Enhancing Sentimental analysis using Multimodal data

Raj Kishor Verma[1] , Divyansh Mittal[2] , Manish Sharma[2] , Ram Jindal[2]

[1]Department of CSE-DS, ABES Institute of Technology, Ghaziabad, 201009, Uttar Pradesh, India
{rvrajverma77@gmail.com}
[2]Department of CSE-DS, ABES Institute of Technology, Ghaziabad, 201009, Uttar Pradesh, India
{ divyanshmittal7225@gmail.com ,manishsharma3515@outlook.com. ramjindal1234@gmail.com}

ABSTRACT

Sentiment analysis is crucial for assessing opinions regarding a variety of topics. It includes text, audio, and video. It assesses brand sentiment and customer happiness in detail. The emergence of multimodal Sentiment Analysis, which incorporates data streams beyond text, can be attributed to the advent of social media. Multimodal models aspire to cultivate a more profound and comprehensive understanding of the data, paving the way for novel insights and unlocking diverse applications. It is a powerful tool for sentiment identification that includes audio, photos, voice expressions, and more. Applications include everything from human-machine interactions to video blogs. The proposed sentiment analysis model aims to detect sentiments in video highlights using time-sync comments, departing from the conventional method that relies solely on basic sentiment word combinations and coarse sentiment categorization. This approach seeks to overcome the limitations of traditional machine learning classification techniques. In this article, we propose modern technologies like HMM, CNN, and MFCC for more accurate results. the information we gather from all three categories audio, video, and text so the cluster of all data helps us to find the better results. In this article Sentiment analysis across several domains is a constantly developing field that presents opportunities and challenges for further research.

Keywords: Multimodal sentiment analysis, HMM(Hidden Markov Models), CNN(Convolutional Neural Network), MFCC(Mel-Frequency Cepstral Coefficients), Sentiment identification.

1. INTRODUCTION

Sentiment Analysis is a procedure aimed to analyse people's preferences, habits, and content on platforms like social ones. Sentiment Analysis by using multimodal data is an extension to typical text-based analysis using audio and visual data and containing additional implication of the problem with interpreting the written data. Sentiment is a motivation in which an individual feels concerning a specific object, topic, or entity. Its significance means understanding people's position, process, or evaluation of it[2].The text-based feeling investigation has been the leading figure around here and, as of later, examined different modalities, like audio and vision[3] Sentiment analysis involves automatically detecting four key components of a concept: entity, viewpoint, entity holder, and the feeling associated with the viewpoint[4].A robust sentiment analysis framework should be capable of accurately isolating and discerning these four components. A new improvement in multimodal sentiment analysis is textual, audio, visual assumption investigation. Web-based means of communication 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. Regarding vocal and visual modalities, recordings provide multimodal information. Along with the written content, spoken word balances and facial expressions in the form of images

© CINEFORUM

provide critical clues to determine the opinion holder's true state of emotions. As a result, a combination of textual and graphic information aids in accelerating the creation of an improved assumption and sentiment assessment model.

Sentiment analysis has become a critical tool for understanding user sentiments and preferences in the era of social media and multimedia content. This paper explores the background, significance, methodologies, and applications of multimodal sentiment analysis, emphasizing its recent development in visual sentiment analysis. The integration of text, audio, and visual data provides a more comprehensive view of user emotions and sentiments, making it invaluable for various domains, from marketing to mental health[1] Real-time sentiment analysis has become prominent, particularly in applications such as social media monitoring, customer service, and mental health support. Researchers are fervently engaged in the development of efficient algorithms and models capable of analysing sentiment in real-time. The utilization of streaming data and innovative edge computing technologies has become increasingly prevalent in the architecture of real-time sentiment analysis systems. Audio data,

encompassing speech and non-verbal vocal cues, presents a treasure trove of emotional insights. Recent studies have ardently focused on enhancing emotion recognition within audio data through the adept deployment of deep learning techniques. For instance, deep neural networks (DNNs) have been harnessed to discern emotional states based on acoustic features like pitch, intensity, and spectral characteristics. This pioneering research holds immense value for applications such as voice assistants and call centre analytics. Visual sentiment analysis transcends mere facial expression recognition by delving into the sentiment conveyed by complete scenes or images. Researchers have pioneered models capable of scrutinizing the composition, objects, and contextual elements within an image to deduce its overall sentiment. Notably, a study by [2]introduced a novel visual sentiment analysis approach that takes into 6 account not only facial expressions but also the presence of objects and scene attributes to arrive at a holistic assessment of an image's sentiment. The burgeoning field of multimodal sentiment analysis has found a particularly promising application in the realm of mental health assessment. Recent studies are steadfastly aimed at identifying early indicators of mental health issues by scrutinizing sentiment patterns present in textual, audio, and visual data gathered from individuals. Machine learning models are meticulously trained to discern shifts in sentiment that may signify emotional distress or mood disorders. Such groundbreaking research has the potential to usher in a change in basic assumptions in the way we monitor and provide support for mental health concerns. Recent research has focused on developing models that can effectively learn from multiple modalities and fuse information intelligently. Cross-modal sentiment analysis aims to leverage the complementary nature of different modalities to enhance the accuracy of sentiment predictions. For example, a proposed a cross-modal sentiment analysis framework that incorporates both text and visual data. The model learned to associate textual sentiment expressions with corresponding visual cues, leading to improved sentiment classification[3] The development and rapid development of sentiment analysis are paralleled by the proliferation of social media channels on the internet, encompassing sites that include blogs, forums, reviews, microblogs as, Facebook, and social networks. The amazing amount of opinionated data that is now accessible in digital media is the primary driver of this increase, which represents a critical turning point in human history. In the realm of natural language processing (NLP), the analysis of sentiment has emerged as a hot topic for research. It is being investigated in the areas of data mining, web mining, text mining, and information retrieval, among various other areas. Because of its significant influence on business and society, its importance goes beyond computer science to include fields like management sciences and social sciences, such as marketing, finance, political

© CINEFORUM

philosophy, communications, medical research, and even history.

1.1 Text Sentiment: The aim to extract evaluative meaning, an alternative to topic detection in the realm of sentiment analysis has emerged. The most promising advancement in analysing text sentiment stems from the utilization of deep learning. Deep learning can exploit vast datasets to capture word embeddings that are pertinent for sentiment analysis, yielding naturally enriched lexicons. While deriving word categories based on deep learning techniques is yielding results remarkably close to those of human annotators, recent investigations have revealed that extrapolating word sentiment inherent factors based on word embeddings still necessitates considerable refinement. Profound Recurrent Neural Networks have been deployed for the task of subjectivity detection, and word vector representations can amalgamate supervised and unsupervised learning approaches when applied to sentiment analysis. In their research on Text Sentiment, the authors employed SVM; however, I propose utilizing alternative techniques, such as leveraging word embeddings with the Attention Mechanism.

