




NUMERIC FEATURE EXTRACTION + XGBOOST

Extracted Features:

- Zero Crossing Rate
- Mel-Frequency Cepstral Coefficients
- Root Mean Square
- Mel-spectrogram
- Chromogram
- Spectral Centroid
- Spectral Bandwidth
- Spectral Contrast
- Spectral Rolloff
- Tonnetz



XGBOOST RESULTS

	precision	recall	f1-score	support
ANG	0.61	0.72	0.66	309
DIS	0.41	0.30	0.35	327
FEA	0.44	0.38	0.41	310
HAP	0.45	0.40	0.42	331
NEU	0.42	0.50	0.46	264
SAD	0.50	0.59	0.54	312
accuracy			0.48	1853
macro avg	0.47	0.48	0.47	1853
weighted avg	0.47	0.48	0.47	1853

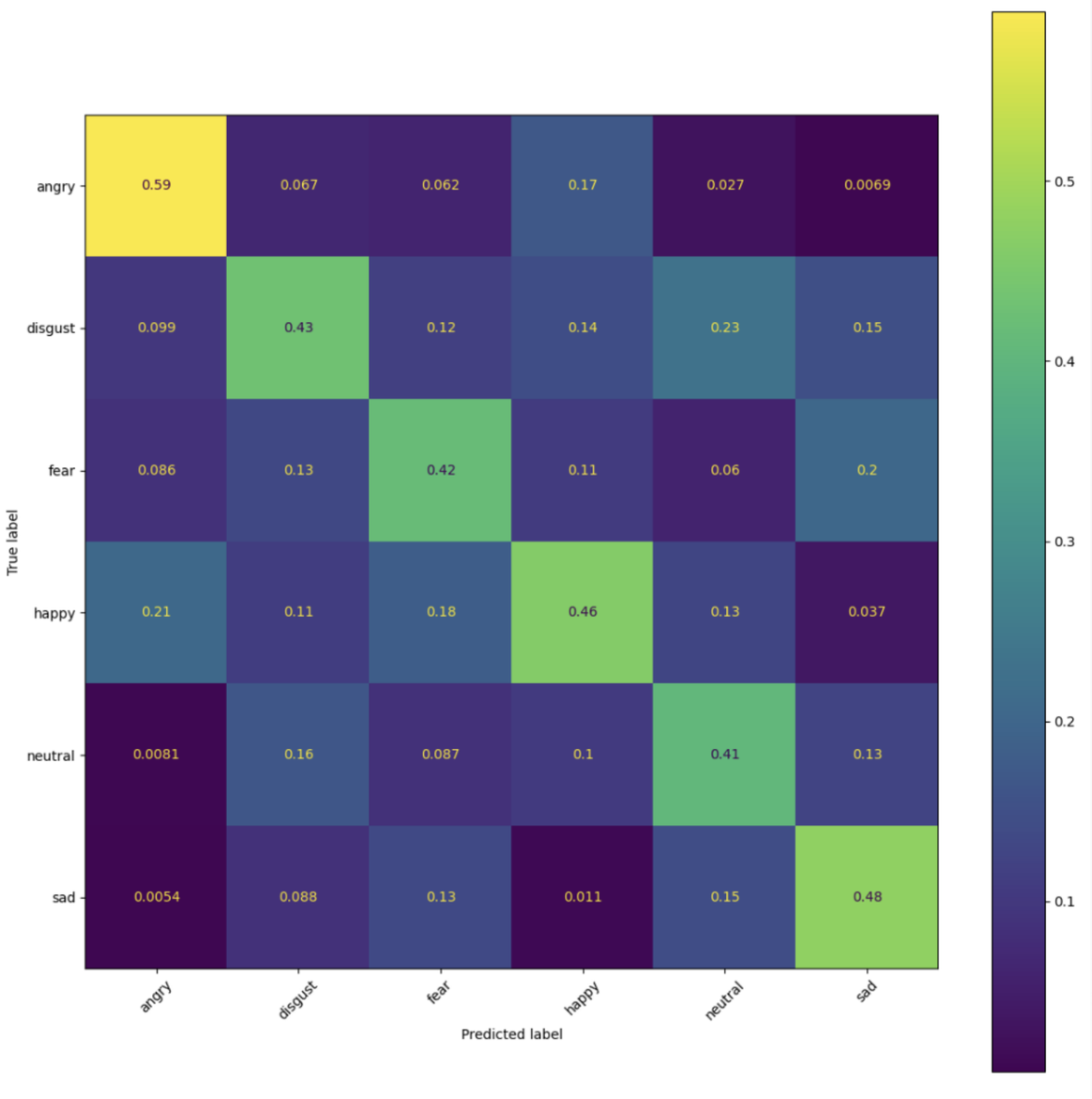
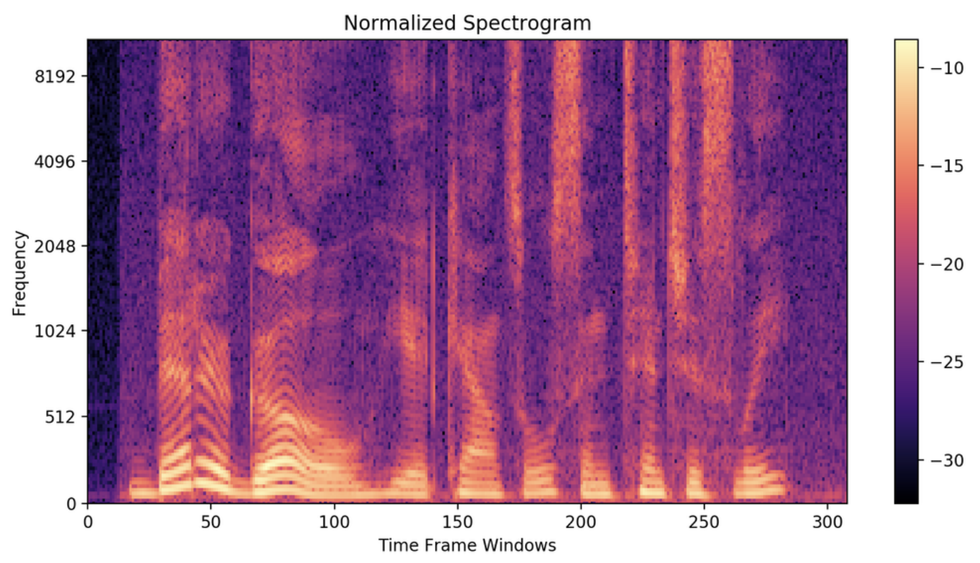
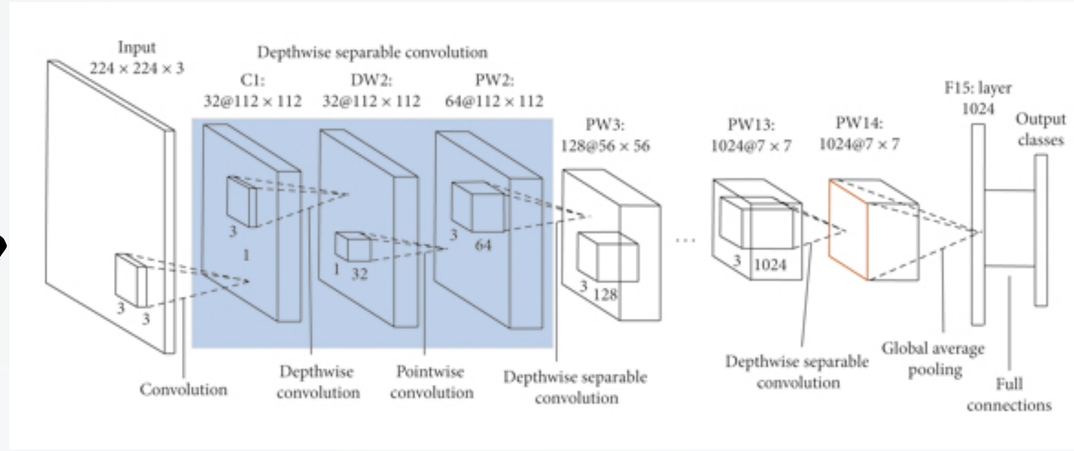


