MULTIMODAL SENTIMENT ANALYSIS

A
Report submitted
in the partial fulfilment of the requirements for the award of the degree of

Bachelor of Technology
in
Information Technology

BY:

BHAVYA SHARMA (1805213018)
HIMANI JAYAS (1805213023)
KOKIL GUPTA (1805213027)

Under the guidance of

PROF. MANIK CHANDRA
MR. ABHISHEK SINGH

Department of Computer Science and Engineering
Institute of Engineering and Technology, Lucknow
Dr. A.P.J. Abdul Kalam Technical University, Lucknow, Uttar Pradesh

TABLE OF CONTENTS

DECLARATION

We hereby declare that this submission is our own work and that, to the best of our belief and knowledge, it contains no material previously published or written by another person or material which to a substantial error has been accepted for the award of any degree or diploma of university or other institute of higher learning, except where the acknowledgement has been made in the text. The project has not been submitted by us at any other institute for the requirement of any other degree.

Date: 21st May 2022

Submitted by: -

(1) Name: Bhavya Sharma

Roll No.: 1805213018

Branch: IT

Signature: 

(2) Name: Himani Jayas

Roll No.: 1805213023

Branch: IT

Signature: 

(3) Name: Kokil Gupta

Roll No.: 1805213027

Branch: IT

Signature: 

CERTIFICATE

This is to certify that the project report entitled "Multimodal Sentiment Analysis" presented by Bhavya Sharma, Himani Jayas and Kokil Gupta in the partial fulfillment for the award of Bachelor of Technology in Information Technology, is a record of work carried out by them under my supervision and guidance at the Department of Computer Science and Engineering at Institute of Engineering and Technology, Lucknow.

It is also certified that this project has not been submitted at any other Institute for the award of any other degrees to the best of my knowledge.

01- June-2022.

(Mr. Abhishek Singh)

Department of CSE

IET, Lucknow

01-06-2022

(Dr. Manik Chandra)

Department of CSE

IET, Lucknow

ACKNOWLEDGEMENT

I am highly indebted to Dr. Manik Chandra and Mr. Abhishek Singh, and I want to thank them for giving us the freedom to operate and experiment with new ideas. I want to make a move to our significant thanks to them for their educational direction and advantage in our task and steady help combined with certainty boosting and propelling meetings that demonstrated extremely productive and were instrumental in injecting confidence and trust inside us. The sustaining and blooming of the current work is primarily because of their significant direction, ideas, adroit judgment, productive analysis, and an eye for flawlessness. Our mentor consistently addressed a horde of our questions with grinning thoughtfulness and enormous tolerance. They never make us feel like we're on our backsides by constantly listening to our perspectives, respecting and developing them, and allowing us a free hand in our project. It is simply because of their staggering interest and accommodating disposition; the current work has achieved its stage. Finally, I am grateful to our Institution and colleagues whose constant encouragement served to renew our spirit, refocus our attention and energy, and carry out this work.

Bhavya Sharma

Kokil Gupta

Himani Jayas

ABSTRACT

Sentiment Analysis intends to naturally reveal the hidden mentality that we hold towards an entity. The total of this assumption over a populace addresses sentiment surveying and has various uses. At present text-based sentiment analysis depends on the development of word embeddings and Machine Learning models that take in conclusion via enormous text collection. Text based Sentiment Analysis is presently generally utilized as consumer loyalty appraisal and brand insight investigation. When the online media expanded, multimodal assessment investigation is going to carry new freedoms with the appearance of integral information streams for upgrading and going past text-based feeling examination using the new transforms methods. Multimodal investigation offers good roads for vocal articulations notwithstanding the printed or record content and this compelling follows supposition that distinguishes it. Recurrent Neural Networks (RNNs) along with the Long-Short Term Memory modes are the methodologies that are used to increase the performance. In multimodal examination, we characterize issues and the feeling in advancements in ongoing audits and investigation of multimodal assessment which generally includes video websites, human-human connections, pictures, human-machine and spoken surveys. Multimodal feeling investigation helps us in promoting our theory which holds the undiscovered critical potential and the arising field is examined for challenges and difficulties.

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

Sentiment Analysis opens up various freedoms relating to web-based media to under- stand clients' inclinations, propensities, and substance.

1.1 BACKGROUND INFORMATION

Multimodal Sentiment Analysis is another dimension of the customary text-based assessment investigation, which goes past the test of writings, and incorporates different modalities like sound and visual information. The sentiment is evoked when an individual experiences a particular topic, person, or element. Understanding individuals' position, disposition, or assessment towards a specific feature has numerous applications. The text-based feeling investigation has been the leading figure around here and, as of late, has examined different modalities, like audio and vision, started to be thought of in use.

Liu and Zhang [1] characterized sentiment analysis as an issue of automatic detection of four segments of a notion including, entity, viewpoint, entity holder, viewpoint's feeling. A good sentiment analysis framework ought to have the option to disengage this load of four segments accurately.

A new improvement in multimodal sentiment analysis is visual assumption investigation. Web-based media clients regularly share instant messages with pictures/recordings, and these visible sights and sounds are extra direct data in communicating client notions. Mid-level visual supposition portrayals are one valuable development for separating feeling and elements in text-based notion investigation.

Recordings give multimodal information as far as vocal and visual modalities. The vocal balances and looks in the visual information, alongside text information, give significant prompts better to recognize genuine emotional conditions of the assessment holder. Consequently, a mix of text and video information assists with making a better feeling and assumption examination model.

1.2 MOTIVATION

Understanding emotion using text became so common throughout the years. Thus, introducing other models like audio is necessary and provides a broad domain in sentiment analysis. We will be doing the Text Analysis by using LSTM and Bidirectional LSTM [8]. Audio data will be used to create Spectrograms or MFCC's using the Librosa library, which can predict the label using the spectrograms images with the CNN network or the MFCC values combined with the classification model. Multimodal Sentiment Analysis can be used in chatbots, call centers that can tell the customers' satisfaction after talking to a bot or even an employee.

1.3 PROJECT OBJECTIVE

To create a Multimodal Sentiment Analysis that will extract the sentiments using the three modes, i.e., Audio, Text, and Video.

Evaluating the datasets by checking the loss and accuracy of our model.

1.4 OUR CONTRIBUTION

Text Sentiment Analysis is done by filtering the dataset, like reducing every word to its stem and passing the corpus through models with LSTM and Attention Layers to predict the sentiment.

Audio Sentiment Analysis is done by taking the Real-time Audio of a user and then calculating MFCC's to predict the user's emotion [7].

Video Sentiment Analysis is the simple use of detecting human facial expressions in real-time video using some Transfer Learning Techniques.

A local server website will be created to deploy all these three modes in one.

1.5 REPORT LAYOUT

- Literature Review describes all the previous works done in this field.
- Methodology describes the architecture of the project.
- Implementation Details describes the tools that are used and the process that needs to be followed in each mode.
- Results shows the final outputs that are done in the course of this project.
- Conclusion specifies the limitations and future scope of this project.

2. LITERATURE REVIEW

Multimodal Sentiment Analysis is a dynamic theme in Natural Language Processing (NLP). It automatically removes individuals' perspectives or emotional states from numerous correspondence channels [6] (e.g., text, voice, and facial expressions). Furthermore, it has different applications. The center test displays the complex intra-modular and between modular cooperations, where multimodal highlights are being intertwined. Yanan Jia and Sony SungChu proposed the idea of Multimodal Sentiment Analysis in which they used two modes, i.e., Audio and Text for Sentiment Analysis; we here will add another method, i.e., of video mode that will use facial expressions.

2.1 Text Sentiment

As an aim to extract evaluative meaning, an alternative to topic detection in the field of sentiment analysis was started.

Maybe the most encouraging improvement in text sentiment is due to the use of deep learning. Deep learning can leverage massive scope datasets to register word embeddings that are relevant for feeling examination, delivering naturally extended lexical. While the derivation of word classes dependent on deep learning [13] strategies is accomplishing results exceptionally near those of human annotators, ongoing work found that extrapolating word sentiment consistent factors dependent on word embeddings still requires significant work. Profound Recurrent Neural Networks have been applied to the errand of subjectivity detection, and word vector representations can join administered and unaided learning when applied to feeling analysis.

For Text Sentiment the authors of implemented SVM in their research, but I thought of using different techniques, and so we will work with Bidirectional LSTM's with the Attention Mechanism. Though specialists have stretched out LSTM cells and doors to learn fleeting collaboration designs among multimodal successions and also Pham-et-al proposed consideration-based RNNs [9] to learn multimodal portrayals with a cyclic interpretation misfortune among modalities. Still, we give a chance to a Bidirectional LSTM that will help us upbeat these mechanisms significantly.

2.2 Audio Sentiment

Notwithstanding, targeting opinion unequivocally solely from spoken expressions is an equivalently youthful field. Zeroing in on the acoustic side of communication in language, the line among opinion and feeling investigation is regularly extremely frail, as, e. g. In Mairesseet al. zero in on pitch-related provisions and saw that addition- ally, without text-based signals [3], pitch contains data on feeling. Various further works center around feeling examination solely from the text-based substance as present in the discourse. For example, Costa Pereira et al's proposed approach takes a verbally expressed inquiry and recovers reports whose conclusions look like the question. Likewise, Pérez-Rosas and Mihalcea focus on the semantics of spoken audits in the wake of utilizing discourse acknowledgment. Kaushik et al.and its extension observe that feeling examination on normal unconstrained discourse information can be acknowledged in any event when confronted with low word acknowledgment rates — a pattern that has been seen additionally in the acknowledgment of valence from an unconstrained discourse by Metze et al.

The Audio Sentiment implement by authors of used KNN for their purpose. We will be calculating the MFCC's for carrying out our work in the Audio Field. Sequence models can be fitted dependent on channel banks, MFCCs, or any other low-level descriptors removed from crude discourse without highlight designing. In any case, this methodology, for the most part, requires exceptionally effective calculation and huge explained sound records. It used an audio dataset with the meantime for calls to be 4 seconds for the sentiment analysis in audio. Still, we will try to increase its mean to >7 seconds to check its progress for large audios as they haven't explored that region [10].