1.2 Audio Sentiment: Audio is mode of communication in language, the line among opinion and feeling investigation 7 is regularly extremely frail, as, e. g. In on pitch-related provisions and saw that addition- ally, without text-based signals pitch contains data on feeling. Various further works centre around feeling examination solely from the text-based substance as present in the discourse. The Audio Sentiment implement by authors of used KNN for their purpose. We will be calculating the MFCC's for conducting 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 stores furthermore, predicts fleeting cross- modular collaborations. It 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.

1.3 Video Sentiment: They analyzed numerous shading highlights, such as sharpness, immersion, and lightning highlights linked to support vector relapse, in order to forecast the recurrence of these sets, including as brilliant– gloomy, warm–cool, and vibrant–desolate. All of these efforts to advance and use visual perception analysis demonstrate the possibility of both reached inclusion and improved precision approaches, such as CNNs and multilingual and alternative material sources. Moreover, the steadily increasing quantity of publicly available personal computer vision models/libraries and visual perception datasets means that visual analysis of opinion is
poised for improvement in both directions as well.

The complex idea of feeling shows that visual feeling investigation alone cannot gauge and additionally portray our experiential attitude and sentiments in interactive media information. For instance, visible substance will not 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.

1.2 RESEARCH OBJECTIVE

The research aimed to achieve the following objectives:

a) To develop a Multimodal Sentiment Analysis system that framework interprets emotions from text,

© CINEFORUM

speech, and visual content, offering a wide-ranging approach for sentiment analysis across various communication modes and data types.

b) Designed a user-friendly graphical user interface (GUI) aimed at enhancing sentiment analysis.
c) A unified system capable of interpreting sentiments seamlessly across different forms of human expression.

2. LITERATURE REVIEW

- **Rui Zhang et al. 2023[2]-**The bimodal multi-head attention network (bman) proposes a solution to the alignment problem in multimodal representation learning, aiming to harmonize different modal information and address vector offset issues. The bman framework receives text and audio sequences as input, comprising two crucial components: unimodal encoders and a bimodal decoder.
- **Jiangfeng L et al. 2023 [4]-** Sentiment Highlight Extraction involves finding fragments with similar sentiments, assigning highlight scores based on TSC density, and merging relevant fragments. The process results in a set of sentiment highlights, reflecting the video's emotional dynamics. Sentiment Intensity Calculation quantifies sentiment strength for each highlight, considering emotional words, adverbs, and negative words in Time-Sync Comments. The sentiment intensity is calculated based on linguistic analysis, providing a nuanced measure of sentiment for each sentiment type.
- **Chuanmin Mi et al. 2022[5]-** The research framework, as shown in Fig. 1, involves predicting video views for web series based on comment sentiment. Input variables are categorized into drama-related, marketing, phased, and word-of-mouth factors. Sentiment analysis on web series comments yields sentiment scores, while data normalization is applied. Utilizing base learners like Random Forest, GBDT, XG Boost, and Light-GBM, an improved stacking ensemble model is established for prediction. The predictive model works in stages, with eleven independent variables, including three static and eight dynamic ones, for each stage. The dependent variable is the cumulative video views (Drama View) at the end of each week, transitioning to an independent variable in the next stage.
- **Arif Ullah et al. 2022 [6]-** The research emphasizes the importance of comparative studies in evaluating products or services based on textual data from reviewers. The focus is on extracting information from reviewer texts to assess the risks and marketing strategies. A relationship between sentiment comparison and target entities is explored, involving words and aspects for opinion extraction. Sentiment analysis is classified into binary, ternary, unary, or thumbs up/down categories. Machine learning and lexicon-based approaches are employed, with machine learning further categorized into supervised, unsupervised, and semi-supervised methods. Various classifiers such as Decision Tree, Support Vector Machine, Neural Networks, Naive Bayes, and Maximum Entropy are utilized.
- **Jyoti Yadav et al.2023[7]-**This paper employs social media mining and sentiment analysis to scrutinize global companies, examining data from diverse continents. Using label details and storyboards, it explores sentiment analysis on various social media applications worldwide, visually presenting data distribution through pie charts. The study underscores the potency of extracting brand characteristics from social media messages, revealing insights applicable to numerous brands.
- **Nik Nur Wahidah Nik Hashim et al. 2023[8]-** This research focused on detecting depression in Bahasa Malaysia text using Natural Language Processing (NLP) through classification and sentiment analysis techniques.

© CINEFORUM

Three distinct questions were posed to depressed and healthy individuals, revealing that depressed respondents frequently referenced school and study-related experiences, indicating a prevalence among students.

•**Adil Baqach et al. 2022[9]-** The research explores various machine learning and deep learning approaches for sentiment analysis across diverse domains. It employed the Naïve Bayes algorithm to analyse sentiment in children's fairy tales, outperforming other methods by extracting thirty features. In contrast, utilized Support Vector Machines (SVMs) to discern emotions in text, achieving superior results with Rough Set Theory augmentation. Real-time student feedback was leveraged by, employing Naïve Bayes, Compliment Naïve Bayes, Maximum Entropy, and SVMs, with SVM emerging as the top classifier with 94% accuracy. developed a sentiment prediction engine for Twitter responses, revealing Compliment Naïve Bayes as the most effective classifier.

•**Kiran V. Sonkamble et al. 2023[1]-** This research focuses on leveraging modern computer-assisted text analysis tools, namely LIWC and Empath, for the analysis of text data in the domain of scholarly writing, particularly within children's storybook reviews. The distinctive feature of scholarly writing is conciseness and a comprehensive understanding of the subject matter. The chosen tools, LIWC and Empath, operate on dictionary- based word counts, with Empath offering a broader spectrum of categories and the ability to create and validate new categories through unsupervised language modelling. In contrast, LIWC provides a meticulously validated dictionary for analysis.

•**Zhi Li et al. 2022[10]-** This paper introduces a sentiment analysis method, SD-NB (Sentiment Dictionary and Naïve Bayes), for danmaku videos, harnessing the burgeoning realm of network media-sharing. By constructing a specialized danmaku sentiment dictionary, incorporating 161 main emoticons from the danmaku video platform, and applying Naïve Bayes classification, the study investigates the time distribution of seven sentiment dimensions and two-polarity sentiment values. The results reveal dynamic changes in danmaku review volume and sentiment values over time and video content.