IMAGE CLASSIFICATION + MOBILENET



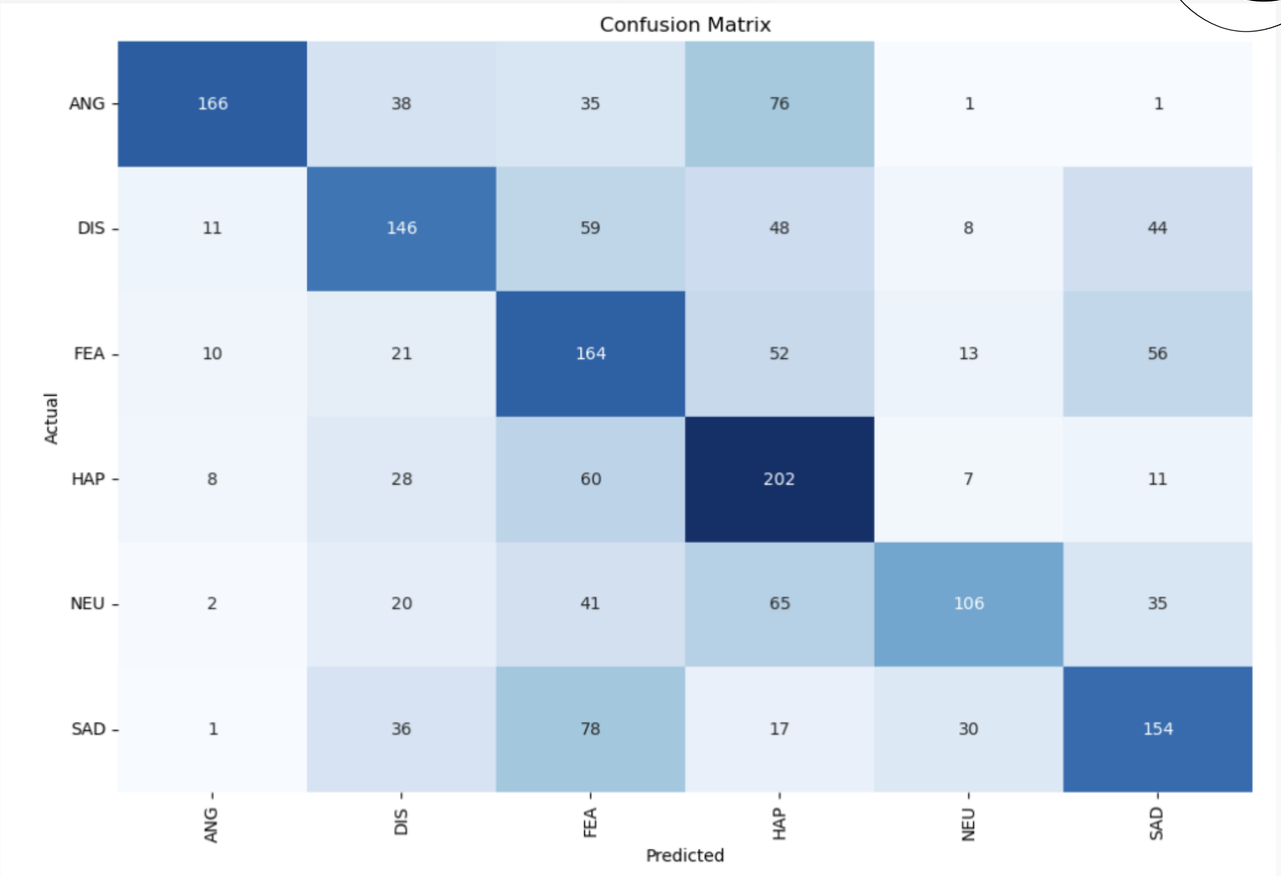


MOBILENET RESULTS

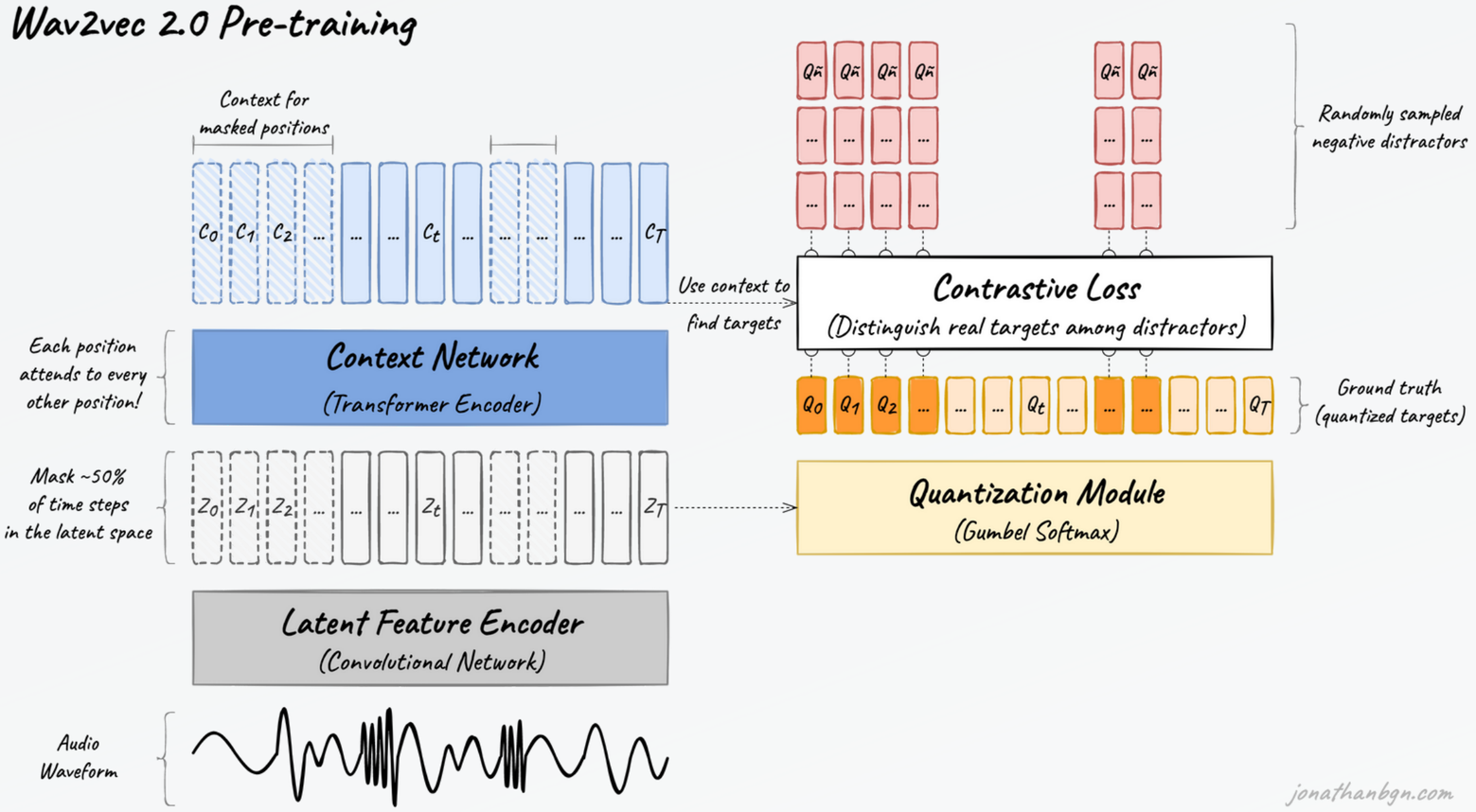
```
Classification Report:
----------------------
              precision    recall  f1-score   support

         ANG     0.8384    0.5237    0.6447       317
         DIS     0.5052    0.4620    0.4826       316
         FEA     0.3753    0.5190    0.4356       316
         HAP     0.4391    0.6392    0.5206       316
         NEU     0.6424    0.3941    0.4885       269
         SAD     0.5116    0.4873    0.4992       316

    accuracy                         0.5070      1850
   macro avg     0.5520    0.5042    0.5119      1850
weighted avg     0.5499    0.5070    0.5125      1850
```



