Major Project Report on

Emotion Recognition and Personality Assessment using Multimodal Data

submitted in partial fulfillment of the
requirements for the award of the degree of

Bachelor of Technology

in

Computer Science & Engineering

By:

Orendra Singh (00176807219 | CSE-3 | 2018[LE])
Jasmeet Singh(02376802718 | CSE-3 | 2018)
Medha (40676802718 | CSE-3 | 2018)
Harsh Shawait Singh (00376807218 | CSE-3 | 2018)

under the guidance of

Dr. Aashish Bhardwaj

(HOD,CSE)

Department of Computer Science & Engineering Guru Tegh Bahadur Institute of Technology(Affiliated to Guru Gobind Singh Indraprastha University Dwarka, New Delhi)

Class of Batch- 2018-2022

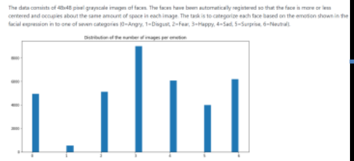
ABSTRACT

- The techniques devised under the domain of Artificial Intelligence, have diverse utilization and hence a huge impact in the process to simplify day-to-day activities, in general.
- One such problem, which has always been the basis of landmark research and multi-fold development amongst technical researchers, aims to identify human emotion and assess human behavior and personality, accurately, with an objective to support and enhance general human observation or even go beyond it to analyze certain characteristic traits for given multimodal data input.
- The project presents an effective analysis of three distinct media inputs aiming at recognition and classification facial expressions, analysis of human emotion , perceived through voice and retrieval of psychological traits from Real-Time visual, audial and textual input respectively, deployed through an interactive web application.
- Here accuracy has been used as a metric to evaluate the performance of the two mathematical models designed. The basis and design of each one, is rooted at Convolutional Neural Networks and LSTMs, widely used as a Deep Learning methodology to provide bench-mark results for Image Classification ,Speech Emotion Analysis and Sentiment Analysis for text.
- Further, Transfer Learning has been used as an additional technique to enhance the accuracy of result and present an improved model to preserve resources. The output has been analyzed and visualized through different functions and techniques as GRAD-CAM to verify the learning of the Neural Network for Emotion Recognition.

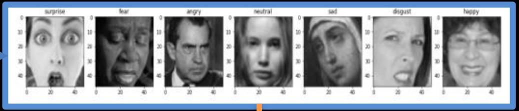
Methodology

Parameter	Facial Emotion Retrieval & Analysis	Vocal Emotion Retrieval & Analysis	Textual Emotion Retrieval & Analysis
Dataset	For the video data sets, we are using the popular FER2013 Kaggle Challenge data set. The data consists of 48x48 pixel grayscale images of faces. The faces have been automatically registered so that the face is more or less centered and occupies about the same amount of space in each image.	For audio data sets, we are using the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). This database contains 7356 files (total size: 24.8 GB). The database contains 24 professional actors (12 female, 12 male). Speech includes calm, happy, sad, angry, fearful, surprise, and disgust emotions.	For the text input, we are using the Stream-of-Consciousness dataset that was gathered in a study by Pennebaker and King [1999]. It consists of a total of 2,468 daily writing submissions from 34 psychology students (29 women and 5 men whose ages ranged from 18 to 67 with a mean of 26.4).
Deep Learning Approach	The model we have chosen is a Res-Net model as it outperformed the base model. The model utilizes the approach of Transfer Learning, where a pre-trained Convolutional Neural Network is utilized and the top-most layer is modified according to the dataset and classification categories. Also the attention features are visualized using GRAD_CAM techniques. Live prediction on outputs for each category has been discussed.	The model we have chosen is a Time Distributed Convolutional Neural Network. The main idea of a Time Distributed Convolutional Neural Network is to apply a rolling window (fixed size and time-step) all along the log-mel-spectrogram. Four Local Feature Learning Blocks (LFLBs) exist and the output of CNs are fed into RNN composed by 2 cells LSTM (Long Short Term Memory)	We have selected a Neural Network Architecture based on both one-dimensional Convolutional Neural Networks and Recurrent Neural Networks. The one-dimensional convolution layer plays a role comparable to feature extraction : it allows finding patterns in text data. The Long-Short Term Memory cell is then used in order to leverage on the sequential nature of natural language.

1.1 Video Analysis Pipeline



The data consists of 48x48 pixel grayscale images of faces. The faces have been automatically registered so that the face is more or less centered and occupies about the same amount of space in each image. The task is to categorize each face based on the emotion shown in the facial expression in to one of seven categories (0=Angry, 1=Disgust, 2=Fear, 3=Happy, 4=Sad, 5=Surprise, 6=Neutral).



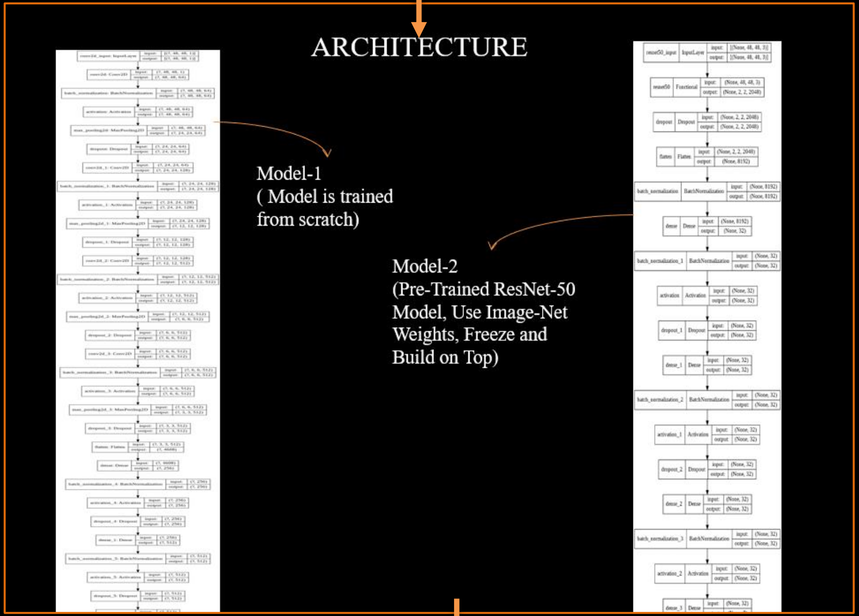
Minor Project Overview

- Explore the Dataset
- Generate Training and Validation Batches
- Create a Convolutional Neural Network (CNN) Model
- Train and Evaluate Model
- Save and Serialize Model as JSON String
- Create a Flask App to Serve Predictions
- Create a Class to Output Model Predictions
- Design an HTML Template for the Flask App
- Use Model to make a prediction for the Facial Expressions in each frame of the video as in the data.

Note:- (0=Angry, 1=Disgust, 2=Fear, 3=Happy, 4=Sad, 5=Surprise, 6=Neutral).

- Also identify the number of blinks on the facial landmarks on each picture
- Analyze the output for both the models for Accuracy and Loss, Confusion Matrix, and Top-K Accuracy.
- Visualize the Output by GRAD-CAM

Note:- We plot class activation maps, which display the pixels that have been activated by the last convolution layer. We notice how the pixels are being activated differently depending on the emotion being labeled. The happiness seems to depend on the pixels linked to the eyes and mouth, whereas the sadness or the anger seem for example to be more related to the eyebrows.

