Infinity AI
Team – Mind Techies

Objective

Scope

Approach

Input:

1. Textual questionnaire
2. Audio file
3. Audio+video

Note : 1&2 is compulsory

1. **For audio - wav2vec2 and NN architechture using audio features from extracted from wav aduio files**
2. **For video – extracting inception v3 or vggnet features to NN based model**
3. **Baseline model – consuming both audio and video features (extracting inception v3 features)**

We contenated all answers from audio as one record and now send that audio based model. Meanwhile the videio itself was scrutinized to create features.

Ensemble model – taking avg of audio & video model, comparing thay against xgboost model and now taking avg of prediction probabilites from rule based .

Conclusion f1 score – 75%

```
Wav2Vec2ForSequenceClassification
```

1. **Wav2vec2**
   We obtained a base-line model using default training arguments, but since it was highly imabalanced ckass problem – as

   ```
   neutral 0.471466 joy 0.174509 surprise 0.120645 anger 0.111033 sadness
   0.068382 disgust 0.027133 fear 0.026832
   ```

2. **We observed that classfication report has 0 values for all metrics for all classes, Except for neutral. Rectification -**
   ------Underfitting
   1. **Adding stratify in train-test split**
   2. **Comput sample weight from sklear.utils**

   3. **Still it did now work, therfore used keras tuner Bayseain optimision**

3. **NN BASED MODEL USING DEFAULT AUDIO FEATURES**
4. **XGBOOST MODEL WITH DEFAULT AUDIO FEATURES**

I. **MFCC Features**
https://www.analyticsvidhya.com/blog/2021/06/mfcc-technique-for-speech-recognition/

Mel-frequency cepstral coefficients(MFCC):

In simpler terms, MFCCs are a set of coefficients that capture the shape of the power spectrum of a sound signal. They are derived by first transforming the raw audio signal into a frequency domain using a technique like the Discrete Fourier Transform (DFT), and then applying the mel-scale to approximate the human auditory perception of sound frequency. Finally, cepstral coefficients are computed from the mel-scaled spectrum.

MFCCs are particularly useful because they emphasize features of the audio signal that are important for human speech perception while discarding less relevant information. This makes them effective for tasks like speaker recognition, emotion detection, and speech-to-text conversion.

II. **Spectral features**

Spectral features in audio analysis refer to characteristics extracted from the frequency domain representation of a sound, like the distribution of energy across different frequencies, providing insights into the sound's timbre, pitch, and other auditory qualities.

III. **Pitch**

In the context of audio analysis, "spectral features" refer to characteristics derived from the frequency content of a signal, providing insights into its tonal qualities, timbre, and other acoustic properties.

Here's a breakdown of what spectral features entail:

Frequency Domain Representation:

Spectral features are obtained by transforming audio signals from the time domain to the frequency domain, typically using techniques like the Fast Fourier Transform (FFT).

Key Features:

- **Spectral Centroid:** Represents the "center of mass" of the spectrum, indicating the average frequency of the signal's energy.
- **Spectral Rolloff:** Measures the frequency below which a certain percentage (e.g., 85%) of the total spectral energy lies, indicating the presence of high-frequency content.
- **Spectral Bandwidth:** Quantifies the range of frequencies in the signal, reflecting the spread of energy across different frequencies.
- **Spectral Flatness:** Indicates the tonality of a sound, with lower values suggesting noisy sounds and higher values indicating voiced sounds.
- **Spectral Contrast:** Measures the difference between the peak and trough amplitudes of sound energy across the frequency spectrum.
- **Spectral Flux:** Quantifies the variation in the power spectrum of an audio signal across consecutive frames, useful for identifying onsets and abrupt changes.
- **Root Mean Square Energy (RMSE):** A measure of the energy of an audio signal, representing the average power over a specified window.
- **Mel-Frequency Cepstral Coefficients (MFCCs):** A widely used set of coefficients that represent the spectral characteristics of an audio signal, providing a detailed representation of its frequency content.

Output: Shall be based on ensemble model chosen.

Data:

Priority – 1) Audio 2) Questioniarre 3)

DAIC-WOZ Database

Brainstorming:

https://github.com/maelfabien/Multimodal-Emotion-Recognition

https://www.geeksforgeeks.org/wav2vec2-self-a-supervised-learning-technique-for-speech-representations/

https://www.analyticsvidhya.com/blog/2022/06/automatic-speech-recognition-using-wav2vec2/

https://huggingface.co/docs/transformers/v4.49.0/en/model_doc/wav2vec2#transformers.Wav2Vec2ForSequenceClassification

Future Scope