- **Gina Khayatun Nufus et al.2021[11]-** This study investigates aspect-based sentiment analysis for user satisfaction with the Netflix application in Indonesia, employing the Long Short-Term Memory (LSTM) model. Deep learning, particularly LSTM, has shown superiority in text classification tasks like sentiment analysis due to its ability to overcome issues like the vanishing gradient problem encountered by traditional Recurrent Neural Networks (RNNs). The implementation involves classifying review data into positive or negative sentiments, focusing on customer feedback regarding the Netflix application.

- **Asher Prescott et al. 2023[12]-** This research paper explores about sentiment classification in short videos using different modalities and fusion models, particularly focusing on the GSO-2016, GIFGIF, and MOSI datasets. Results on the GSO-2016 dataset demonstrate that incorporating textual information into short video analysis significantly enhances performance metrics, with improvements of 12.36% in precision (P), 13.55% in recall (R), 12.43% in F1 score, and 17.80% in accuracy (Acc) compared to visual analysis alone.

- **Yuxin Jin et al. 2023[10]-** This review offers a thorough examination of how natural language processing is utilized in analysing sentiment in text. Sentiment characteristics can be efficiently derived from text through various pre-processing techniques and methods for feature extraction. Diverse algorithms for sentiment classification and techniques for analysing sentiment intensity contribute to gaining insights into the sentiment trends and strength within texts. Additionally, the inclusion of multilingual sentiment analysis and case studies across various application domains highlights the broad scope of sentiment analysis applications.

© CINEFORUM

- **Ping He et al. 2023[3]-** The proposed model addresses sentiment analysis in unaligned data scenarios by employing a cross-modal approach, primarily focusing on text features, and using video features as auxiliary information. The emotional feature extraction module involves BERT word embedding for text information and Bi-GRU for contextual encoding. For video feature extraction, facial expression features are obtained using FACET and Bi-GRU. The unimodal features are then enhanced through a cyclic memory enhancement network across time steps. This network utilizes a cross-modal hierarchical attention module to progressively strengthen cross-modal fusion features.
- **Md Abu Rumman Refatet al.2021[13]-**The model outperforms established architectures like VGG16 and ResNet50 in terms of accuracy, precision, recall, and F1- score. Additionally, it achieves competitive results against related methods on FER2013, KDEF, and JAFFE datasets, demonstrating the effectiveness of its shallow CNN design. The proposed model strikes a balance between accuracy and computational efficiency, highlighting superior performance in facial expression recognition tasks.
- **Feiyang Chen et al. 2019[14]-** The proposed DFF-ATMF, a Dynamic Fusion Framework with Audio-Text Multi-Feature, introduces attention mechanisms into an encoder-decoder framework for sentiment analysis. Inspired by neural machine translation, it enhances multimodal attention through a multi-feature fusion strategy, capturing sentiment information from audio and text. The model, evaluated on CMUMOSI, CMU-MOSEI, and IEMOCAP datasets, surpasses state-of-the-art models in accuracy and F1-score, highlighting robust generalization. Attention weight distribution heatmaps reveal the model's effective feature focus and learning behaviour.
- **Anay Ghosh et al. 2022[15]-** The provided text describes a multimodal sentiment analysis system for text and image data. In text-based sentiment analysis, comments are pre-processed and transformed into a feature matrix using a CNN architecture. The matrix is split into training and testing sets for classification using LR, KNN, CART, and SVM classifiers. The system performs better with SVM and LSTM classifiers, particularly for 2-class problems. The proposed system's performance surpasses existing methods, as indicated by F1-score and accuracy. In image-based sentiment analysis, a CNN architecture is applied to extracted facial regions, achieving improved performance with specific batch sizes and epochs.
- **Songning Lai et al.2023[16]-** This review explores multimodal sentiment analysis, emphasizing its significance in natural language processing and computer vision. It covers the research background, definition, and development process. Common benchmark datasets are summarized, and recent state-of-the-art models are compared. The review identifies challenges and suggests research directions, including the creation eleven of large multimodal sentiment datasets in multiple languages, solving domain transfer issues, building a unified model with strong generalization, reducing model parameters, addressing multilingual hybridity, optimizing modal fusion weights.
- **Harika Abburi et al.2018[17]-** The paper discusses audio sentiment detection using Maximum Entropy modelling, Part of Speech tagging, and features like pitch and intensity. Text sentiment analysis leverages machine learning and semantic orientation. The study then introduces multimodal sentiment analysis, combining text and audio features, achieving improved accuracy. The proposed system is implemented on Hindi product reviews, demonstrating its effectiveness in capturing sentiments across different modalities.

OPEN ACCESS © CINEFORUM

Table 1: Evaluating the Strengths and Weaknesses of multimodal analysis methods in Recent Research.

S. No	**Author & Date**	**Proposed & methodology**	**Pros & Cons**
1	Rui Zhang et al. 2023[2]	Bimodal Fusion Network with Multi- Head Attention for Multimodal Sentiment Analysis	**Pros-** Multimodal Integration, Attention Mechanism, Positional Embeddings **Cons-** Complexity, Data Dependency, Generalization
2	Jiangfeng L et al. 2023 [4]	Sentiment Analysis on Online Videos by Time-Sync Comments	**Pros-** Real-Time Emotion Tracking: TSCs provide real-time emotional insights, allowing for the identification of emotional peaks and shifts in videos. **Cons-**Noise and Irrelevant Comments: TSCs can also contain irrelevant or spammy comments that may not accurately reflect the sentiment of the video segment.
3	Chuanmin Mi et al. 2022[5]	Predicting video views of web series based on comment sentiment analysis and improved stacking ensemble model	**Pros-** The study highlights the importance of word-of-mouth factors, marketing strategies, and star power in determining web series popularity. This can guide decision-makers in focusing on these critical aspects to attract a wider audience and increase viewership. **Cons-** Even with advanced predictive models, there is always a degree of uncertainty in forecasting web series popularity.
4	Arif Ullah et al. 2022.[6]	Review on sentiment analysis for text classification techniques	**Pros-** It Cleans and prepares the text data for analysis. **Cons-**It is Limited to detecting depression based on text data.