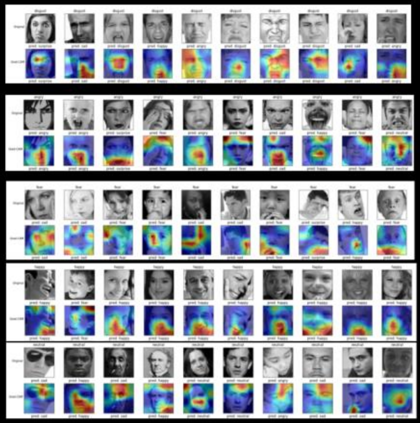


ARCHITECTURE

Model-1
(Model is trained from scratch)

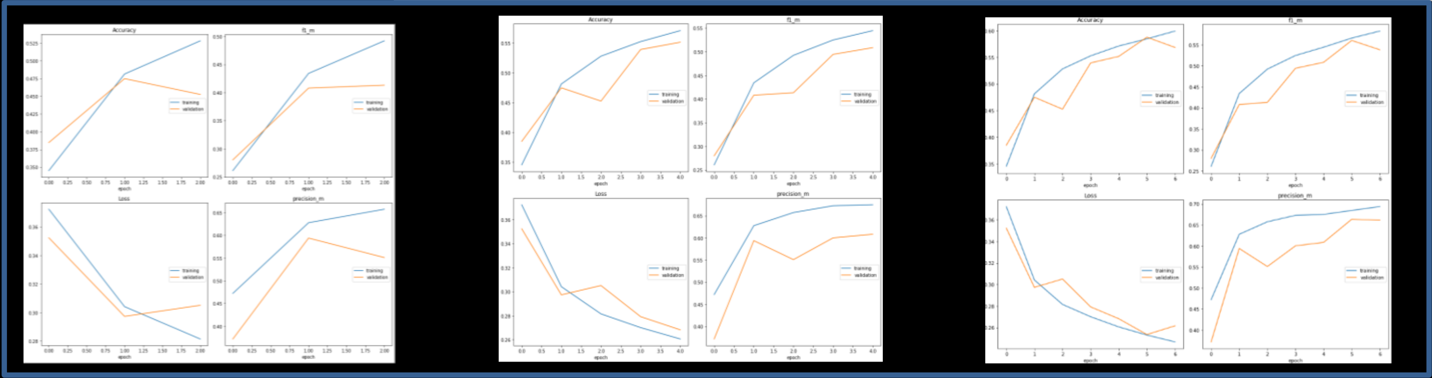
Model-2
(Pre-Trained ResNet-50 Model, Use Image-Net Weights, Freeze and Build on Top)

GRAD-CAM Visualizations

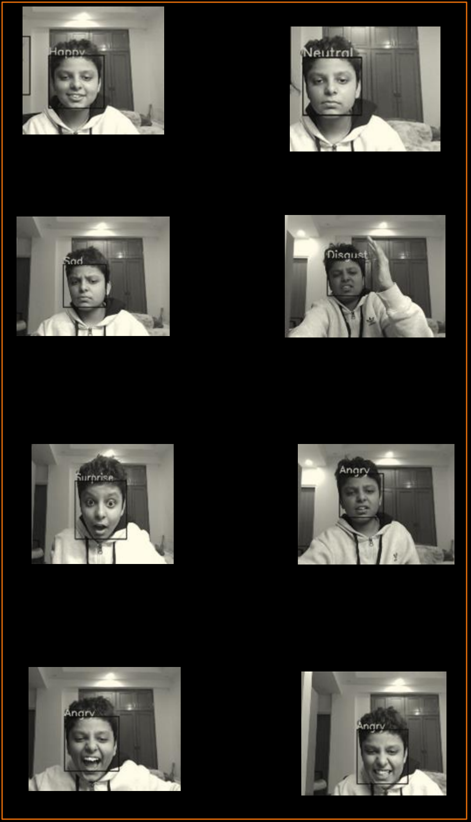


1. 2 Real-Time Facial Emotion Prediction

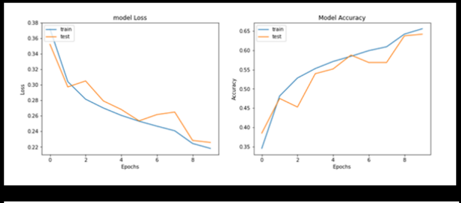
(Minor Project Report)



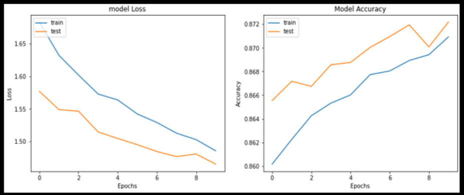
Real-Time Output Predictions



Accuracy and Loss-Comparison

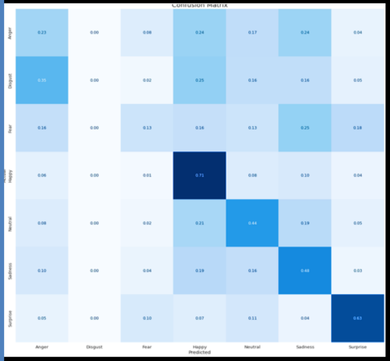


Model-1
(Model is trained from scratch)



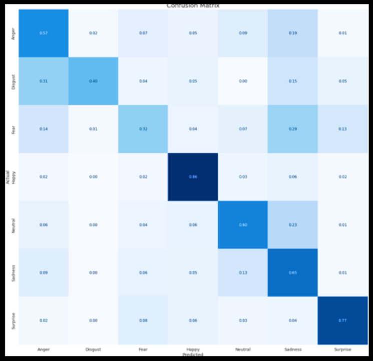
Model-2
(Pre-Trained ResNet-50 Model, Use Image-Net Weights, Freeze and Build on Top)

Confusion Matrix



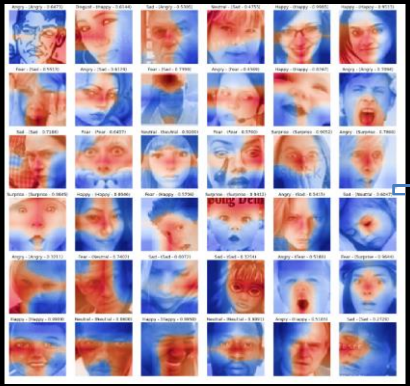
Confusion Matrix

Actual \ Predicted	Anger	Disgust	Fear	Happy	Neutral	Sadness	Surprise
Anger	0.23	0.00	0.08	0.24	0.17	0.24	0.04
Disgust	0.35	0.00	0.02	0.25	0.16	0.16	0.05
Fear	0.16	0.00	0.13	0.16	0.13	0.25	0.18
Happy	0.06	0.00	0.01	0.71	0.08	0.10	0.04
Neutral	0.08	0.00	0.02	0.21	0.44	0.19	0.05
Sadness	0.10	0.00	0.04	0.19	0.16	0.48	0.03
Surprise	0.05	0.00	0.10	0.07	0.11	0.04	0.63