Zadeh et al. planned a multiview gated memory unit that neural organizations constrain. It stores furthermore, predicts fleeting cross-modular collaborations. Tsai et al. used transformer consideration systems to learn both cross-modular arrangements furthermore, collaborations. Albeit neural organizations extraordinarily work on the presentation over conventional techniques, and their unpredictable engineering genuinely influences the model interpretability.

2.3 Video Sentiment

While there have been connected lines of examination in vision-based emotion [2] acknowledgment for quite a while, e.g., directing sentiment investigation by computer vision is a somewhat ongoing region of research. The chief examination undertakings in "visual opinion analysis" spin around displaying, distinguishing, and utilizing sentiment expressed through facial or accurate signals or feeling associated with visual sight and sound.

Among the soonest work in visual opinion examination, Wanget al. investigated descriptor affiliations coordinated into 12 adjective–modifier word sets more than 100 pictures commented on by 42subjects. They utilized an assortment of shading high- lights, including lightness, immersion, and sharpness highlights related to support vector relapse to anticipate the presence of these sets like warm-cool, brilliant–gloomy, and vibrant–desolate

Every one of these work in promoting and applying visual feeling examination highlight the potential in the higher precision methods, as with CNNs, just as expanded inclu- sion, as with multilingual [12] and different substance source methods. Furthermore, with the expanding number of freely accessible PC vision models/libraries and visual feeling datasets, visual opinion examination is ready to see development in both of these bearings. The complex idea of feeling shows that visual feeling investigation alone can not wholly gauge and additionally portray our experiential attitude and sentiments in interactive media information. For instance, visible substance probably won't have the option to comprehend the unique circumstance or concentrate the element.

In Video Sentiment we will be using simple face expressions to identify sentiments like Happy, Angry, Disgust, etc. As of late, neural network techniques are well known to demonstrate the perplexing interaction between images. The authors of improved methods for Faster CNN [11], which are well known as to be Transfer Learning Techniques.

3. METHODOLOGY

The methodology is a relevant structure for research. We have multiple sections that cover the Architecture, Working, and Tools Used in the project.

3.1 System Architecture

The accompanying address the System Architecture and essential working of the Web Application of the Sentiment Analysis. The system is designed to provide sentiments using text, audio, and video. The ends of the system consist of a list of emotions that can be predicted using any of the three models with probabilities assigned to each one of them individually.

3.2 Word Embeddings: [17] The text data to be predicted for sentimental analysis is provided to the Word Embeddings module, which is equipped for catching the setting of a word in an archive, semantic and syntactic likeness, connection with different words.

3.3 MFCC's [18]: In solid preparation, the mel-recurrence cepstrum (MFC) is a depiction which shows momentary force span of a sound in light of a direct cosine change of a log power range on a nonlinear mel size of recurrence.

Mel-recurrence cepstral coefficients (MFCCs) are basically some coefficients that aggregately make up an MFC. They can be obtained from a kind of cepstral display of the brief snippet (a nonlinear "range-of-a-range").

The difference between the mel-recurrence cepstrum and cepstrum is that the recurrence groups are separated similarly on mel scale that resembles the reaction of body hearable frameworks more closely than the recurrence groups that are divided straight utilized in conventional range. For example, in sound pressure the distortion in recurrence can take into account improved portrayal of sound.

MFCCs are decided by means of following:

- For a sign, take Fourier transform.
- Guide the forces of the range got above onto the mel scale, utilizing three-sided covering windows or, on the other hand, cosine surrounding windows.
- For each mel frequencies make a log of the forces.
- Of the rundown of log powers of mel, calculate the discrete cosine, assuming it is anything but a sign.
- MFCCs are amplitudes of the succeeding range.

3.4 Transformers [14]: The figure given below depicts transformers and which is also called a sequence-to-sequence architecture. Sequence-to-Sequence architecture is a neural network which changes a specified succession of components, like grouping words in a sentence, into

another grouping. These models are admissible for interpretation, in which the grouping in words from one language is changed to a series of different words in some other dialect.

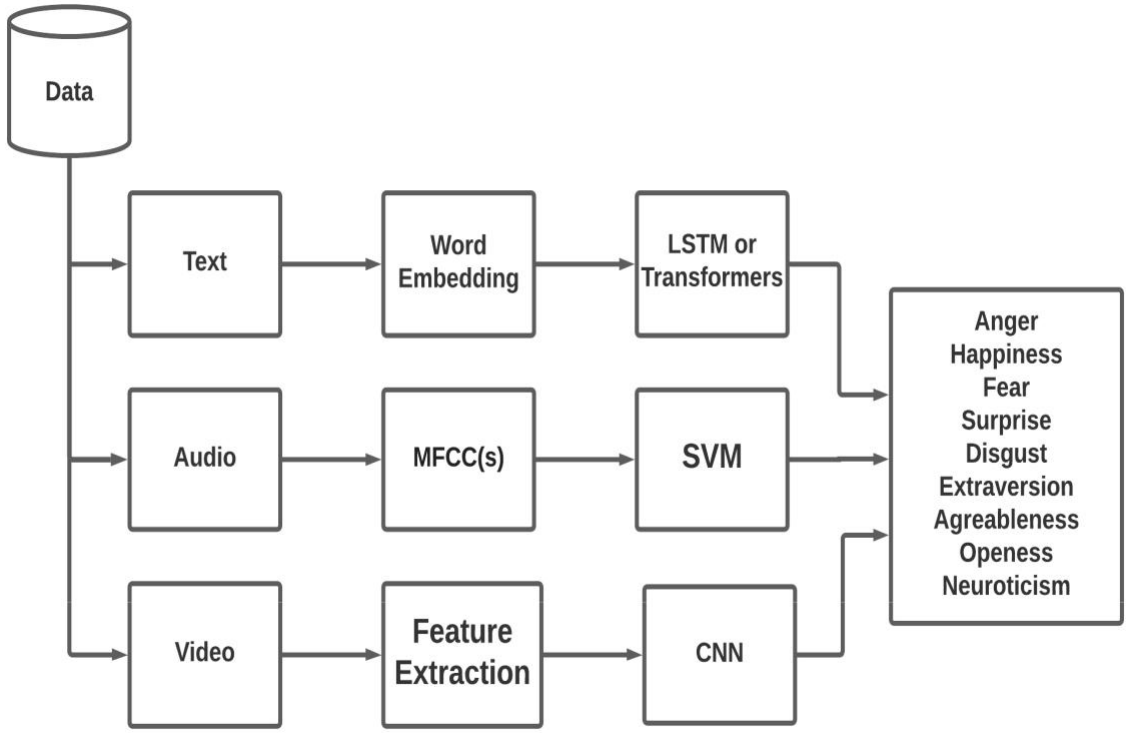

Fig 1: System Architecture

Above figure proposes the System Architecture of our project that deals with the three modes of data, i.e., Text, audio and video.

The walkthrough of the Architecture is as follows:

- The Text Data is cleaned and pre-processed. This Bag Of Words or Embedding Matrix is created to send it to the LSTM model that will predict the label or the maximum probability of sentiment in the text.
- The Audio Data is cleaned and pre-processed. Using this audio, we calculated the Spectrograms or MFCC, gave it to Neural Networks od Classification models, respectively, and predicted the label accordingly.

- The Video Data is cleaned and pre-processed as discussed in 3.2.2. After this, landmark points are extracted, which then is used by the Transfer Learning Techniques to predict the label.

4. PROPOSED MODEL

Our point is to foster a model ready to furnish live sentiment with a visual UI utilizing Tensorflow and Js innovation. Consequently, we have chosen to isolate three kinds of information sources:

1. Textual Information: It has been developed to interview an individual that will help us determine the Personality Traits of the individual. We can also get these using a cover letter of an individual and analyze them accordingly.

2. Audio Information: It has been developed to take audio input of about 15 sec and visualize the sentiments like Angry, Happy, Disgust, Sad and Neutral over the period. This can be used in customer satisfaction detection after the call gets ended in the Call Centers.

3. Video Information: It will take an individual's live video feed and help us identify the sentiment in a live form using a webcam.

4.1 DATASET SOURCES

4.1.1 Text: For the text input, we are using data which was gathered in a study by King and Pennebaker [19]. It has 2,468 daily writing submissions given by 34 psychology [5] scholars (five men and 29 women from 18 to 67 years of age).

4.1.2 Audio: For sound informational collections, we are using the "Ryerson Audio-Visual Database of Emotional Speech and Song". RAVDESS contains 7356 voice clips (size: 24.8 GB). These records contain 24 audio clips (12 females, 12 guys), showing two lexically coordinated explanations in a nonbiased North American speech. Discourse incorporates quiet, glad, miserable, sore, unfortunate, shock, and repugnance articulations, and the tune contains quiet, cheerful, dismal, sad feelings.

4.1.3 Video: For the video informational collections, we are utilizing the well-known FER2013 Kaggle Challenge informational index. The information comprises 48x48 pixel grayscale pictures of countenances. The informational collection remains very testing to use since there are vacant pictures or wrongly ordered pictures.

4.2 DATA PRE-PROCESSING

This comprises two different variety of data namely Audio and Video. We will discuss the pre-processing of all the data formats.

4.2.1 Text Pre-processing:

The pre-processing is the initial step of our NLP pipeline. This is the place where we convert crude content records to cleaned arrangements of words. To finish this interaction, we first

need to tokenize the corpus. This implies that sentences are parted into a rundown of single words, likewise called tokens. Other pre-processing steps remember using standard articulations for a request to erase undesirable characters or reformat comments. At last, there are strategies accessible to supplant words by their linguistic root: the objective of both stemming and lemmatization is to decrease derivationally related types of a comment to a typical base structure.