© CINEFORUM

5	Jyoti Yadav et al.2023[7]	Sentiment Analysis on social media	**Pros-** Provides valuable insights, enables real-time monitoring, and utilizes rich data sources. **Cons-**Data can be noisy, changing trends, interpreting neutral sentiments
6	Nik Nur Wahidah Nik Hashim et al. 2023[8]	Sentiment Analysis and Text Classification for Depression Detection	**Pros-**The project aims to detect depression based on text data, potentially helping individuals who may need mental health support. **Cons-**The sentiment analysis using Vader and TextBlob may not be accurate for Malay language text, potentially limiting the tool's effectiveness in non-English languages.
7	Adil Baqach et al. 2022[9]	Text-Based Sentiment Analysis	**Pros-**Deep learning models have demonstrated better performance compared to traditional machine learning algorithms, indicating the potential for more accurate sentiment analysis solutions. **Cons-**Choosing the most suitable deep learning architecture or model configuration for a specific sentiment analysis task can be complex and may require expertise in neural network design.
8	Kiran V. Sonkamble et al. 2023[1]	Sentiment Analysis Using Computer-Assisted Text Analysis Tools	**Pros-**CATA tools provide a systematic way to assess linguistic and emotional characteristics in text data, enhancing the understanding of text content. **Cons-**LIWC dictionaries are created and validated rigorously, while Empath's categories can be generated with seed words, potentially affecting accuracy.

© CINEFORUM

9	Zhi Li et al. 2022[10]	Sentiment Analysis of Danmaku Videos Based on Naïve Bayes and Sentiment Dictionary	**Pros-**The use of sentiment analysis on Danmaku reviews helps in understanding the emotional tendencies of viewers towards video content. **Cons-**Analysing danmaku data requires handling unstructured, real-time data with dynamic changes, making it complex and resource-intensive.
10	Gina Khayatun Nufus et al.2021[11]	Sentiment Analysis for Video on Demand Application User Satisfaction with Long Short-Term Memory Model	**Pros-**Companies can consider user sentiment to meet customer expectations and desires, leading to improved service offerings. **Cons-**The accuracy rate of 73.90% suggests that the model can reliably predict user sentiment, aiding in decision-making.
11	Asher Prescott et al. 2023[12]	A Multi-modal Fusion-based Sentiment Analysis Model for Short Videos	**Pros-**The combination of text and video information improves the classification of emotions in short videos, leveraging the directness of textual emotional expression. **Cons**-In some cases, incorporating textual information may lead to slight decreases in precision, recall, and values
12	Yuxin Jin et al. 2023[10]	A Review of Text Sentiment Analysis Methods and Applications	**Pros-**The effectiveness of pre-processing may vary based on the quality of the text data. **Cons-**They may not manage complex semantic and contextual relationships well.

© CINEFORUM

13	Ping He et al. 2023[3]	Cross-Modal Sentiment Analysis of Text and Video Based on Bi-GRU Cyclic Network and Correlation Enhancement	**Pros-**Provides valuable insights for cross-modal sentiment analysis in applications such as social media analysis. **Cons-**It Computational complexity increases in unaligned settings, potentially affecting real-time performance.
14	Md Abu Rumman Refatet al.2021[13]	A Nonverbal Facial Sentiment Analysis Using Convolutional Neural Network	**Pros-**Trained on three benchmark datasets for diverse expression recognition. **Cons-**Limited to recognizing seven basic facial expressions; may not manage subtle or nuanced emotions well.
15	Feiyang Chen et al. 2019[14]	Audio-Text Sentiment Analysis using Deep Robust Complementary Fusion of Multi-Features and Multi- Modalities	**Pros-**The paper addresses the zero-probability problem in sentiment lexicons by using Laplace smoothing, which can enhance the accuracy of sentiment analysis.
			Cons-The model involves various components, including sentiment analysis, stacking ensemble, and feature engineering, which may make it complex to implement and maintain.
16	Anay Ghosh et al. 2022[15]	A multimodal sentiment analysis system for recognizing person aggressiveness in pain based on textual and visual information	**Pros-** The study illustrates the overpowering advantages of the suggested strategy by contrasting the outcomes of the suggested approach with innovative methods. This plays a role in proving the developed system's efficiency. **Cons-**The potential for misinterpretation or biased predictions could have real-world consequences, emphasizing the need for thorough ethical considerations and precautions in deploying such systems, especially in sensitive domains.

© CINEFORUM

17	Songning Lai et al.2023[16]	Multimodal Sentiment Analysis	**Pros-** The paper outlines meaningful research directions, offering a roadmap for future work in multimodal sentiment analysis. This can help researchers focus on critical issues and promote collaboration in addressing these challenges. **Cons-** Balancing model performance and interpretability are crucial, and the paper does not delve deeply into strategies for achieving this balance.
18	Harika Abburi et al.2018[17]	Multimodal Sentiment Analysis Using Deep Neural Networks	**Pros-** Experimentation with different configurations, such as varying layers and nodes for DNN and mixture components for GMM. **Cons-** It lacks specific evaluation metrics, making it challenging to assess the model's performance comprehensively.

3. Methodology and Proposed Work

3.1 Methodology

The goal of this research is to unleash the incredible potential that exists within. Fundamentally, the research aims to explore the complex process of converting spoken language into written form while revealing the hidden feelings and character attributes that are incorporated into the speaker's voice. To accomplish this challenging objective, a carefully selected dataset of several audio recordings will be put together. Every entry in this repository will include both the uncut speech samples and the transcripts that go along with them, which are painstakingly written to preserve each subtlety and variation of the spoken words. Furthermore, each item will have painstaking comments affixed to it that carefully outline the speakers' rich tapestry of emotions and psychological features. But auditory analysis is just one aspect of this investigation. To gain a more complete picture, synchronised video recordings will also be included in the investigation, providing a multimodal view that considers both auditory and visual information.

The study expects to gain significant insights into the complex interactions between speech, emotion, and personality through this interdisciplinary method. We intend to illuminate the intricacies of human expression and communication via the prism of innovative speech recognition technology, opening the door for ground- breaking developments in the area.

3.1.1 Methodology for Enhancing Sentimental analysis using Multimodal data:

a: Dataset Collection and Preprocessing:

Assemble a diverse dataset of sentimental analysis, covering a spectrum of conditions and their outputs. Annotate the dataset to label regions of interest, such as voice modulation, text, video analysis ensuring a well- annotated dataset for model training. *b:*

© CINEFORUM

Dataset Collection and preparation:

To create the proposed enhancing sentimental model, we started by using a valuable collected datasets named called as the "Ryerson Audio-Visual Database of Emotional Speech and Song", FER2013 Kaggle Challenge informational index and Eliciting Personality Traits in Large Language Models.

- Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): this section comprises 1440 files, resulting from 24 actors each performing 60 trials. The recordings feature 24 professional actors, evenly split between genders, with a neutral North American accent. These actors deliver two lexically matched statements in the RAVDESS dataset, embodying a range of emotions such as peace, happiness, sadness, anger, fear, surprise, and disgust through spoken word. Each emotional expression is presented at two intensity levels - strong and normal - alongside a neutral expression.
- FER2013 Kaggle Challenge informational index: The 48x48 pixel grayscale portraits in the collection are all faces. By using automatic authorization, the images have been organised, guaranteeing that the face is in the centre and takes up the same amount of space in every picture. The purpose of this exercise is to categorise each face under one of seven categories based on the emotion portrayed in the expression: 0 represents Angry, 1 indicates Disgust, 2 indicates Fear, 3 indicates Happy, 4 indicates Sad, 5 indicates Surprise, and 6 indicates Neutral.
- Eliciting Personality Traits in Large Language Models: Text analysis is divided into two main approaches: closed and open vocabulary methods. Closed vocabulary techniques rely on predefined lists to infer personalities from online platforms. For instance, the General Inquirer model targets concepts such as power and wellbeing when predicting personalities based on Twitter data. On the other hand, open vocabulary approaches in natural language processing (NLP) utilize algorithms to generate word vectors, which identify clusters of words from the data to make predictions.

3.1.2 Model Architecture Selection:

This research will investigate the capabilities of speech recognition systems by analysing their ability to not only transcribe spoken language into text but also to infer speaker emotions and personality traits.

Data Acquisition:

- Audio data: A diverse dataset of labelled audio recordings containing speech with corresponding transcripts and emotional/personality annotations will be collected from public databases or created through controlled recordings.
- Text data: Existing text corpora may be used to train and improve language models for context understanding.
- Video data (Optional): If investigating the impact of visual cues, a subset of recordings with synchronized video data will be included.

Feature Extraction:

- **Audio:** The raw audio will be processed to extract features like Mel-Frequency Cepstral Coefficients (MFCCs) representing pitch, loudness, and timbre.
- **Text:** Words will be converted into numerical representations (embeddings) capturing their meaning and

© CINEFORUM

relationships.

- **Video:** Convolutional Neural Networks (CNN's) will be employed to extract visual features such as facial expressions and mouth movements.

Modelling and Analysis:

- **Acoustic Modelling:** Statistical models will be trained on the extracted audio features to identify the sequence of sounds (phonemes or words) corresponding to the spoken input.
- **Language Modelling:** Language models, potentially using architectures like LSTMs or Transformers, will be trained on existing text data to understand the context and predict the most probable word sequence based on the recognized words and overall language structure.
- **Emotion/Personality Recognition:** Additional models will be implemented to analyse extracted features (audio and/or visual) and predict the speaker's emotions or personality traits. This may involve training specific classifiers or employing multi-modal learning approaches.

a: Implementation:

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 challenging task as the machine cannot understand humour, 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 LSTMs, and Transformers. I used Bidirectional-LSTMs 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, 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 textbox and were also given an option of uploading the Cover-Letter that can also be used to predict the individual's Personality Traits.

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 managing the audio in small chunks is a significant improvement for the predictions. Literature is centred on just around six feelings., happy, sad, angry, disgusted, fear, and surprise. Albeit the feeling classifications are more plentiful and complex.

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.

© CINEFORUM

3. After all this MFCCs 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.

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 Exception transfer learning model.

1. First, 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 in those frames.
4. The image is flattened and fed to the Exception Model for an output.

3.2 Proposed Model

Our point is to foster a model ready to furnish live sentiment with a visual UI utilizing Tensorflow and JavaScript 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 analyse 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 Centres.

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.

3.2.1 DATASET SOURCES

3.2.1 Text: The text input we're utilizing originates from a study conducted by King and Pennebaker [19]. This dataset comprises 2,468 daily writing submissions provided by 34 psychology scholars, consisting of five men and 29 women aged between 18 and 67 years.

3.2.2 Audio: To gather sound information, we're utilizing the "Ryerson Audio-Visual Database of Emotional Speech and Song" (RAVDESS). This database encompasses 7,356 voice clips, with a total size of 24.8 GB. These recordings consist of 24 audio clips, evenly split between 12 females and 12 males, each presenting two lexically coordinated statements in a neutral North American accent. The speech recordings include expressions of calm, happiness, sadness, anger, fear, surprise, and disgust, while the song recordings depict emotions of calm, happiness, sadness, and sorrow.

3.2.3 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.

3.3 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.

3.3.1 Text Pre-processing:

The first stage of our NLP pipeline is pre-processing. This is where we translate impure content recordings into neat word groupings. Tokenizing the corpus of data is a prerequisite for completing this interaction. This suggests that sentences are divided into a list of single words, also referred to as

© CINEFORUM

tokens. Additional pre-
processing procedures include employing standard expressions to ask for the removal of unwanted characters or the reformatting of comments. Finally, methods for replacing words through their linguistic root are available: the goal of lemmatization and stemming is to reduce derivatively related word types to a standard base structure.

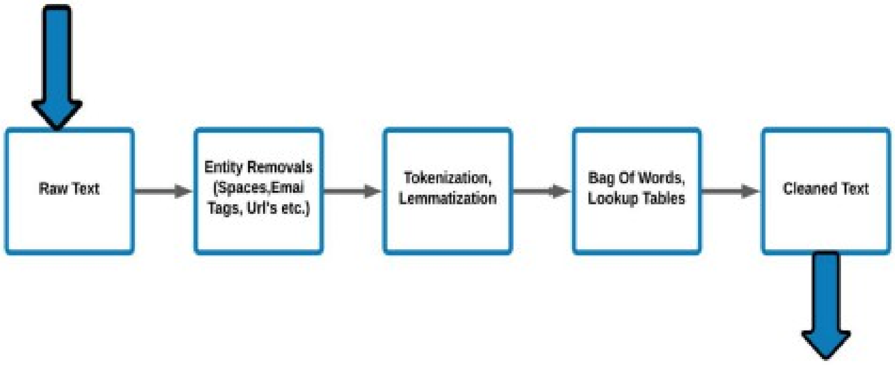

Raw Text
Entity Removals (Spaces,Emai Tags, Url's etc.)
Tokenization, Lemmatization
Bag Of Words, Lookup Tables
Cleaned Text

Fig 1: Text Cleaning Pipeline
Fig 1 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.

3.3.2 Audio Pre-Processing:

To begin with, before starting feature extractions, it is 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.