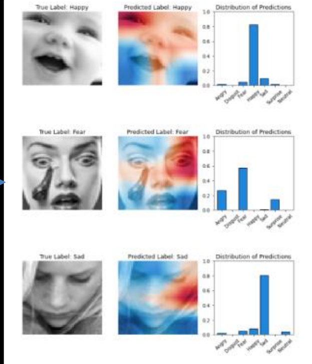


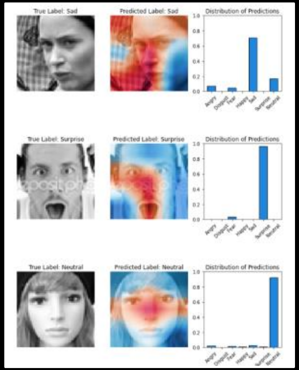
Confusion Matrix

Actual \ Predicted	Anger	Disgust	Fear	Happy	Neutral	Sadness	Surprise
Anger	0.57	0.02	0.07	0.05	0.09	0.19	0.01
Disgust	0.31	0.40	0.04	0.05	0.00	0.15	0.05
Fear	0.14	0.01	0.32	0.04	0.07	0.29	0.13
Happy	0.02	0.00	0.02	0.86	0.03	0.06	0.02
Neutral	0.06	0.00	0.04	0.06	0.60	0.23	0.01
Sadness	0.09	0.00	0.06	0.05	0.13	0.65	0.01
Surprise	0.02	0.00	0.08	0.06	0.03	0.04	0.77

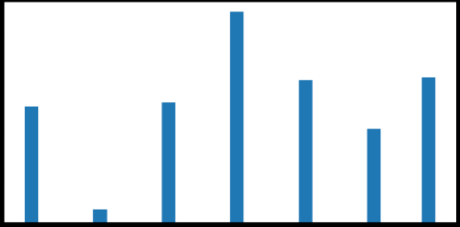


GRAD-CAM Visualisation





Data Distribution



2.1 Audio Analysis Pipeline

The speech emotion recognition pipeline was built the following way :

- Voice recording
- Audio signal discretization
- Log-mel-spectrogram extraction
- Split spectrogram using a rolling window
- Make a prediction using our pre-trained model

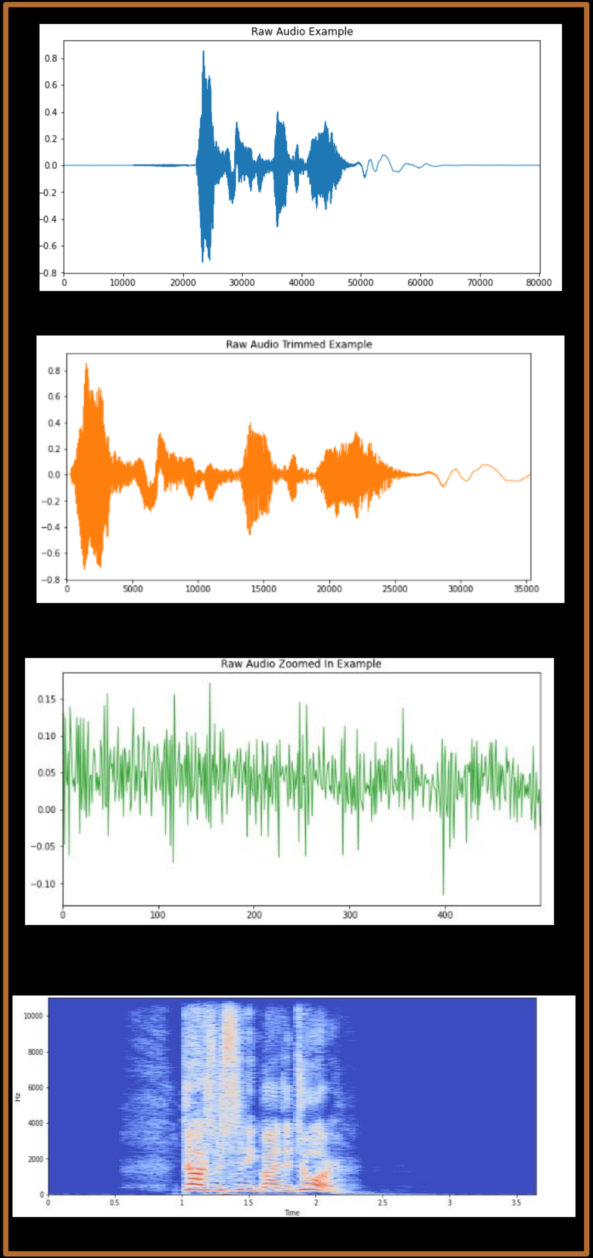
Dataset –

- This portion of the RAVDESS contains 1440 files: 60 trials per actor x 24 actors = 1440. The RAVDESS contains 24 professional actors (12 female, 12 male), vocalizing two lexically-matched statements in a neutral North American accent.
- Speech emotions includes calm, happy, sad, angry, fearful, surprise, and disgust expressions. Each expression is produced at two levels of emotional intensity (normal, strong), with an additional neutral expression.
- File naming convention is specific to the dataset.
- Each of the 1440 files has a unique filename. The filename consists of a 7-part numerical identifier (e.g., 03-01-06-01-02-01-12.wav).

The model we have chosen is a Time Distributed Convolutional Neural Network.

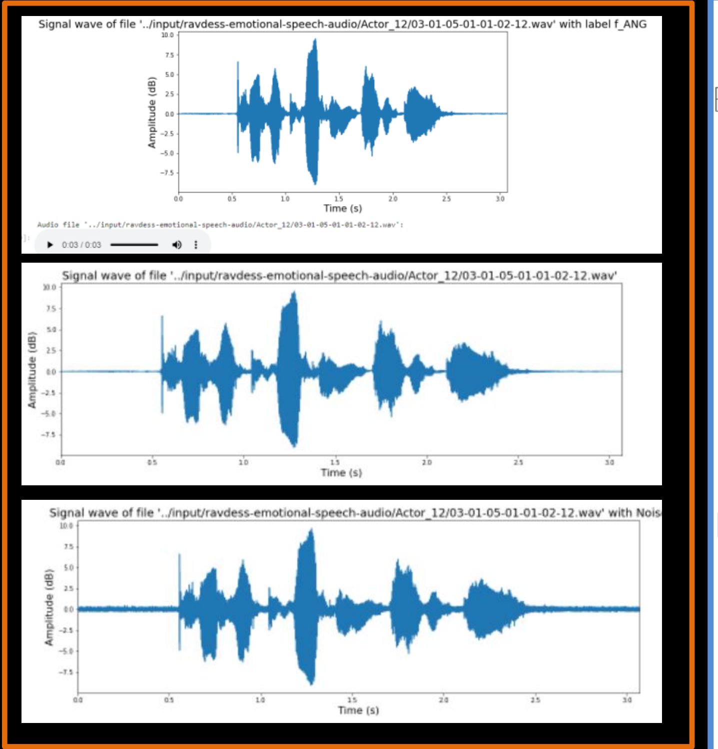
The main idea of a Time Distributed Convolutional Neural Network is to apply a rolling window (fixed size and time-step) all along the log-mel-spectrogram. Each of these windows will be the entry of a convolutional neural network, composed by four Local Feature Learning Blocks (LFLBs) and the output of each of these convolutional networks will be fed into a recurrent neural network composed by 2 cells LSTM (Long Short Term Memory) to learn the long-term contextual dependencies. Finally, a fully connected layer with softmax activation is used to predict the emotion detected in the voice.