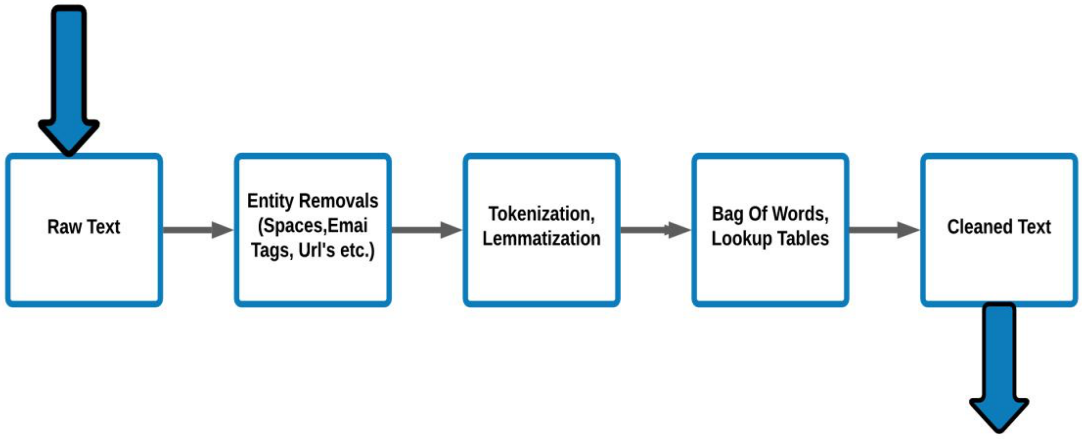

Fig 2: Text Cleaning Pipeline

Fig explains the Text Cleaning Pipeline and how the text is converted to its basic stem and fed to the model for the training and testing purposes.

4.2.2 Audio Pre-Processing:

To begin with, before starting feature extractions, it's fitting to apply a pre-emphasis filter on the sound sign to intensify every one of the significant frequencies. After the pre-emphasis filter, we need to part the sound sign into transient windows called frames. We duplicate each case by a Hamming window work in the wake of parting the movement into different casings. It permits decreasing spectral spillage or any sign discontinuities and working on signal lucidity.

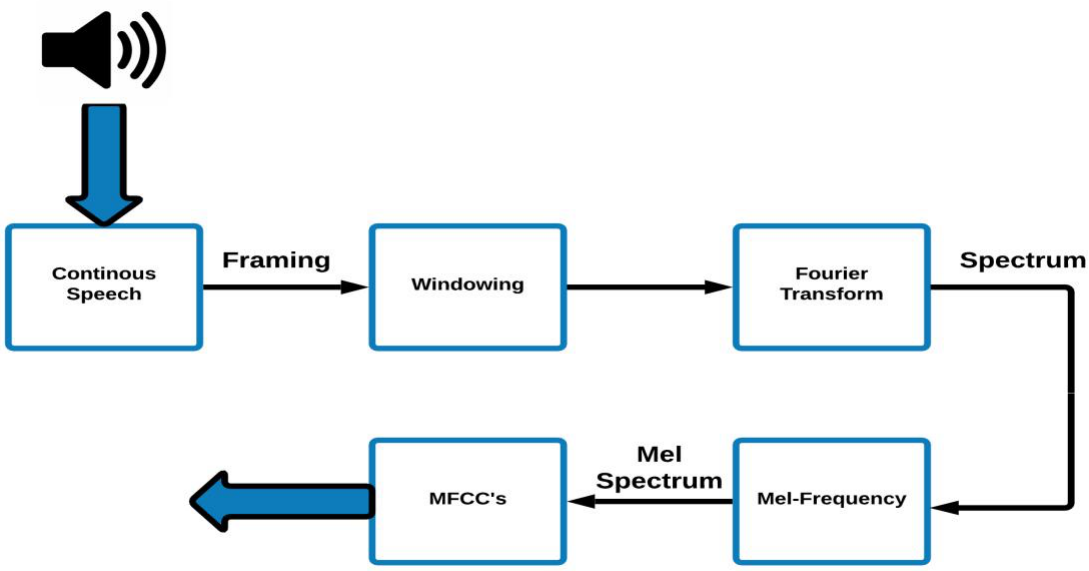

Fig 3: Audio Pre-processing

Fig explains the Audio Cleaning and conversion of those into the MFCC's that will be used as the input for the model and is used for training and testing purposes.

4.2.3 Video Pre-Processing:

Starting by analyzing the video frame by frame, then applying filters using some of the convolution techniques and making fewer inputs to identify the face then and adequately zoom on it, reducing pixel density to the same pixel density as that of the train set. Getting landmarks points is a part of feature extraction that is processed during this stage. We are transforming the input image to a model readable input to predict the emotion of the information.

5. IMPLEMENTATION DETAILS

5.1 LANGUAGES AND LIBRARIES

- OpenCv [22]: OpenCv is a ML package library and associates ASCII text file laptop vision. It's a library of programming functions chiefly geared toward period laptop vision. We have used this mostly in my video phase, where most of the work of face detection and augmentation is done by this library.

- HTML: HTML or HyperText Markup Language is a markup language that allows web users to create and structure various parts of a web page such as headers, tables and links using elements, tags, and attributes. It tends to be helped by innovations like Cascading Style Sheets (CSS) and prearranging dialects like JavaScript [23]. This would be based on the language for my website that I would create to deploy my three models of Sentiment Analysis.

- NumPy [24] : It is a open source Python library. NumPy works with Python objects called multi-dimensional arrays. Arrays are basically collections of values, and they have one or more dimensions. NumPy array data structure is also called *ndarray*, short for n-dimensional array. Datasets are usually built as matrices and it is much easier to open those with NumPy instead of working with lists. Numpy here is used for many processes. This library does all the mathematical computation in my project.

- Tensorflow [25]: Tensorflow is an open source framework. It was initially designed to be a neural network library but with advancement it can perform much more functions. It is a machine learning library. It is the base library that we have used to create our model; this is the cover of Keras that helped in model designing and fitting the values.

5.2 SETUP USED

- Flask [26]: It is a small net framework written in Python. It's classified as a microframework. As a result of it doesn't need explicit tools or libraries. It's no information abstraction layer, type validation, or the other parts wherever pre-existing third- party libraries give standard functions. It has been used to create an interface between the website and models and also is responsible for returning the HTML pages accordingly to the output.

- Google Colab [27]: Colab is a free Jupyter notebook climate that runs altogether in the cloud. In particular, it doesn't need an arrangement, and the notebooks that you make can be at the same time altered by your colleagues - how you vary reports in Google Docs. The free GPU of Colab is used in this project for the training purpose of

the Video and Audio modes for providing fast results.

- Spyder [28]: It is a free and open-source logical climate written in Python and planned by and for researchers, specialists, and information examiners. It includes a remarkable mix of an exhaustive advancement instrument's high-level altering, examination, troubleshooting, and profiling usefulness. I have used this to work on all my python files, and all the mathematical work is done over here.

- VS Code [29]: VS Code is a lightweight text editor, one of the best for coding in all most all languages. It provides you to code in any programming language for example Python, Java, C++, JavaScript, and more. Visual Studio Code is a source code editor, which helps businesses build and debug web applications running on Windows, Linux, and macOS. It is a source-*code* editor text editor program design.

5.3 TEXT SENTIMENT

Text modal used the Pennebaker and King dataset for Text Sentiment Analysis that usually predicts the Personality Traits that we will use to check over an individual that can be used in an interview process. Sentiment Analysis is always a difficult task as the machine cannot understand humor, anger, happiness, and sadness. Day by day, NLP is growing, and we are getting many models that are improving and solving this problem. Initially, RNN models were used, but the problem was that it could not see the future data as a word by word inputs were given to the model. Thus, new models came up like the LSTM's, Bidirectional LSTM's, and Transformers. I used Bidirectional-LSTM's in the process that helped to improve the accuracy and decision by the model.

The steps that we will go through this module are:

1. First of all, the text is cleaned, and unnecessary words are removed using the Tokenization method, and all symbols are removed, and the whole text is made in lower-case.

2. Then we will create a Bag Of Words that will contain the vocabulary size,i.e., most of the words used in the data.

3. Embedding Matrix is created which is the strong relationship of words that are nearby like King and Queen, or Apple and Mango are strongly related.

4. This embedding matrix data is put as an input to the Attention Based Model that we will custom create with Bidirectional LSTM Encoders, Attention Layer, and the Decoders.

5. Many to One LSTM's are used to predict the label using the text.

We implemented Text Analysis using the text-box and were also given an option of uploading the Cover-Letter that can also be used to predict the individual's Personality Traits.

5.4 AUDIO SENTIMENT

Audio modal used the RAVDESS data for the Audio Sentiment Analysis. It uses 15- second audio provided by the user in the portal; the runtime is less for less computational work as training and handling the audio in small chunks is a significant improvement for the predictions. Literature is centered on just around six feelings,i.e., happy, sad, angry, disgusted, fear, and surprise. Albeit the feeling classifications are more plentiful and complex, in actuality.

The steps that we went through this module were:

1. Extract 15 seconds audio and add some noise to the data so that model can also be used in the real-life process.

2. Signal Pre-processing will be done in the next stage,like amplifying high-frequency and splitting audio in frames.

3. After all this MFCC'is calculated, which are the input data that will be used for the model.

4. Classification models can be used to predict one of the six labels of sentiment.

5. Printing a bar plot of the sentiments achieved by using Argmax computation.

5.5 VIDEO SENTIMENT

The work that is done on the Facial Expressions has been trained over FER2013 Kaggle Challenge dataset and has obtained a good accuracy while using the Xception transfer learning model.

1. First of all, the video is split into frames, and the analysis is done step by step.

2. Filters are used after getting the frames, and Convolution Operations are per- formed.

3. Features Extraction is done, and landmark points are located in those frames.

4. The image is flattened and fed to the Exception Model for an output.

5.6 DEPLOYMENT

The project's primary purpose is to have a place where we can test all the capabilities of a project. This deployment is the last stage that will help us to do this work.

I created a website on the local server that will run all the three modules of our project, i. e., Audio, Text, and Video. All the three models that we have created during the training time will be used up by Flask which will help us to run our project on Local Network so that all the dependencies can be used in one go.

All the steps discussed here were implemented and below are the results of that implementation with the final local web server.

6. RESULTS

The Web-App has deployed all three models in a local server and ran it using Flask; each of the Modes' results is present below.
Fig 4 shows the Home Page of the deployed model in the local server.