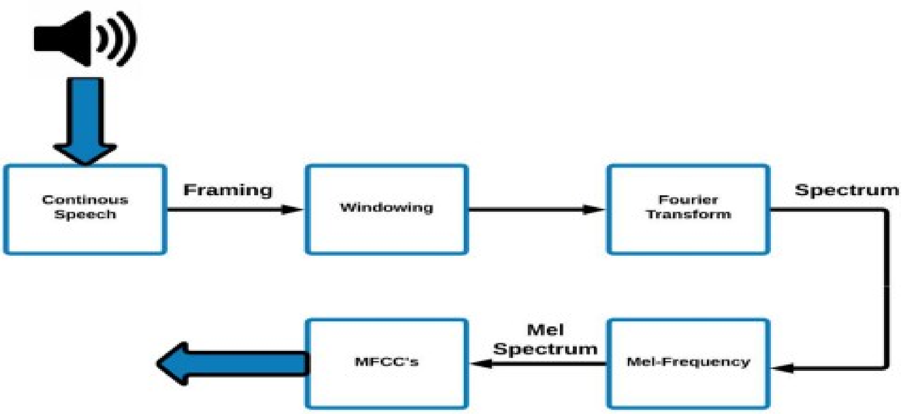


Fig2: Audio Preprocessing
Fig2 explains the Audio Cleaning and conversion of those into the MFCCs that will be used as the input for the model and is used for training and testing purposes.

3.3.3 Video Pre-Processing:

Starting by analysing the video frame by frame, then applying filters using some of the convolution

© CINEFORUM

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.

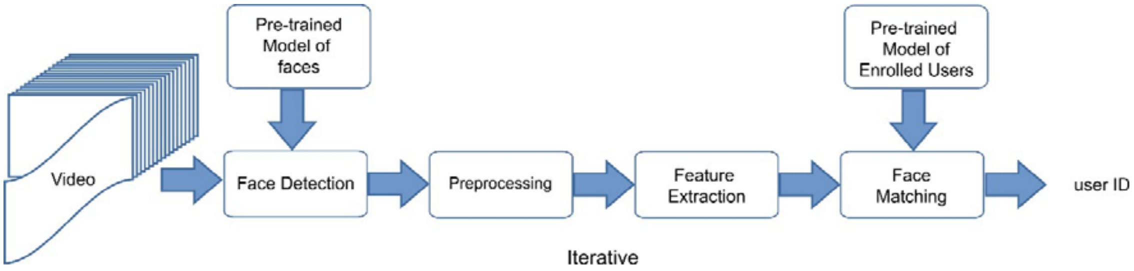
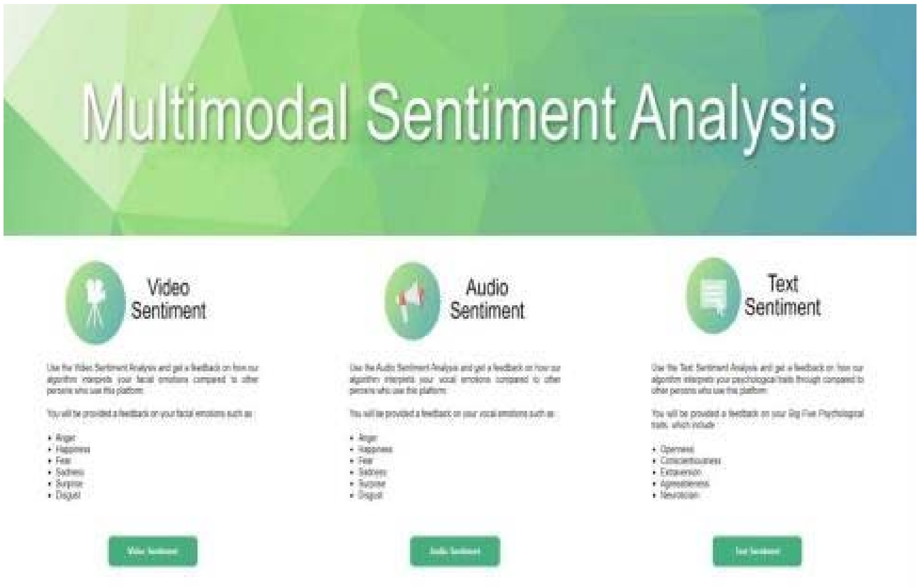


Fig 3:Video Pre-Processing

Fig3 In proposed automatic face recognition system comprises four key stages. Firstly, in the face detection stage, one or more faces are identified within the image or video frame. Subsequently, in the preprocessing stage, various image processing techniques are employed to optimize the face image for subsequent analysis using the machine learning model. Following this, in the feature extraction stage, distinctive facial features are isolated from the prepared face image. Lastly, in the identification stage, the extracted facial features are compared against the known facial features of enrolled users to ascertain person identification

4. RESULT AND COMPARISONS

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.



Video Sentiment

Use the Video Sentiment Analysis and get a feedback on how our algorithm interprets your facial emotions compared to other persons who use this platform.

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

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

Audio Sentiment

Use the Audio Sentiment Analysis and get a feedback on how our algorithm interprets your vocal emotions compared to other persons who use this platform.

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

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

Text Sentiment

Use the Text Sentiment Analysis and get a feedback on how our algorithm interprets your psychological traits through compared to other persons who use this platform.

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

- Openness
- Conscientiousness
- Extraversion
- Agreeableness
- Neuroticism

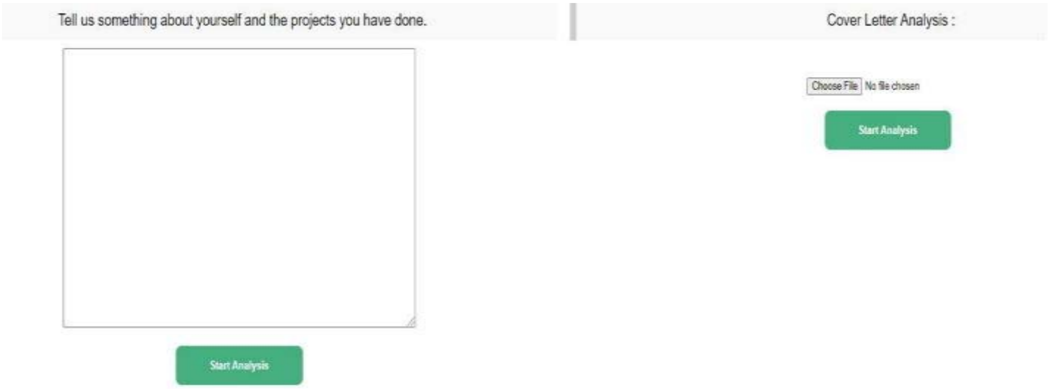
© CINEFORUM

Fig 4:Home Page

Fig 4 shows the Home Page of the deployed model in the local server.

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.

4.1 Text Sentiment Analysis: In this Text Modal, we have implemented Text Analysis for predicting the Personality Traits in a human being used for interview simulation. 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. 14 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.