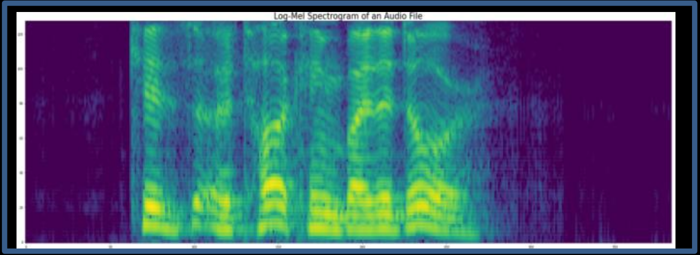
Data Visualisation



2.2 Vocal Emotion Prediction

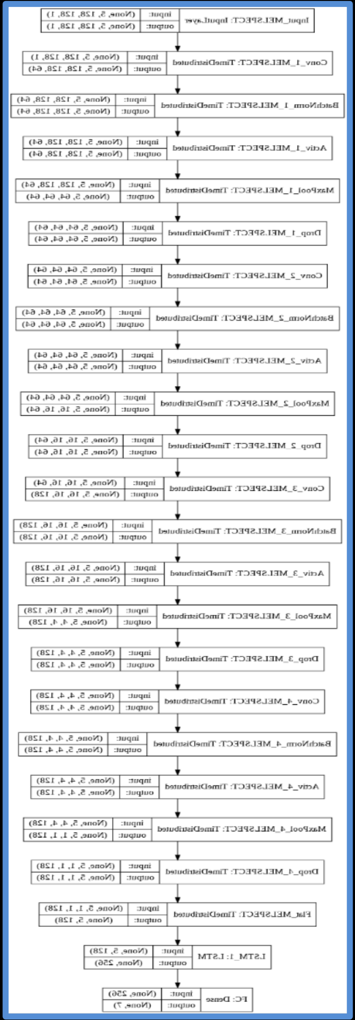
Data Augmentation



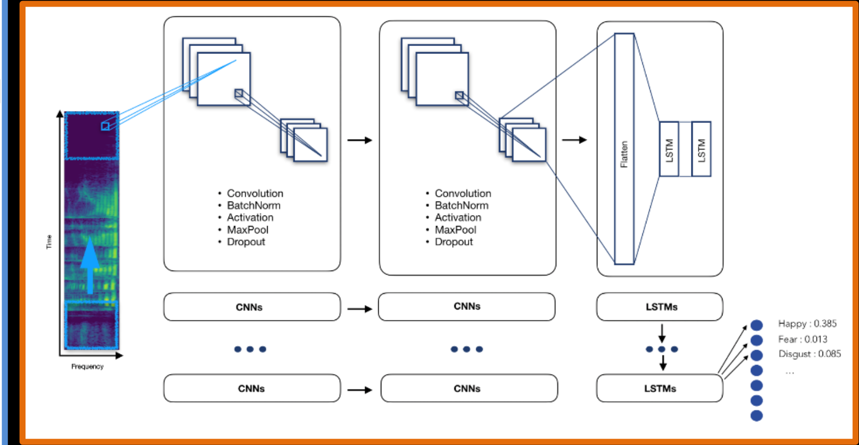


Log-Mel Spectogram Visualisation

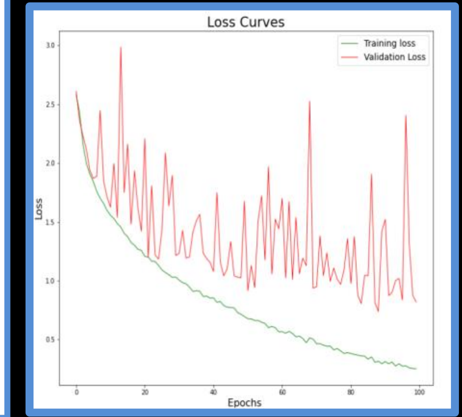
Model Design

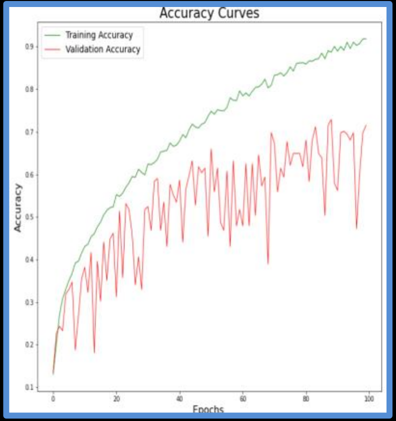

Input_MEL_SPECT: InputLayer | input: (None, 5, 128, 128, 1) | output: (None, 5, 128, 128, 1)
Conv_1_MEL_SPECT: TimeDistributed | input: (None, 5, 128, 128, 1) | output: (None, 5, 128, 128, 64)
BatchNorm_1_MEL_SPECT: TimeDistributed | input: (None, 5, 128, 128, 64) | output: (None, 5, 128, 128, 64)
Activ_1_MEL_SPECT: TimeDistributed | input: (None, 5, 128, 128, 64) | output: (None, 5, 128, 128, 64)
MaxPool_1_MEL_SPECT: TimeDistributed | input: (None, 5, 128, 128, 64) | output: (None, 5, 64, 64, 64)
Drop_1_MEL_SPECT: TimeDistributed | input: (None, 5, 64, 64, 64) | output: (None, 5, 64, 64, 64)
Conv_2_MEL_SPECT: TimeDistributed | input: (None, 5, 64, 64, 64) | output: (None, 5, 64, 64, 64)
BatchNorm_2_MEL_SPECT: TimeDistributed | input: (None, 5, 64, 64, 64) | output: (None, 5, 64, 64, 64)
Activ_2_MEL_SPECT: TimeDistributed | input: (None, 5, 64, 64, 64) | output: (None, 5, 64, 64, 64)
MaxPool_2_MEL_SPECT: TimeDistributed | input: (None, 5, 64, 64, 64) | output: (None, 5, 16, 16, 64)
Drop_2_MEL_SPECT: TimeDistributed | input: (None, 5, 16, 16, 64) | output: (None, 5, 16, 16, 64)
Conv_3_MEL_SPECT: TimeDistributed | input: (None, 5, 16, 16, 64) | output: (None, 5, 16, 16, 128)
BatchNorm_3_MEL_SPECT: TimeDistributed | input: (None, 5, 16, 16, 128) | output: (None, 5, 16, 16, 128)
Activ_3_MEL_SPECT: TimeDistributed | input: (None, 5, 16, 16, 128) | output: (None, 5, 16, 16, 128)
MaxPool_3_MEL_SPECT: TimeDistributed | input: (None, 5, 16, 16, 128) | output: (None, 5, 4, 4, 128)
Drop_3_MEL_SPECT: TimeDistributed | input: (None, 5, 4, 4, 128) | output: (None, 5, 4, 4, 128)
Conv_4_MEL_SPECT: TimeDistributed | input: (None, 5, 4, 4, 128) | output: (None, 5, 4, 4, 128)
BatchNorm_4_MEL_SPECT: TimeDistributed | input: (None, 5, 4, 4, 128) | output: (None, 5, 4, 4, 128)
Activ_4_MEL_SPECT: TimeDistributed | input: (None, 5, 4, 4, 128) | output: (None, 5, 4, 4, 128)
MaxPool_4_MEL_SPECT: TimeDistributed | input: (None, 5, 4, 4, 128) | output: (None, 5, 1, 1, 128)
Drop_4_MEL_SPECT: TimeDistributed | input: (None, 5, 1, 1, 128) | output: (None, 5, 1, 1, 128)
Flat_MEL_SPECT: TimeDistributed | input: (None, 5, 1, 1, 128) | output: (None, 5, 128)
LSTM_1: LSTM | input: (None, 5, 128) | output: (None, 256)
FC: Dense | input: (None, 256) | output: (None, 7)