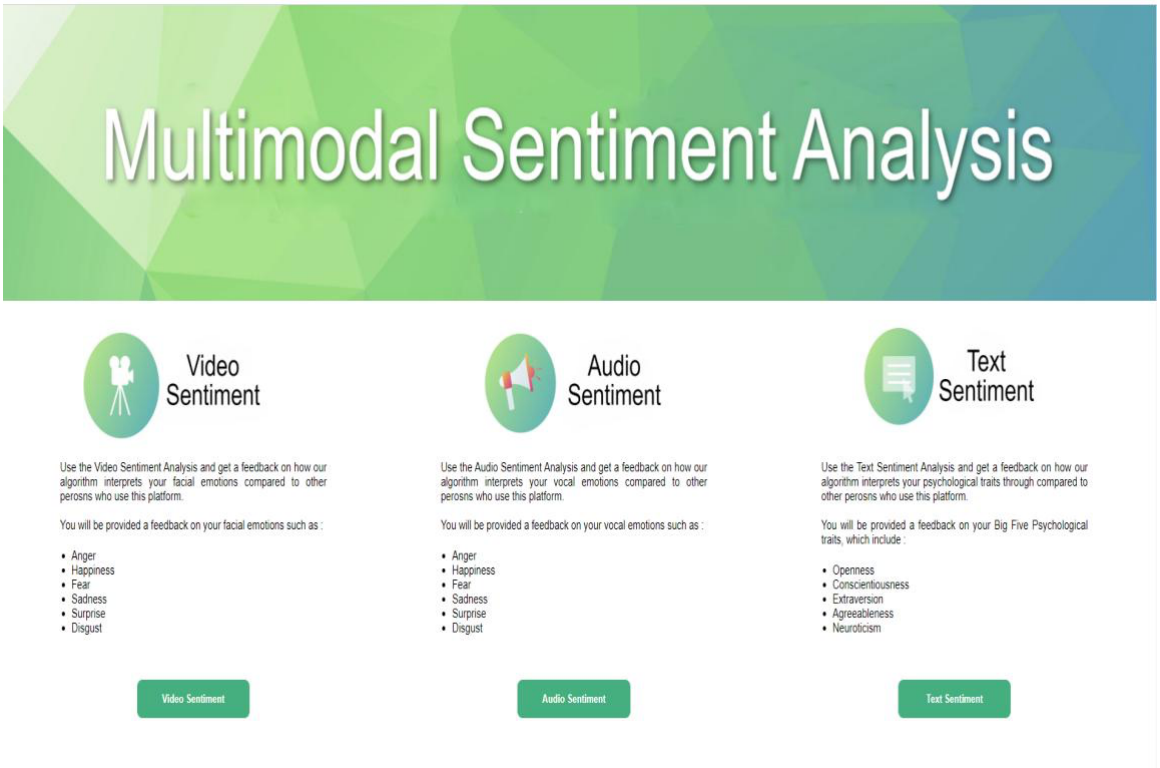

Fig. 4: Home Page

The Web-App is to be designed with three sections with Text, Audio, and Video Sentiment Analysis. The user will type in the Text Sentiment Analysis, which will use the LSTM techniques to predict the Sentiment of the data by a particular label that has been defined during the training. The Audio sections take the audio file as input in a .wav file and predict the Sentiment by calculating the MFCC's and predicting the label used in training. In the video section, real-time camera access is needed for the input of the Sentiment Analysis, and the facial expressions determine the Sentiment.

6.1 TEXT-SENTIMENT

In this Text Modal, we have implemented Text Analysis for predicting the Personality Traits in a human being used for interview simulation [4]. We can help finalize the candidate in an interview.
The dataset that has been used is by Pennebaker, and King for training and testing purposes.

Two options are added one a Dialogue Box and one Pdf Upload that will help us to identify the Personality of an individual and compare it with other candidates by plotting a bar graph.

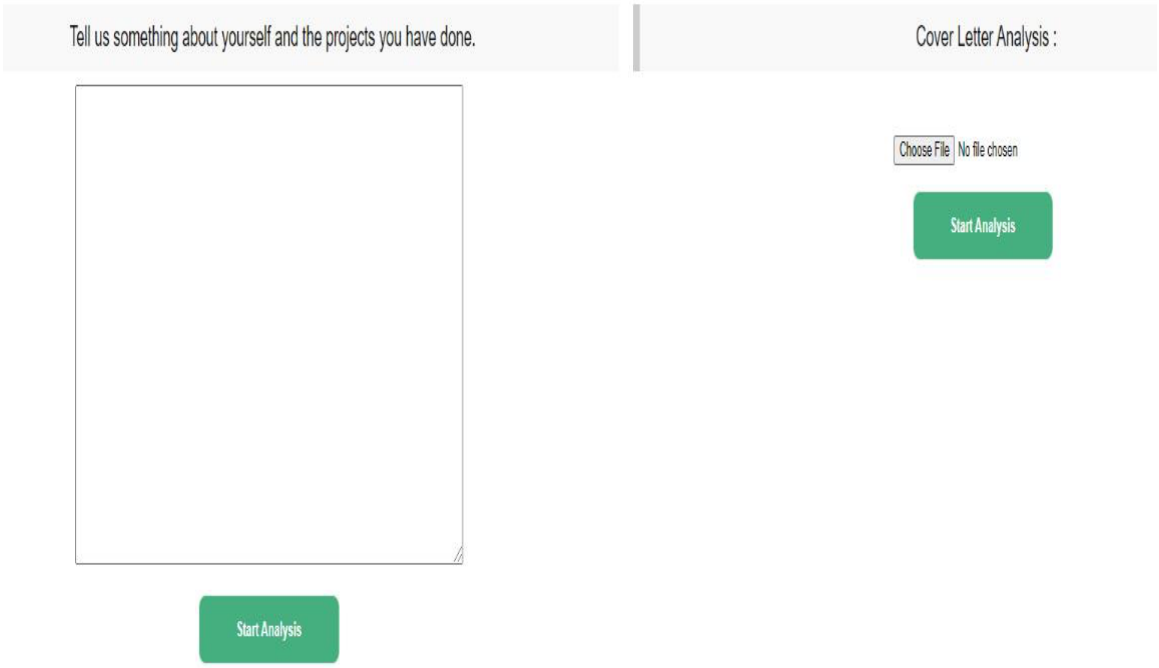

Fig. 5: Text Sentiment Home-Page

Fig 5 shows the two methods we can use in the Text-Sentiment, i.e., Text [15] and Cover Letter upload. Compared with the other candidates, the output bar plots are displayed, and the most common words that appear in the text are also shown on the sidelines. The predicted probability percentage is shown beside the bar plots.

The bar plots with probability are shown below:

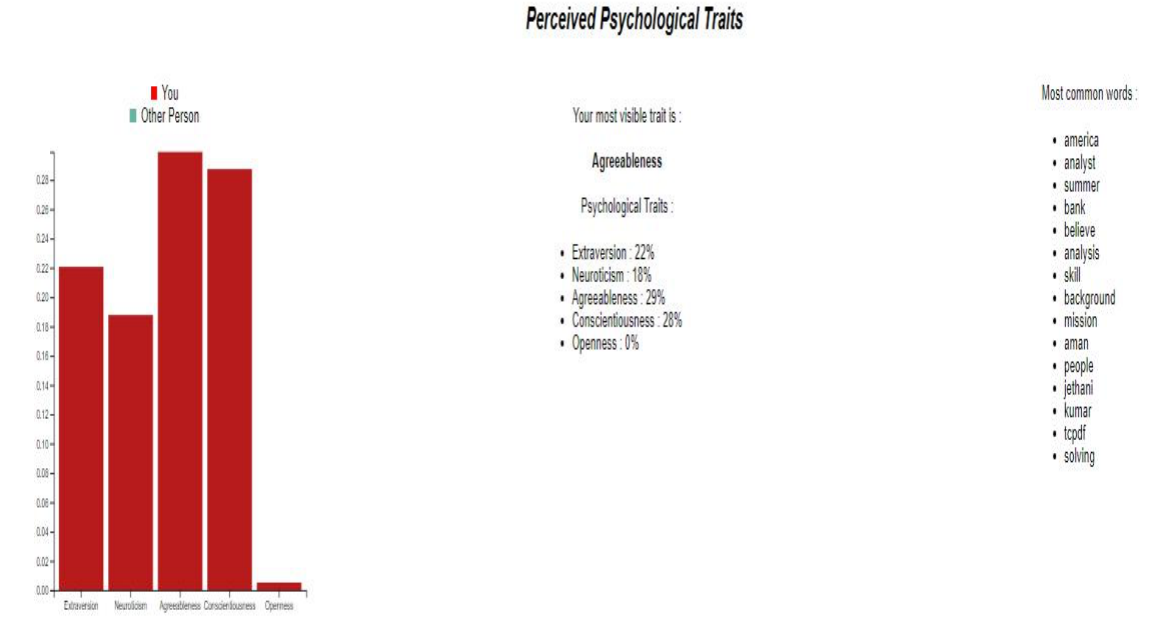

Fig. 6: Probability Bar Plot for Our Input Text

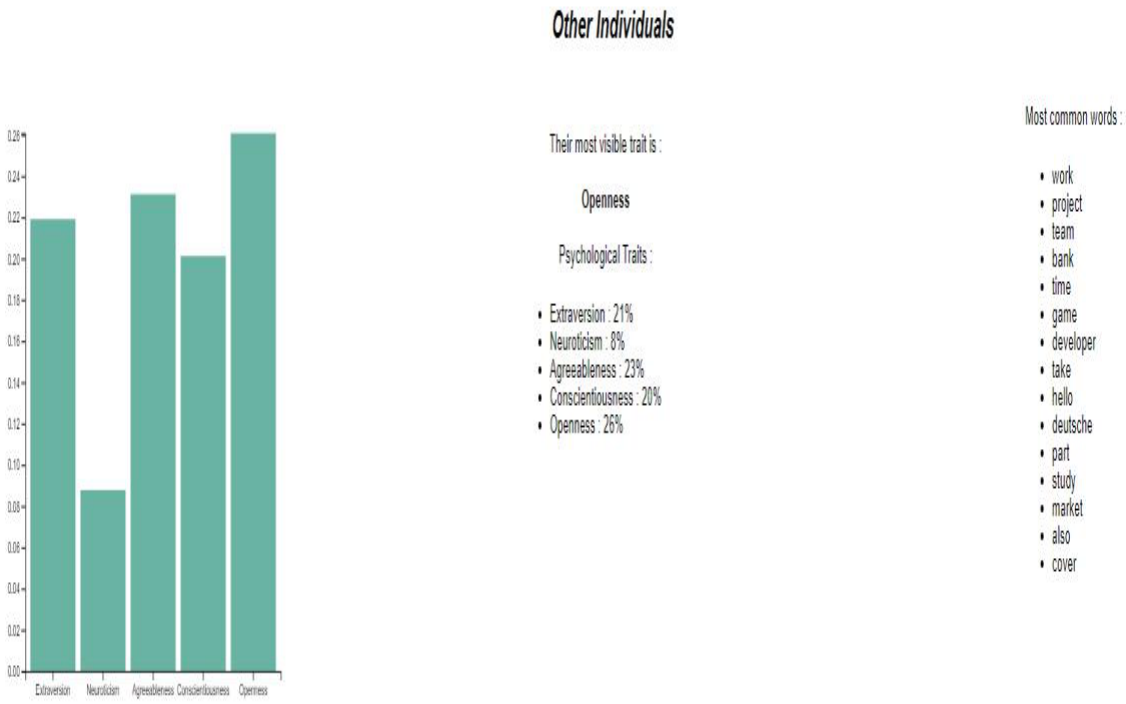

Fig. 7: Probability Bar Plot For Other Individuals

Fig 6 and 7 give us the label prediction, i.e., the emotion with the highest probability of our text input and the comparison with other individuals, respectively.

The accuracy by using di erent models is shown below. The method that has been used is Word-2-Vec embedding with LSTM and SVM models. Both the accuracy of the test set is shown below.

Model	EXT	NEU	AGR	CON	OPN
Word2Vec + SVM	46.18	48.21	49.65	49.97	50.07
Word2Vec + LSTM	55:07	50:17	54:57	53:23	53:84

Table 1: Text Accuracy Confusion Matrix

Table 1 shows the accuracy of labels with two different types of models. LSTM helped us increase the accuracy because LSTM is used as a Bidirectional and can see any independence of the current word with the future.

6.2 AUDIO-SENTIMENT

In this Audio Modal, we have implemented Audio Analysis to predict the Sentiment that takes the live audio of about 15 seconds and runs its prediction on that limited audio. The MFCC and Power-Spectogram are calculated and used in the Neural Networks or classification models.

The labels that are predicted using the Audio-Sentiment are Angry, Happy, Neutral, Sad, Disgust, and Fear, and it also plots a bar plot in the results of all the emotions perceived. The dataset that have been used "Ryerson Audio-Visual Database of Emotional Speech and Song"(RAVDESS)[20] dataset for training and testing purposes.

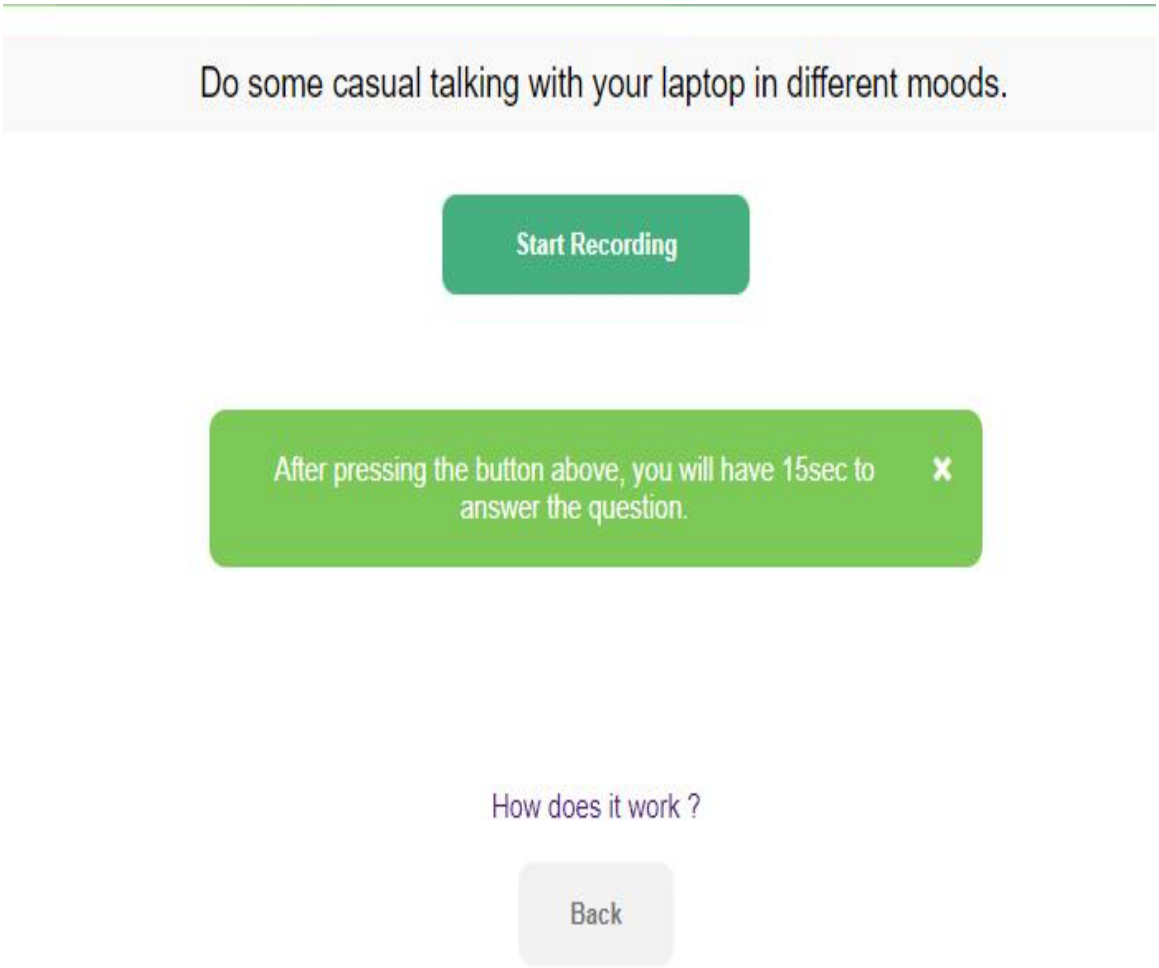

Fig 8: Audio Sentiment Home Page

As soon we click Start Recording as seen in Fig 8 in the Audio Home-Page, it starts running for 15 seconds. As the time is completed, it shows a button Get Emotion Analysis for the results.

After clicking on getting Analysis, we can see the output bar plots and compare them to how a particular person shows emotions in the audio.

The predicted probability percentage is shown beside the bar plots. The image be-low has the emotion analysis for the last two audios that were played while testing the web app.

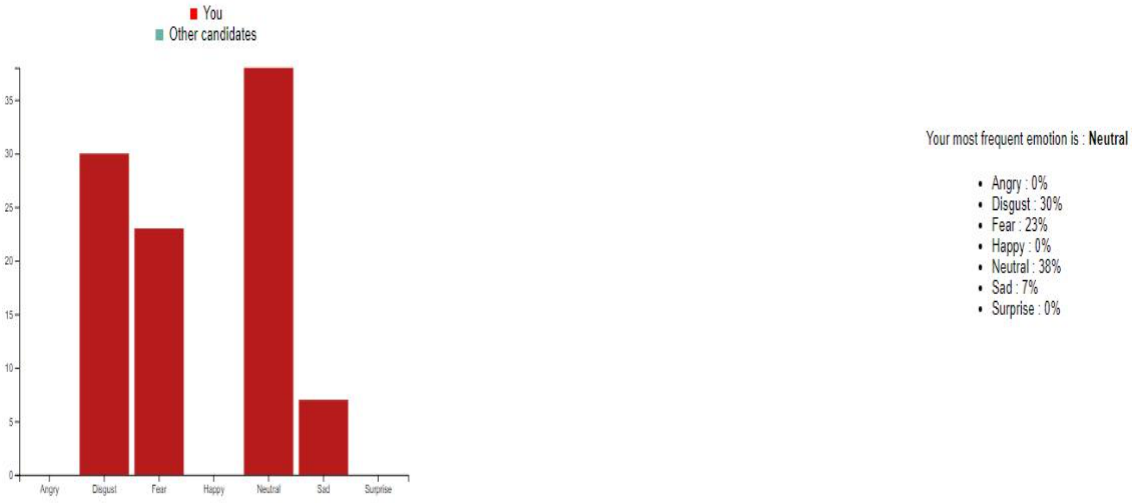

Fig. 9: Label Prediction and Bar Plot For Our Audio

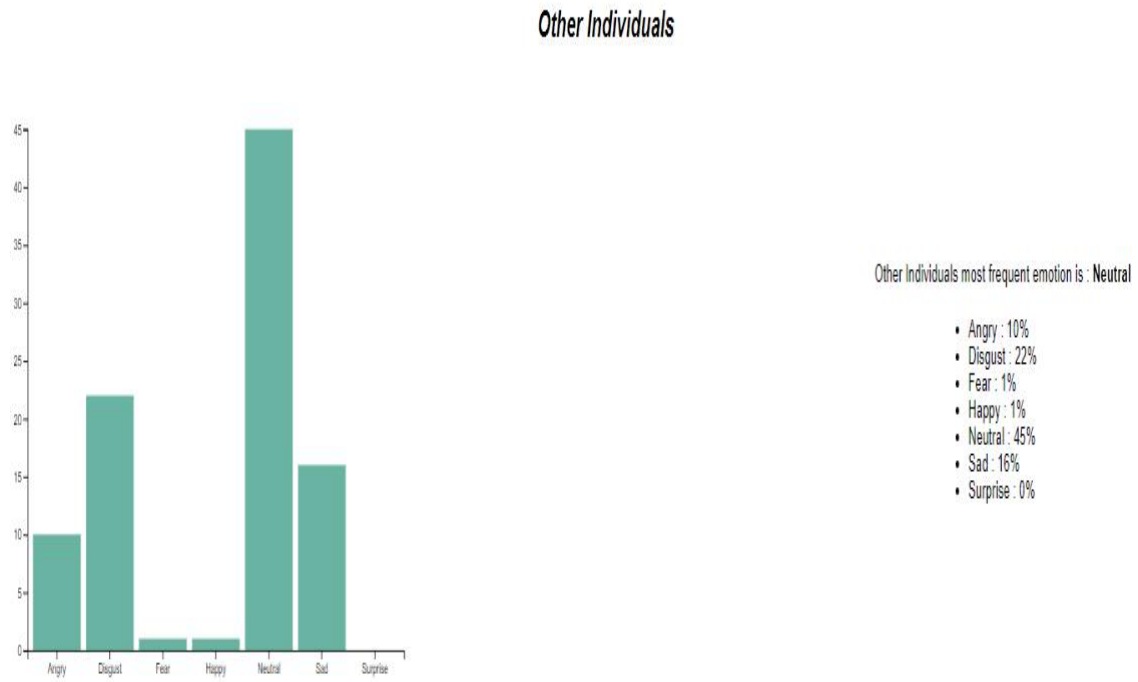

Fig. 10: Label Prediction and Bar Plot Of Others Audio

Fig 9 and 10 give us the label prediction, i.e., the emotion with the highest probability of our audio input and the comparison with other individuals, respectively.