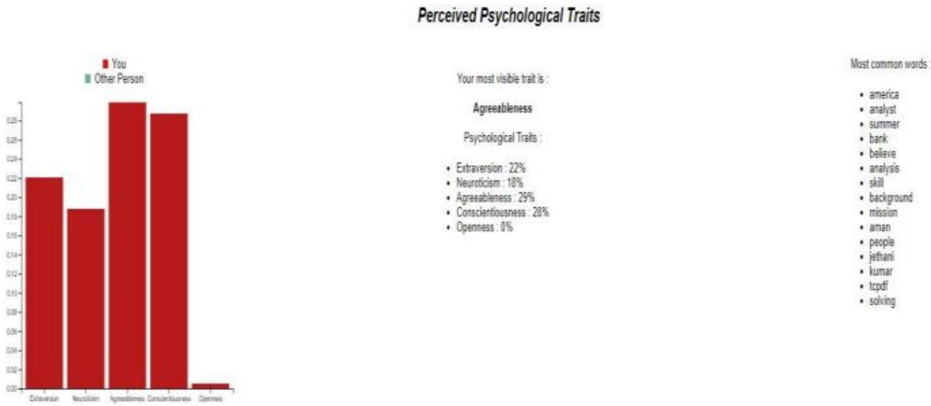
Tell us something about yourself and the projects you have done.
Start Analysis
Cover Letter Analysis :
Choose File No file chosen
Start Analysis

Fig. 5: Text Sentiment Home-Page

Fig 5 shows the two methods we can use in the Text-Sentiment, i.e., Text and Cover Letter upload. Compared with the other candidates, the output bar plots are displayed, and the most frequently used 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:

© CINEFORUM



Perceived Psychological Traits

Your most visible trait is :

Agreeableness

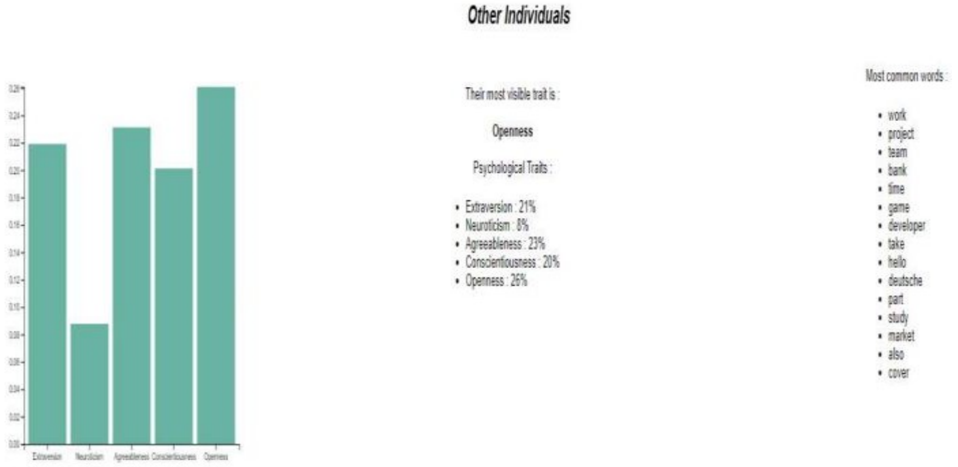
Psychological Traits :

- Extraversion : 22%
- Neuroticism : 18%
- Agreeableness : 29%
- Conscientiousness : 28%
- Openness : 0%

Most common words :

- america
- analyst
- summer
- bank
- believe
- analysis
- skill
- background
- mission
- aman
- people
- jethani
- kumar
- tcpdf
- solving

Fig. 6: Probability Bar Plot for Our Input Text



Other Individuals

Their most visible trait is :

Openness

Psychological Traits :

- Extraversion : 21%
- Neuroticism : 8%
- Agreeableness : 23%
- Conscientiousness : 20%
- Openness : 26%

Most common words :

- work
- project
- team
- bank
- time
- game
- developer
- take
- hello
- deutsche
- part
- study
- market
- also
- cover

Fig. 7: Probability Bar Plot for Other Individuals

Figs 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 different models is shown below. The method that has been used is Word-2-Vec embedding with LSTM and SVM models. The accuracy of the test set is shown below.

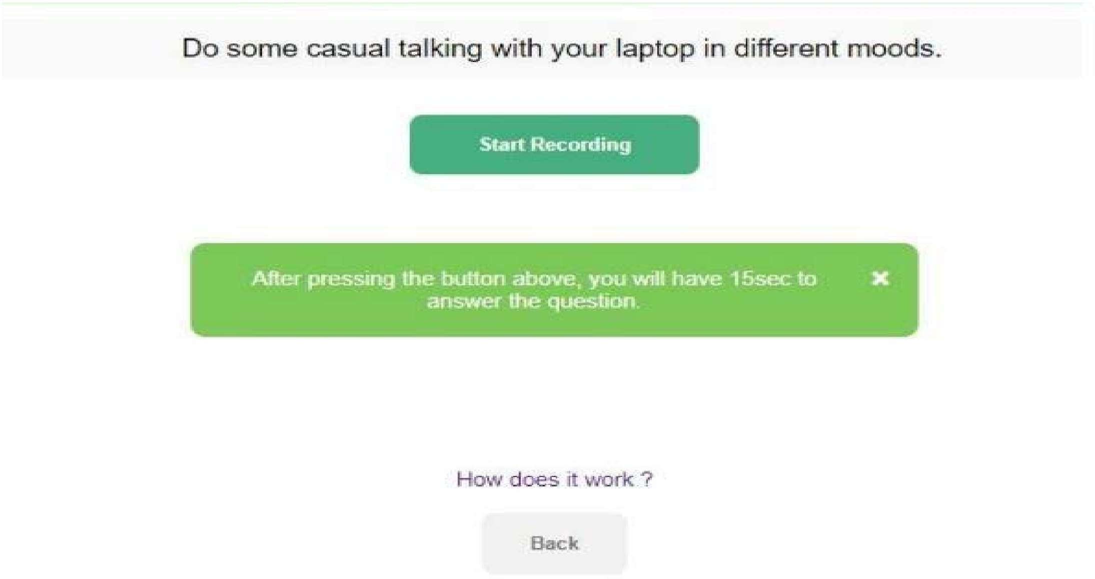
Model	**EXT**	**NEU**	**AGR**	**CON**	**OPN**
Word2Vec+SVM	**46.18**	**48.21**	**49.65**	**49.97**	**50.07**
Word2Vec+LSTM	**55.07**	**50.17**	**54.57**	**53.23**	**53.84**

© CINEFORUM

Table 2: Text Accuracy Confusion Matrix

Table 2 shows the accuracy of labels with two diverse 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.

4.2 Audio Sentiment Analysis: 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-spectrograms 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 has been used "Ryerson Audio-Visual Database of Emotional Speech and Song"(RAVDESS) dataset for training and testing purposes.



Do some casual talking with your laptop in different moods.