Compiled Model Visualisation



Model Prediction Analysis





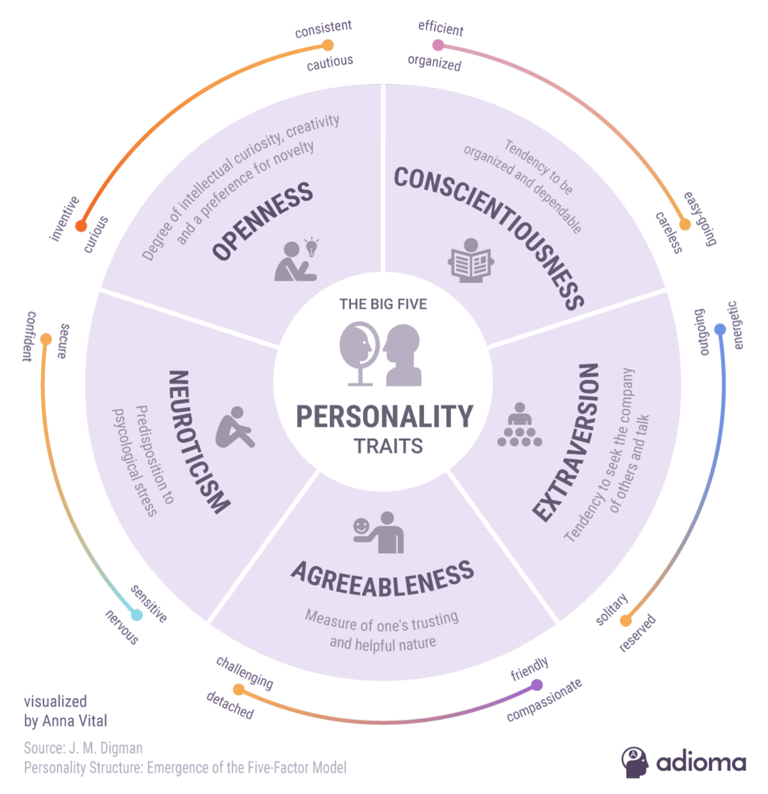
3.1 Personality Assessment

The main motivation behind this choice is to offer a broader assessment to the user : as emotions can only be understood in the light of a person's own characteristics, our approach is that analyzing personality traits would provide a new key to understanding the nature of an individual for the job .
The text-based personality recognition pipeline has the following structure :
Text data retrieving

- Custom natural language preprocessing :
- Tokenization of the document
- Cleaning and standardization of formulations using regular expressions (for instance replacing "can't" by "cannot", "'ve" by "have")
- Deletion of the punctuation
- Lowercasing the tokens
- Removal of predefined stopwords (such as 'a', 'an' etc.)
- Application of part-of-speech tags on the remaining tokens
- Lemmatization of tokens using part-of-speech tags for more accuracy.
- Padding the sequences of tokens of each document to constrain the shape of the input vectors. The input size has been fixed to 30 : all tokens beyond this index are deleted. If the input vector has less than 30 tokens, zeros are added at the beginning of the vector in order to normalize the shape. The dimension of the padded sequence has been determine using the characteristics of our training data. The average number of words in each essay was 652 before any preprocessing. After the standardization of formulations, and the removal of punctuation characters and stopwords, the average number of words dropped to 29 with a standard deviation of 0.5. In order to make sure we incorporate in our classification the right number of words without discarding too much information, we set the padding dimension to 30, which is roughly equal to the average length plus two times the standard deviation.
- 300-dimension Word2Vec trainable embedding
- Prediction using our pre-trained model

3.2 Big-Five Personality Traits Stream-of-consciousness

Many psychology researchers (starting with D. W. Fiske [1949], then Norman [1963] and Goldberg [1981]), believe that it is possible to exhibit five categories, or core factors, that determine one's personality. The acronym OCEAN (for openness, conscientiousness, extraversion, agreeableness, and neuroticism) is often used to refer to this model.