This Audio modal have been implemented with MFCC's calculation and then fed those MFCC's to the Classification Network using the Neural Networks. The confusion matrix accuracy of each label is given below.

		Predicted labels						
		Happy	Sad	Angry	Scared	Neutral	Dis-gusted	Sur-prised
Actual labels	Happy	80.0%	0.0%	5.7%	5.7%	5.7%	2.9%	3.4%
	Sad	8.1%	81.1%	0.0%	0.0%	2.7%	8.1%	1.5%
	Angry	6.3%	6.3%	75%	0.0%	6.3%	6.3%	0%
	Scared	6.7%	0.0%	4.4%	71.1%	8.9%	8.9%	4.7%
	Neutral	11.1%	5.6%	2.8%	8.3%	66.7%	5.6%	0.3%
	Disgusted	0.0%	8.7%	0.0%	4.3%	2.2%	84.8%	2.9%
	Surprised	0.0%	8.7%	0.0%	4.3%	2.2%	84.8%	67.3%

Table 2 : Audio Accuracy Confusion Matrix

Table 2 shows the accuracy of all labels using MFCC's fed to some of the classification methods with the use of Neural Networks.

The Audio model's accuracy and loss graph plot is shown below, and the Final Accuracy can be seen from them predicting those six labels.

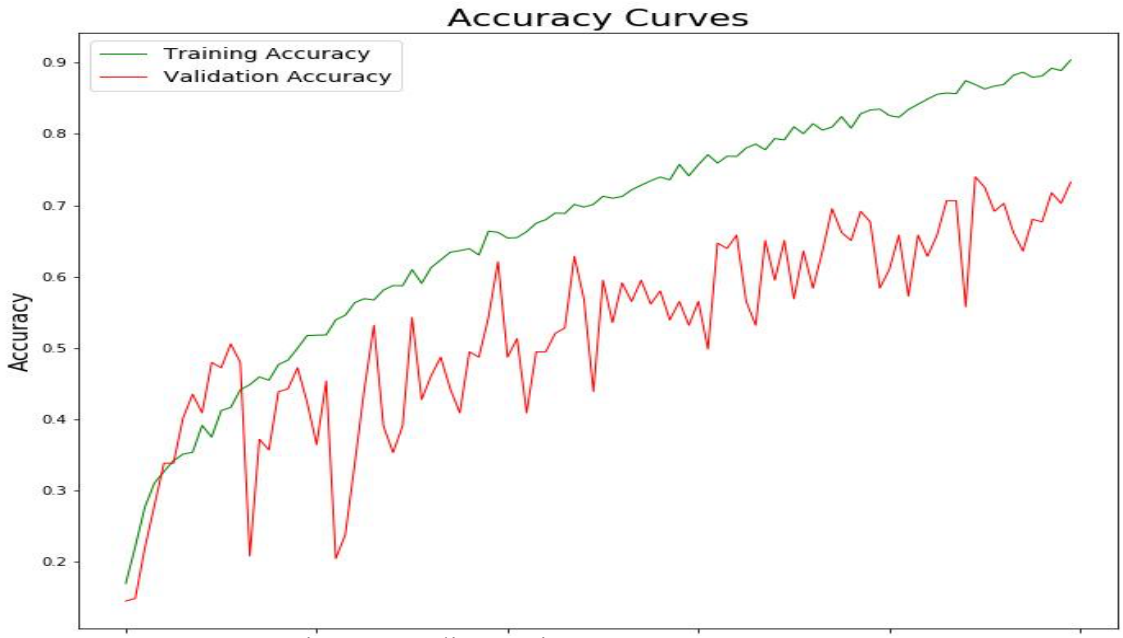

Fig. 11: Audio Sentiment Accuracy Curve

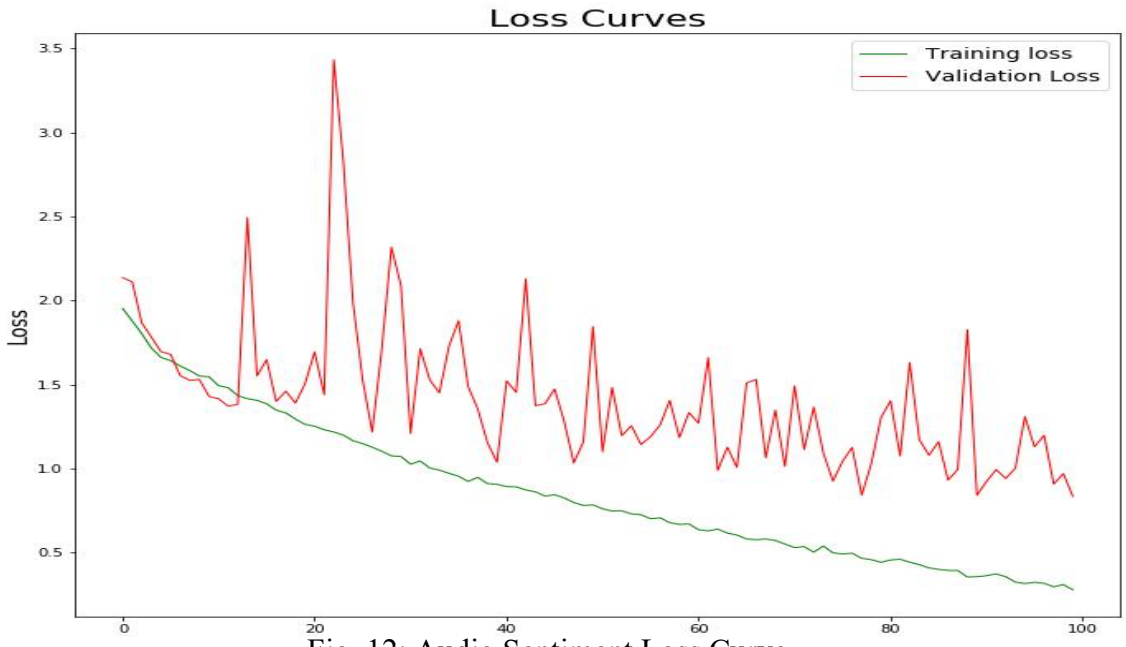

Fig. 12: Audio Sentiment Loss Curve

Note: Keras Early Stopping made the graph stops at 103 Epochs as there was no improvement in the accuracy.

Our model presents reasonably satisfying results. Our prediction recognition rate is around 75% for 7-way (happy, sad, angry, scared, disgust, surprised, neutral) emotions.

6.3 VIDEO-SENTIMENT

In this Video Modal, we have implemented Video Analysis for predicting the Sentiment that takes the live webcam feed and runs its prediction on that live video, detects our emotions, and identifies the number of faces [21]. The process is simple; the video is broken into frames. Each frame is convolved using filters, and landmarks points are obtained using that filtered image to predict sentiments.

The labels that are predicted using the Video-Sentiment are Angry, Happy, Neutral, Sad, Disgust, and Fear, and it also plots a bar plot in the results of all the emotions perceived. It also tells our emotions in a line chart throughout 45 sec.

The dataset that has been used is FER2013 Kaggle Challenge dataset for training and testing purposes.

Fig 13 shows us the Home Page for Video Analysis and has a start recording button that takes us to the new page where sentiment analysis is done on a live webcam, as shown in Fig 14 and Fig 15, respectively.

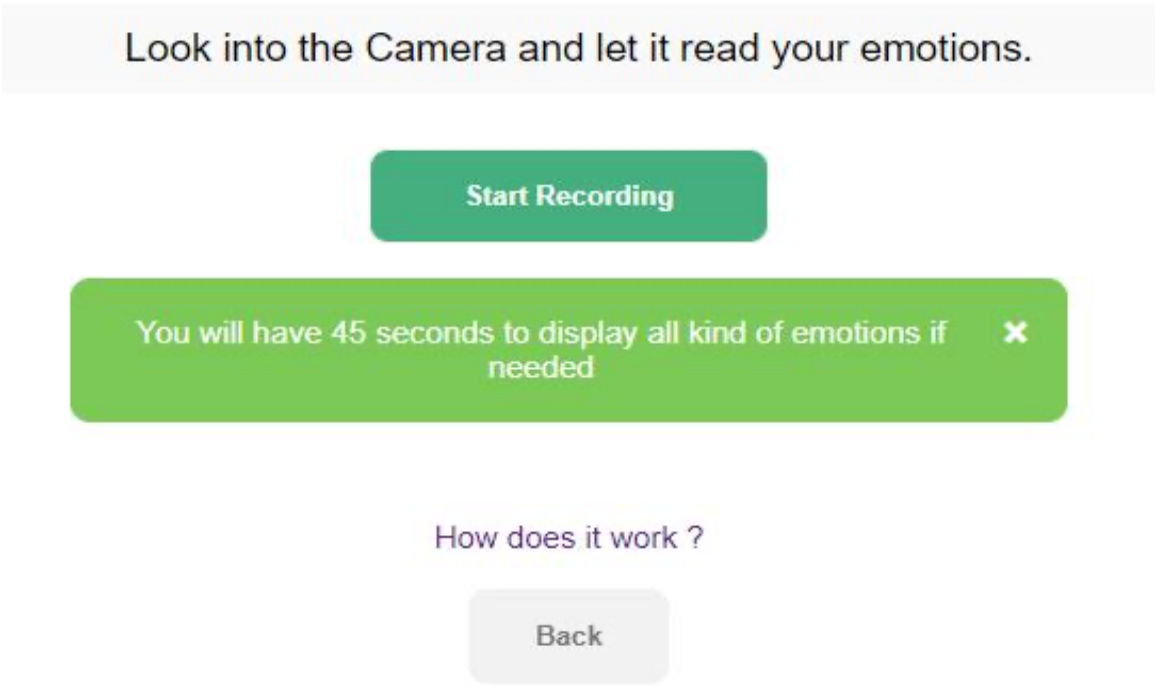

Fig. 13: Video Sentiment Home-Page

As soon we click Start Recording in the Video Home-Page it starts running for 45 seconds and moves us to a another window where live webcam emotions can be detected. The images of the live emotion detection are shown below.