TFWAVE2VEC CLASSIFICATION

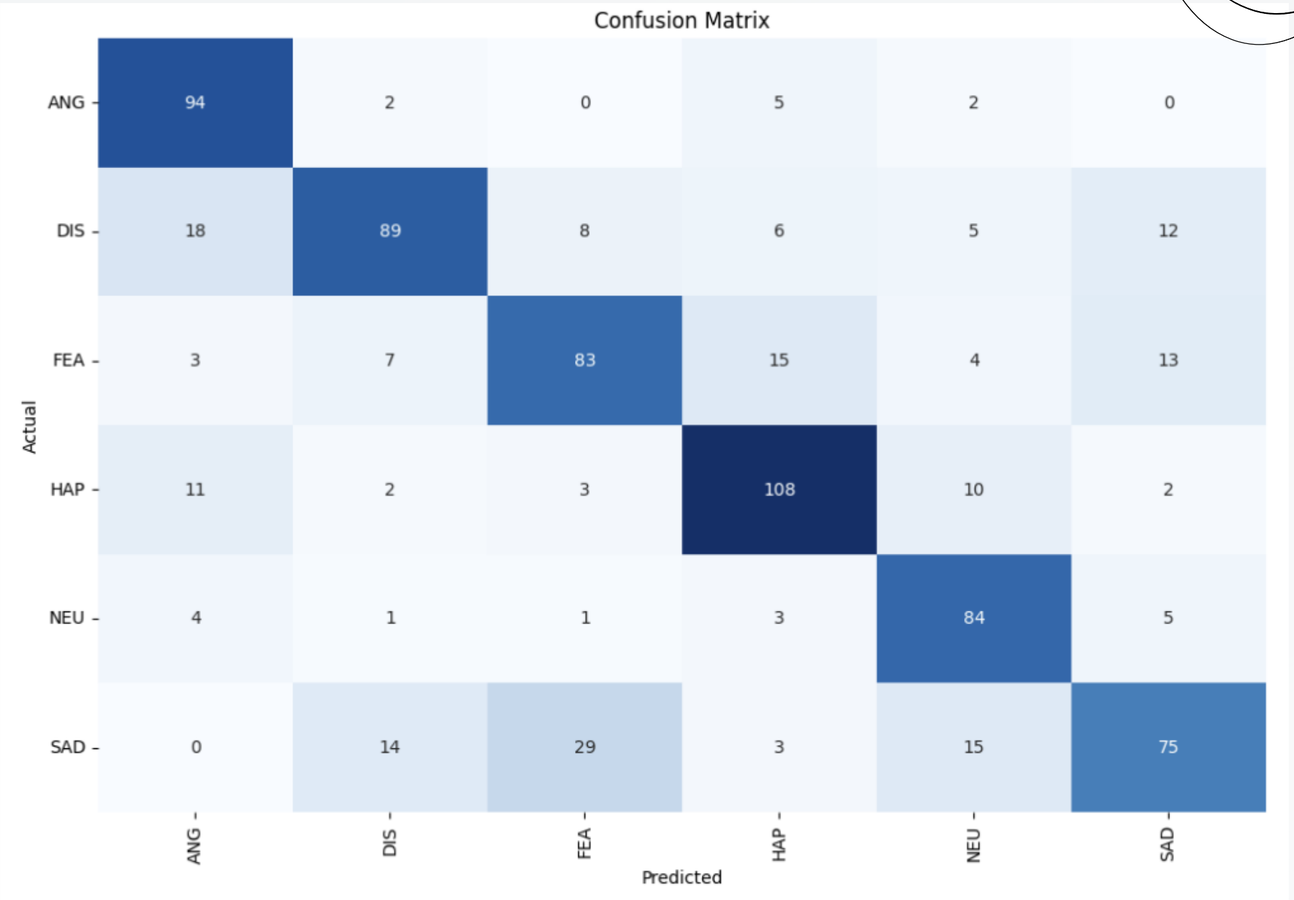


jonathanbgn.com

TFWAVE2VEC CLASSIFICATION RESULTS

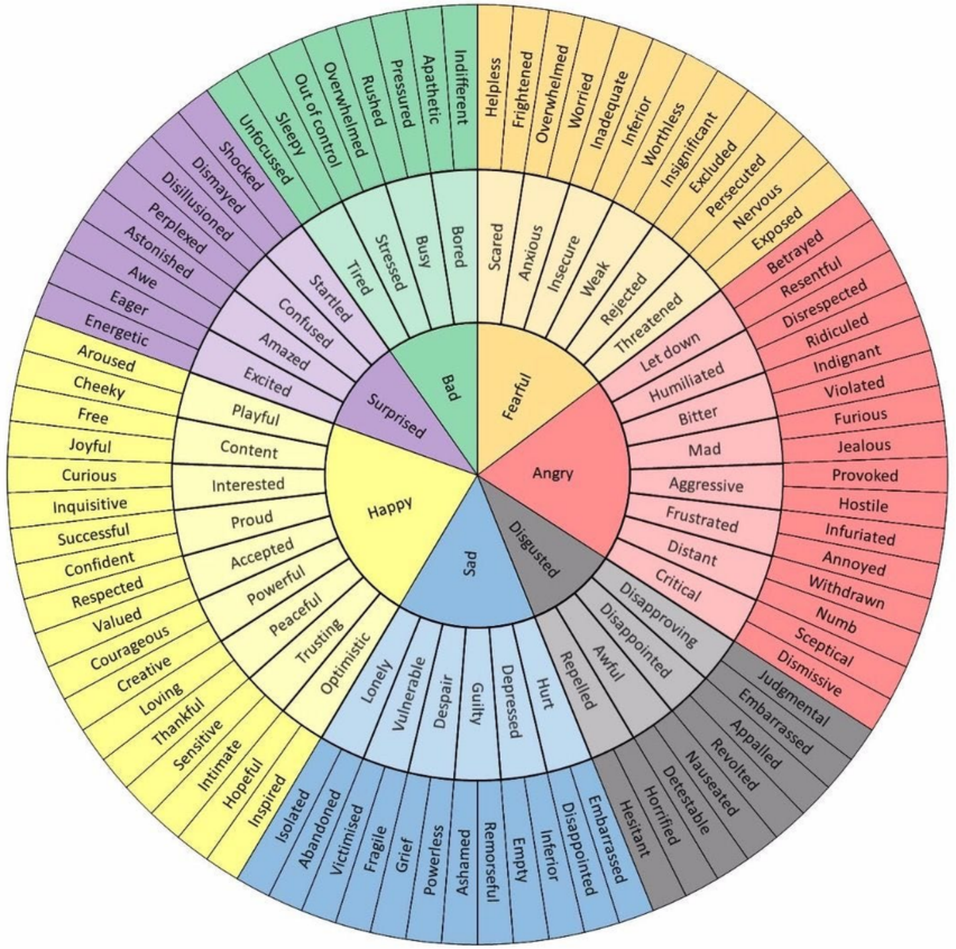
Classification Report:

	precision	recall	f1-score	support
ANG	0.723077	0.912621	0.806867	103
DIS	0.773913	0.644928	0.703557	138
FEA	0.669355	0.664000	0.666667	125
HAP	0.771429	0.794118	0.782609	136
NEU	0.700000	0.857143	0.770642	98
SAD	0.700935	0.551471	0.617284	136
accuracy			0.724185	736
macro avg	0.723118	0.737380	0.724604	736
weighted avg	0.725255	0.724185	0.719348	736



NEXT STEPS

- Instead of taking the aggregation of the extracted numeric features, try keeping all of the feature information from each window to feed into the classification model.
- Use other Hugging Face pretrained models like the Audio Spectrogram Transformer model in image classification.
- Use video clips to classify emotions.
- Use a different datasets or combination of emotion recognition datasets like the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS), Toronto emotional speech set (Tess), or Surrey Audio-Visual Expressed Emotion (Savee).



THANK YOU

biannag98@gmail.com
linkedin.com/in/bianna-gas/
github.com/biannagas