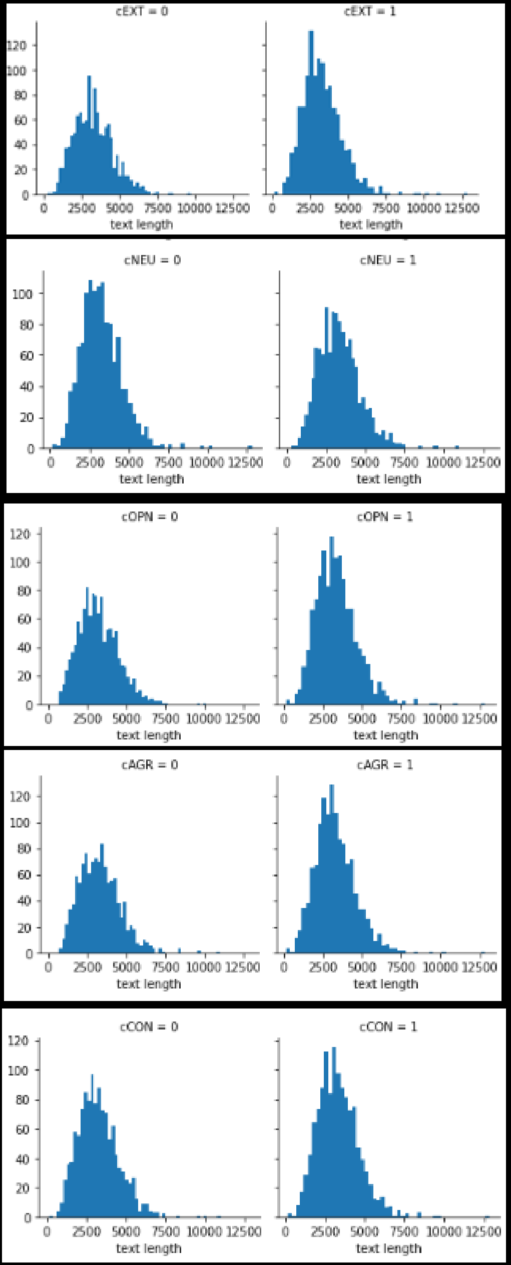
visualized by Anna Vital

Source: J. M. Digman
Personality Structure: Emergence of the Five-Factor Model

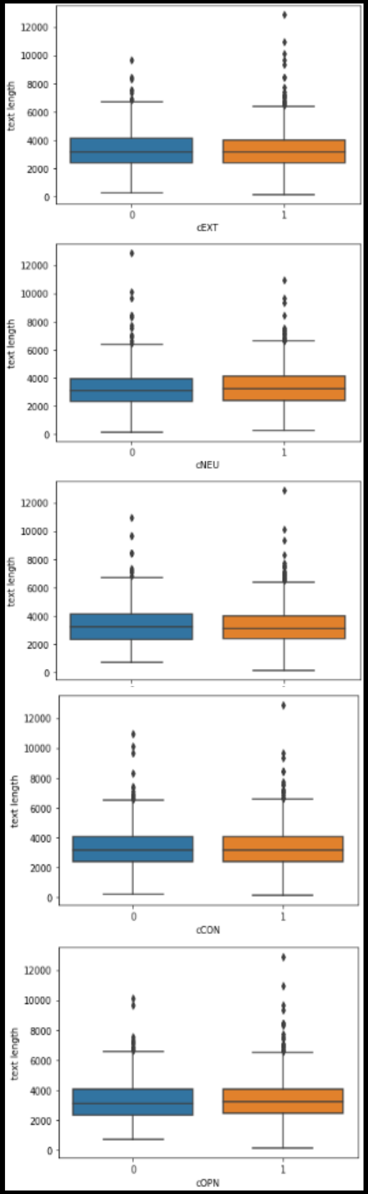
- The writing submissions were in the form of a course unrated assignment.
- For each assignment, students were expected to write a minimum of 20 minutes per day about a specific topic.
- The data was collected during a 2-week summer course between 1993 to 1996. Each student completed their daily writing for 10 consecutive days.
- Students’ personality scores were assessed by answering the Big Five Inventory (BFI) [John et al., 1991]. The BFI is a 44-item self-report questionnaire that provides a score for each of the five personality traits.
- Each item consists of short phrases and is rated using a 5-point scale that ranges from 1 (disagree strongly) to 5 (agree strongly).
- An instance in the data source consists of an ID, the actual essay, and five classification labels of the Big Five personality traits.
- Labels were originally in the form of either yes (‘y’) or no (‘n’) to indicate scoring high or low for a given trait.
- While the five dimensions don't capture the peculiarity of everyone's personality, it is the theoretical framework most recognized by researchers and practitioners in this field.

3.3 Text Analysis Pipeline

Histograms of Text length Distribution for teach Label



Boxplots of text length Distribution for each Label



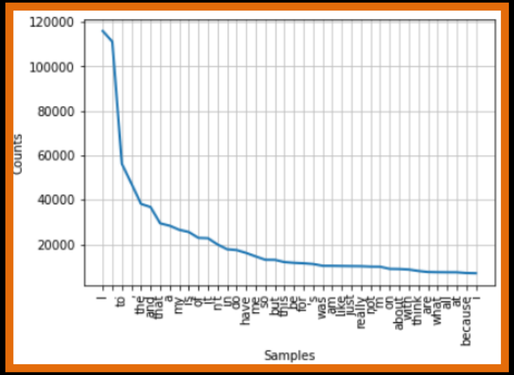
Before Pre-Processing

- The total number of essays is 2467
- The total number of words in all essays is 1608813
- The average number of words in each essay is 652.1333603567085

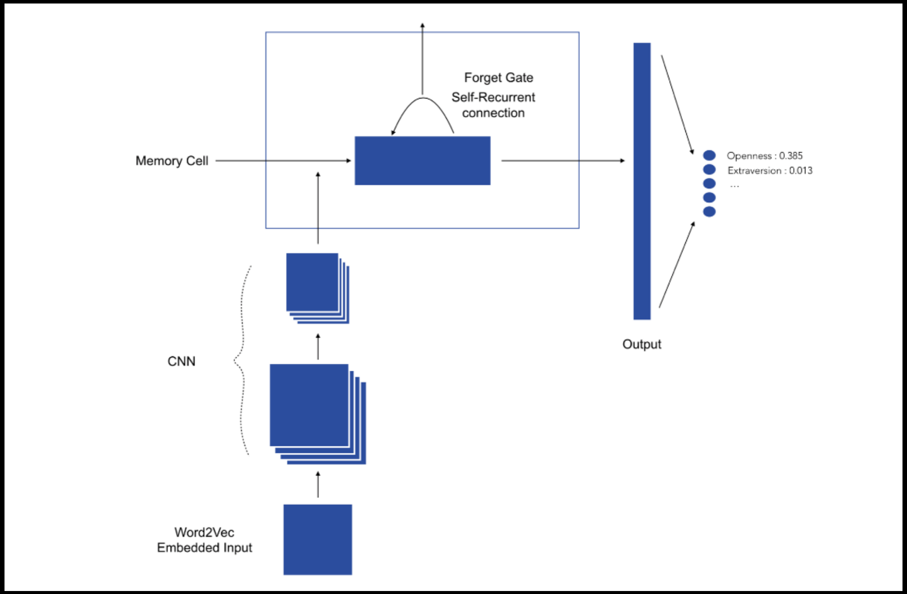
After Pre-Processing

- The average number of words in each preprocessed essay is 29.98256992298338
- The standard deviation of the number of words in each preprocessed essay is 0.5281882006533262
- The average number of words in each preprocessed essay plus 2 standard deviations is 31.038946324290034

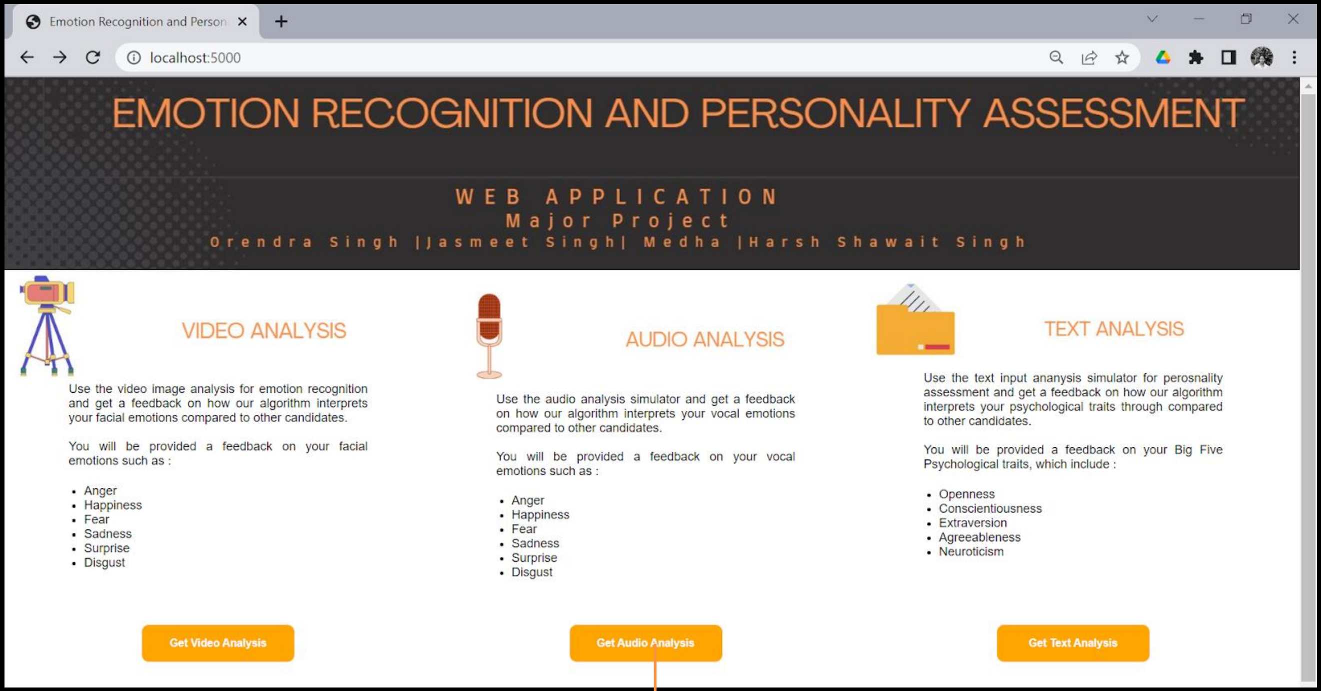
List of 100 most frequent words/counts



Model Design



WEB APPLICATION USER INTERFACE



EMOTION RECOGNITION AND PERSONALITY ASSESSMENT

WEB APPLICATION
Major Project
Orendra Singh |Jasmeet Singh| Medha |Harsh Shawait Singh

VIDEO ANALYSIS

Use the video image analysis for emotion recognition and get a feedback on how our algorithm interprets your facial emotions compared to other candidates.