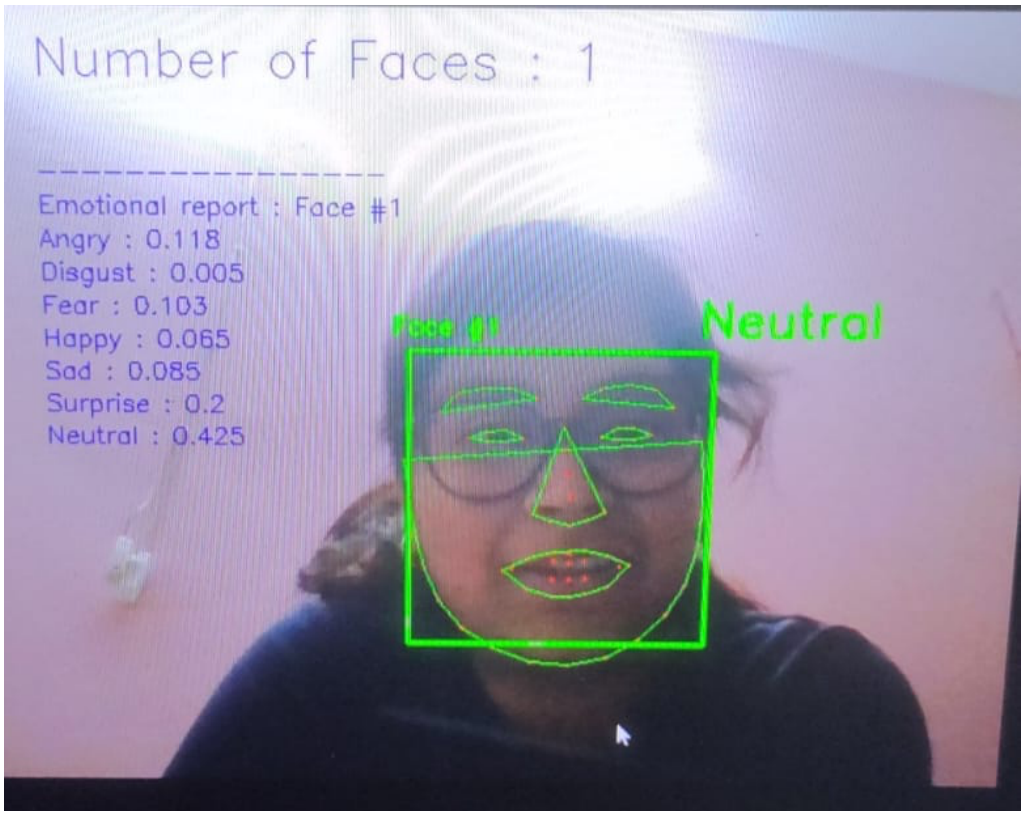

Fig. 14: Emotion Detected(Neutral)

Fig. 15: Emotion Detected(happy)

The figure above shows the emotions in the green box by using the positions of the landmarks and thus making of call for an emotion.

After the video is over recording, we move to the next page with the bar plots with the probability of the expressions over the period and a line chart that shows how our emotions have varied.

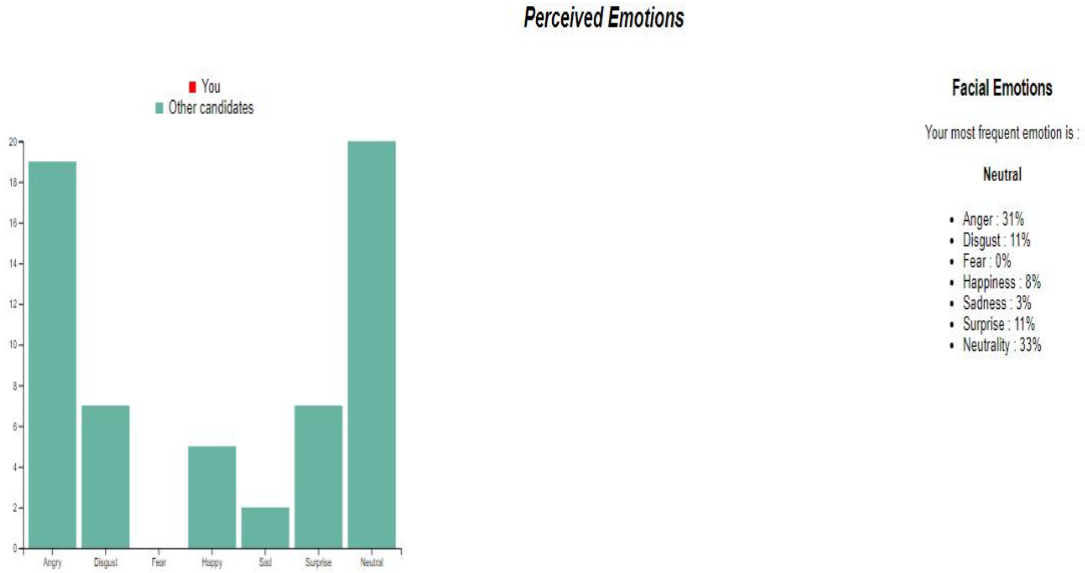

Fig. 17: Probability Bar Plot For Our Input Live Video

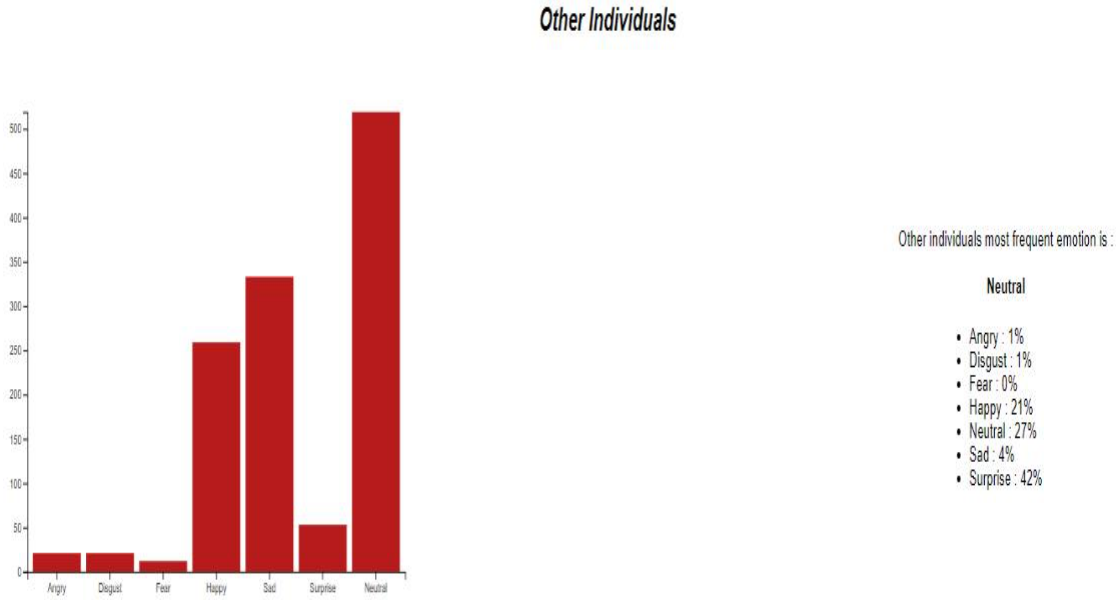

Fig. 18: Probability Bar Plot Of Other Individuals

Fig 17 and 18 gives us the label prediction i.e. the emotion with highest probability of our video feed and also comparison with other individuals respectively.

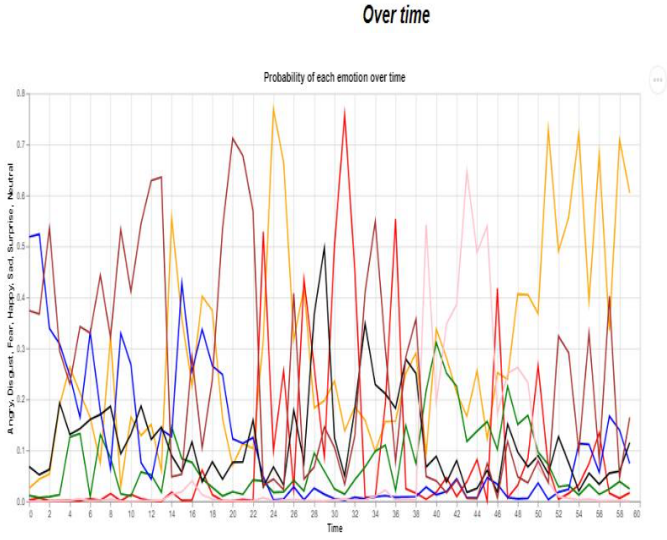

Fig. 19: Line Chart for Varying Emotions

Fig 19 shows us how our emotions vary concerning the time using a line chart that can be used in the long run to get the mean Sentiment. We have used the Xception [30] model that is a Transfer Learning Model and is used in competition for predictions of the 1000 labels.