You will be provided a feedback on your facial emotions such as :

- Anger
- Happiness
- Fear
- Sadness
- Surprise
- Disgust

AUDIO ANALYSIS

Use the audio analysis simulator and get a feedback on how our algorithm interprets your vocal emotions compared to other candidates.

You will be provided a feedback on your vocal emotions such as :

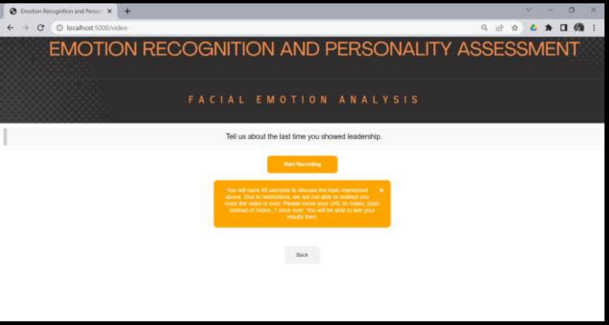
- Anger
- Happiness
- Fear
- Sadness
- Surprise
- Disgust

TEXT ANALYSIS

Use the text input ananysis simulator for perosnality assessment and get a feedback on how our algorithm interprets your psychological traits through compared to other candidates.

You will be provided a feedback on your Big Five Psychological traits, which include :

- Openness
- Conscientiousness
- Extraversion
- Agreeableness
- Neuroticism

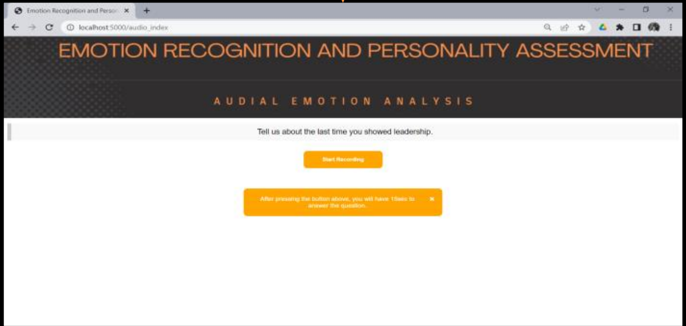


EMOTION RECOGNITION AND PERSONALITY ASSESSMENT

FACIAL EMOTION ANALYSIS

Tell us about the last time you showed leadership.

You will have 45 seconds to discuss the topic mentioned above. Due to restrictions, we are not able to redirect you once the video is over. Please move your URL to /video_dash instead of /video_1 once over. You will be able to see your results then.

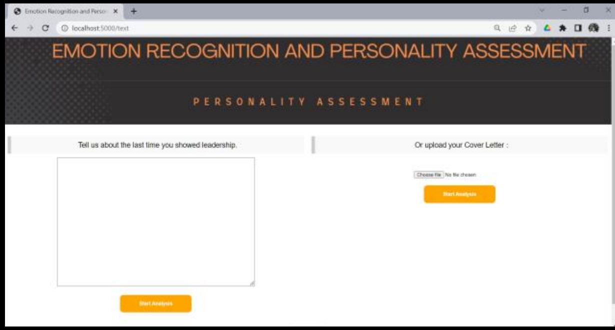


EMOTION RECOGNITION AND PERSONALITY ASSESSMENT

AUDIAL EMOTION ANALYSIS

Tell us about the last time you showed leadership.

After pressing the button above, you will have 15sec to answer the question.



EMOTION RECOGNITION AND PERSONALITY ASSESSMENT

PERSONALITY ASSESSMENT

Tell us about the last time you showed leadership.

Or upload your Cover Letter :

Facial Expression Analysis

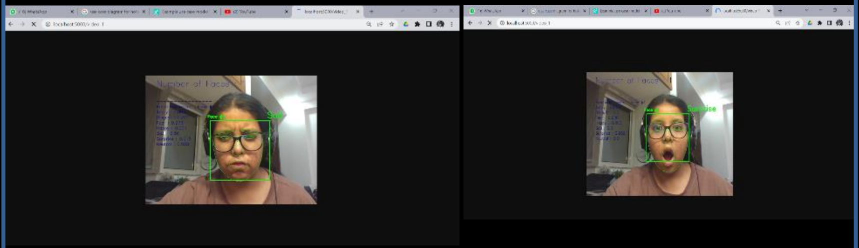
Audial Analysis

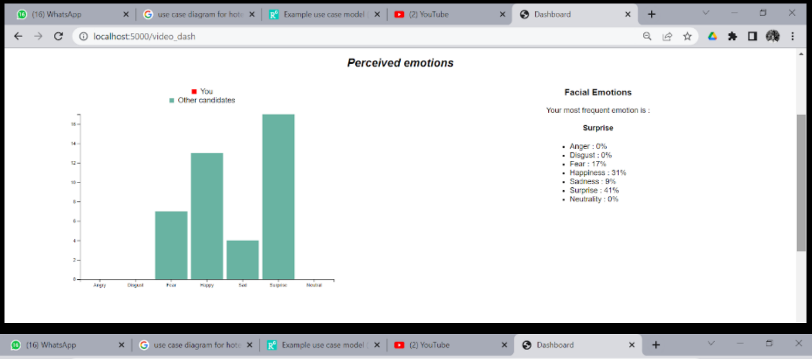
Textual Analysis

- **The interface finds a wide variety of applications. A major area of rising necessity is the need to recognize emotion and personality during online interviews.**
- **It also helps each candidate analyse the behaviour and emotion conveyed before-hand. The application can be used to conduct re-schedulable interviews.**
- **In some context, it may surpass erroneous human judgement to analyse one's personality,**

Dashboard

Facial Expression Analysis





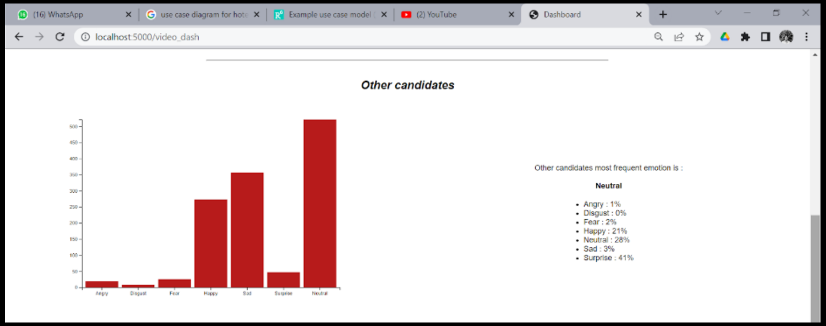
Perceived emotions

Facial Emotions

Your most frequent emotion is :