Layer (type)	Output Shape	Param #
conv2d_2 (Conv2D)	(None, 48, 48, 32)	320
max_pooling2d_2 (MaxPooling2	(None, 24, 24, 32)	0
batch_normalization_1 (Batch	(None, 24, 24, 32)	128
conv2d_3 (Conv2D)	(None, 22, 22, 32)	9248
max_pooling2d_3 (MaxPooling2	(None, 11, 11, 32)	0
batch_normalization_2 (Batch	(None, 11, 11, 32)	128
conv2d_4 (Conv2D)	(None, 11, 11, 32)	9248
max_pooling2d_4 (MaxPooling2	(None, 5, 5, 32)	0
conv2d_5 (Conv2D)	(None, 5, 5, 32)	9248
flatten_1 (Flatten)	(None, 800)	0
dense_1 (Dense)	(None, 512)	410112
dense_2 (Dense)	(None, 7)	3591

Total params: 442,023
Trainable params: 441,895
Non-trainable params: 128

Fig. 20: Figure for Varying Emotions

Fig 20 shows the Keras Xception [16] model summary and all the layers that have been used.

The accuracy and loss graph for that model is shown below.

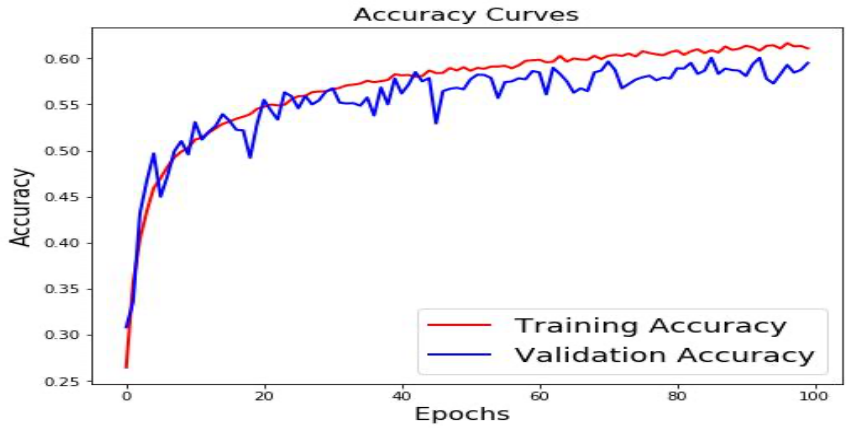

Fig. 21: Xception Accuracy Graph

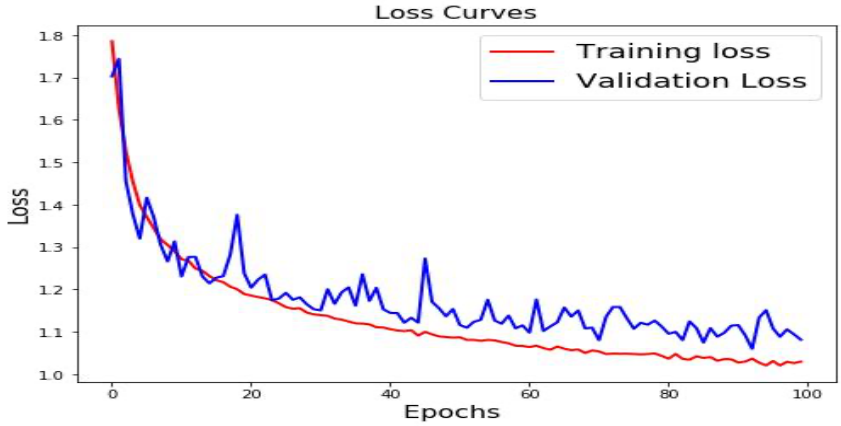

Fig. 22: Xception Loss Graph

Note:Keras Early Stopping made the graph stops at 100 Epochs as there was no improvement in the accuracy. Fig 21 and Fig 22 show the trend of the Accuracy and Loss of the trained and tested model using the Xception Transfer Learning.

This modal was also tried on di erent lengths of videos like 15sec, 30sec, 40sec, but there was no significant impact on the accuracy, so we only implemented it on 45sec.

6.4 APPLICATIONS

In this venture, an online application can be used in Call-Centres to avoid the feedback message provided at the end of the call at the customer service. To upgrade it, the Sentiment can be derived from the audio as both speakers speak continuously.

Text Sentiment can be used in the interviews to detect the person's emotions by making the candidate type or say and getting how confident the candidate is in the discussion. It can also be used by getting the candidate's personality traits by making them type in the portal, and also a cover letter option is available to do so the same.

7. CONCLUSION AND FUTURE SCOPE

This web application helps us in identifying the emotions of an Individual. It is useful if used in all modes of communication, i.e., Text, Audio, and Video. Audio Field can be used in call centres for customer complaint satisfaction, Video and Text combine can be used for many interview purposes. The database for the audio and video sentiment is too large that it requires a great time in training and getting better results.

We can improve the mode of Text by using BERT techniques, and the Audio field can be improved by combining multiple techniques HMM, CNN, and MFCC, together.
We can give more accurate results if we use 2 modes at once like Audio and Video together to get the better accuracy for the predicted labels

8. REFERENCES

1. Zeng, X. Wang, A. Wu, Y. Wang, Q. Li, A. Endert, and H. Qu, "Emoco:Visual analysis of emotion coherence in presentation videos," IEEE Transactions on Visualization and Computer Graphics, p. 1–1, 2019. [Online]. Available:http://dx.doi.org/10.1109/TVCG.2019.2934656

2. Zeng, X. Shu, Y. Wang, Y. Wang, L. Zhang, T. Pong, and H. Qu, "Emotioncues: Emotion-oriented visual summarization of classroom videos," IEEE Transactions on Visualization and Computer Graphics, vol. 27, pp. 3168–3181, 2021 Y.Jia and S. SungChu, "A deep learning system for sentiment analysis of service calls," ArXiv, vol. abs/2004.10320, 2020

3. Fast, B. Chen, and M. S. Bernstein, "Empath: Understanding topic signals in large-scale text," in Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, ser. CHI '16. New York, NY, USA:

4. Association for Computing Machinery, 2016, p. 4647–4657. [Online]. Available:https://doi.org/10.1145/2858036.285853

5. Mandera, E. Keuleers, and M. Brysbaert, "How useful are corpus-based methods for extrapolating psycholinguistic variables?" Quarterly Journal of Experimental Psychology, vol. 68, no. 8, pp. 1623–1642, 2015

6. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," CoRR, vol. abs/1706.03762, 2017. Pham, P. Liang, T. Manzini, L.-P. Morency, and B. Poczos, "Found in translation: Learning robust joint representations by cyclic translations between modalities,"

7. Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 6892–6899, 07 2019

8. M. Chen, S. Wang, P. P. Liang, T. Baltruaitis, A. Zadeh, and L.-P. Morency,"Multimodal sentiment analysis with word-level fusion and reinforcement learning,"Proceedings of the 19th ACM International Conference on Multimodal Interaction, 2017.

9. A . Zadeh, P. Liang, N. Mazumder, S. Poria, E. Cambria, and L.-P. Morency, "Memory fusion network for multi-view sequential learning," Association for the Advancement of Artificial Intelligence, 02 2018

10. E. Sariyanidi, H. Gunes, and A. Cavallaro, "Automatic analysis of facial affect: A survey of registration, representation, and recognition," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 37, no. 6, pp. 1113–1133, 2015.

11. V. Campos, A. Salvador, B. Jou, and X. Giró-i Nieto, "Diving deep into sentiment: Understanding fine-tuned cnns for visual sentiment prediction," 10 2015

12. N. Pappas, M. Redi, M. Topkara, B. Jou, H. Liu, T. Chen, and S. Chang, "Multilingual visual sentiment concept matching," 06 2016

13. Q. You, J. Luo, H. Jin, and J. Yang, “Robust image sentiment analysis using progressively trained and domain transferred deep networks,” 09 2015

14. Y-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov,“Multimodal transformer for unaligned multimodal language sequences,” in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: Association for Computational Linguistics, Jul.

15. 2019, pp. 6558–6569. [Online]. Available: https://aclanthology.org/P19-1656

16. W. W. Lo, X. Yang, and Y. Wang, “An xception convolutional neural network for malware classification with transfer learning,” in 2019 10th IFIP International Conference on New Technologies, Mobility and Security (NTMS), 2019

17. Word embeddings,” https://en.wikipedia.org/wiki/Word_embedding, last accessed on 2021-10-30.

18. “Mfcc(s),” https://en.wikipedia.org/wiki/Mel-frequency_cepstrum, last accessed on 2021-10-30

19. Pennebaker-king,” https://sites.google.com/michalkosinski.com/mypersonality, last accessed on 2021-10-30

20. The ryerson audio-visual database of emotional speech and song (ravdess),” https://smartlaboratory.org/ravdess/, last accessed on 2021-10-30.

21. “The facial emotion recognition challenge from kaggle,” https://www.kaggle.com/ deadskull7/fer2013, last accessed on 2021-10-30.

22. “Opencv,” https://opencv.org/, last accessed on 2021-10-30.

23. “Javascript,” https://en.wikipedia.org/wiki/HTML, last accessed on 2021-10-30.

24. “Numpy,” https://numpy.org/, last accessed on 2021-10-30.

25. “Tensorflow,” https://www.tensorflow.org/, last accessed on 2021-10-30.

26. “Flask,” https://en.wikipedia.org/wiki/Flask_(web_framework), accessed:2021-10-23.

27. “Google colab,” https://www.tutorialspoint.com/google_colab/what_is_google_colab.htm, last accessed on 2021-10-30.

28. “Spyder,” https://en.wikipedia.org/wiki/Spyder_(software), last accessed on 2021-10-30.

29. “Vs code,” https://code.visualstudio.com/docs, last accessed on 2021-10-30.

30. “Xception,” https://keras.io/api/applications/xception/, last accessed on 2021-10-30