Surprise

- Anger : 0%
- Disgust : 0%
- Fear : 17%
- Happiness : 31%
- Sadness : 9%
- Surprise : 41%
- Neutrality : 0%

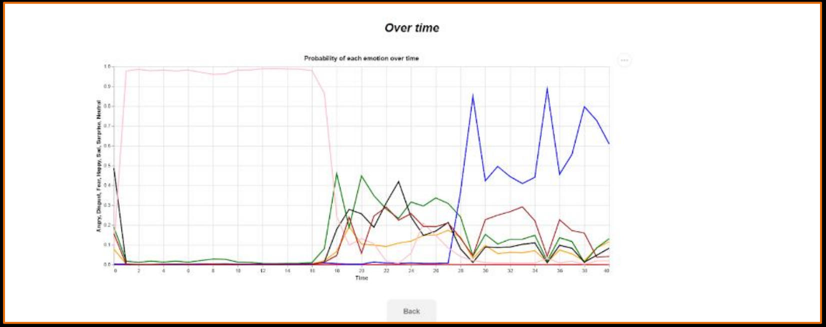


Other candidates

Other candidates most frequent emotion is :

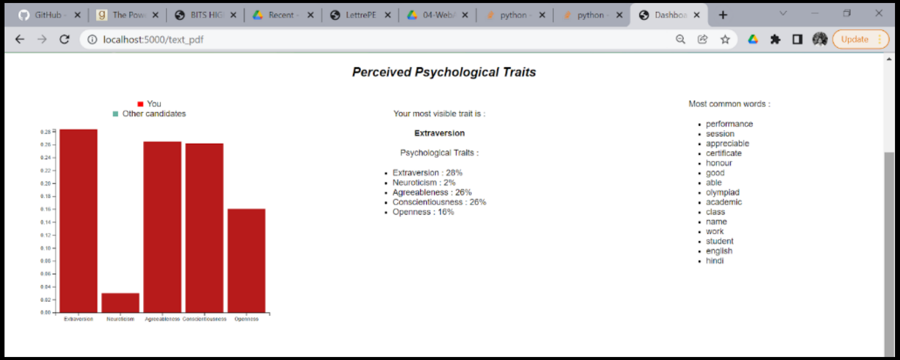
Neutral

- Angry : 1%
- Disgust : 0%
- Fear : 2%
- Happy : 21%
- Neutral : 28%
- Sad : 3%
- Surprise : 41%



Over time

Textual Analysis



Perceived Psychological Traits

Your most visible trait is :

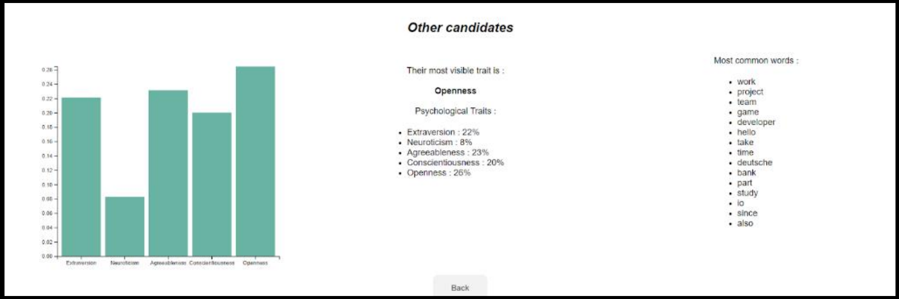
Extraversion

Psychological Traits :

- Extraversion : 28%
- Neuroticism : 2%
- Agreeableness : 26%
- Conscientiousness : 26%
- Openness : 16%

Most common words :

- performance
- session
- appreciable
- certificate
- honour
- good
- able
- olympiad
- academic
- class
- name
- work
- student
- english
- hindi



Other candidates

Their most visible trait is :

Openness

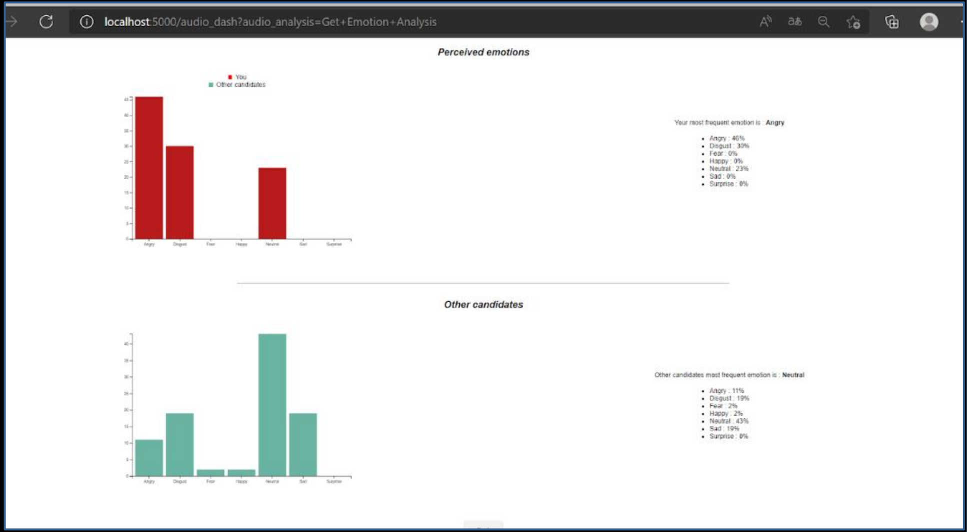
Psychological Traits :

- Extraversion : 22%
- Neuroticism : 8%
- Agreeableness : 23%
- Conscientiousness : 20%
- Openness : 26%

Most common words :

- work
- project
- team
- game
- developer
- hello
- take
- time
- deutsche
- bank
- part
- study
- io
- since
- also

Audial Analysis


localhost:5000/audio_dash?audio_analysis=Get+Emotion+Analysis
You
Other candidates
Angry
Disgust
Fear
Happy
Neutral
Sad
Surprise

Perceived emotions

Your most frequent emotion is : **Angry**

- Angry : 46%
- Disgust : 30%
- Fear : 0%
- Happy : 0%
- Neutral : 23%
- Sad : 0%
- Surprise : 0%

Other candidates

Other candidates most frequent emotion is : **Neutral**

- Angry : 11%
- Disgust : 19%
- Fear : 2%
- Happy : 2%
- Neutral : 43%
- Sad : 19%
- Surprise : 0%

Conclusion and Scope

- This project was a combination of CNN model classification problem where we have to identify and recognize human emotion and get a measure of distinct psychological traits playing a key-factor for personality analysis using Multimodal Data. CNNs and RNNs provide a great scope of improvement in design for better results.
- Additional datasets can be used to train the model for better analysis of facial, vocal and textual input. Other important parameters which can be improved to enhance accuracy of the result are as Data Augmentation, Hyper Parameter Tuning, Early Stopping and Activation Functions.
- The web application user interface can be removed by designing an ensemble model to extract three distinct features from only one stream of input rather than taking a distinct input for each mode of information extraction.
- Additional features can be introduced at frontend to enhance the overall user experience.

THANKS