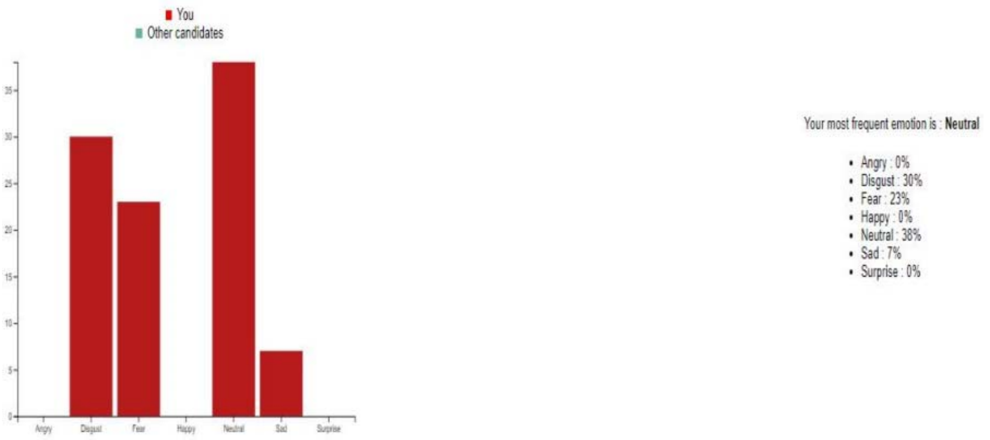
After pressing the button above, you will have 15sec to answer the question.

Fig 8: Audio Sentiment Home Page

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.

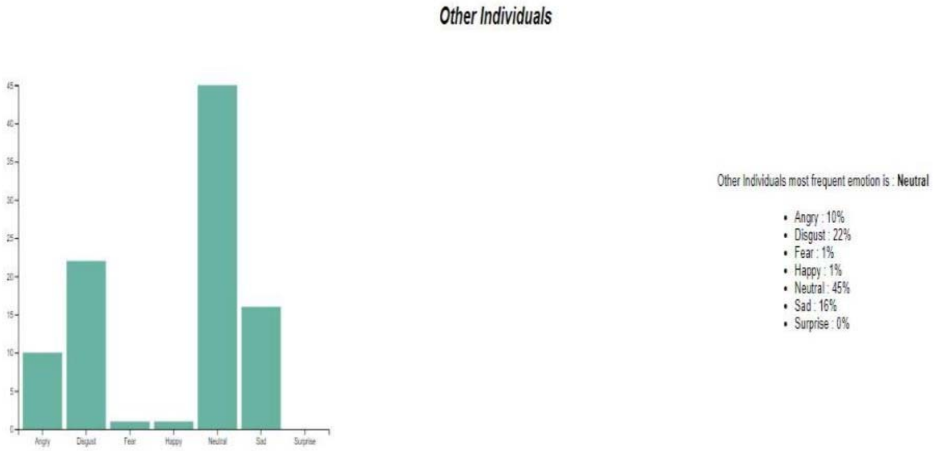
© CINEFORUM



Your most frequent emotion is : **Neutral**

- Angry : 0%
- Disgust : 30%
- Fear : 23%
- Happy : 0%
- Neutral : 38%
- Sad : 7%
- Surprise : 0%

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



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

- Angry : 10%
- Disgust : 22%
- Fear : 1%
- Happy : 1%
- Neutral : 45%
- Sad : 16%
- Surprise : 0%

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

Figs 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.

© CINEFORUM

	Predicted labels							
		Happy	Sad	Angry	Scared	Neutral	Disgusted	Surprised
Actual Labels	Happy	83.0%	0.0%	5.0%	4.3%	10.2%	2.9%	3.4%
	Sad	5.1%	71.1%	0.0%	0.0%	3.7%	10.1%	1.5%
	Angry	6.3%	6.3%	85%	0.0%	6.3%	9.3%	0%
	Scared	6.7%	6.2%	3.4%	81.1%	8.9%	9.0%	4.7%
	Neutral	11.1%	5.6%	0.0%	3.2%	89.7%	5.6%	0.3%
	Disgusted	0.0%	8.7%	0.0%	4.3%	2.2%	84.8%	2.9%
	Surprised	9.4%	7.7%	0.0%	4.3%	2.9%	86.8%	67.3%
	Sad	8.1%	81.1%	0.0%	0.0%	2.7%	8.1%	1.5%

Table 3 : Audio Accuracy Confusion Matrix

Table 3 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.

OPEN ACCESS © CINEFORUM

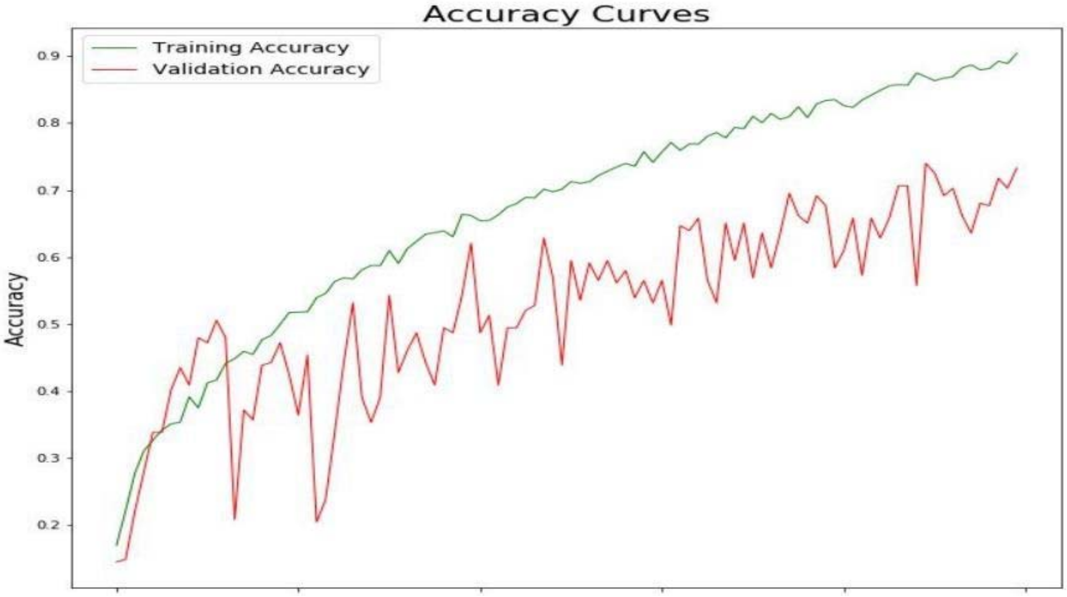


Fig. 11: Audio Sentiment Accuracy Curve

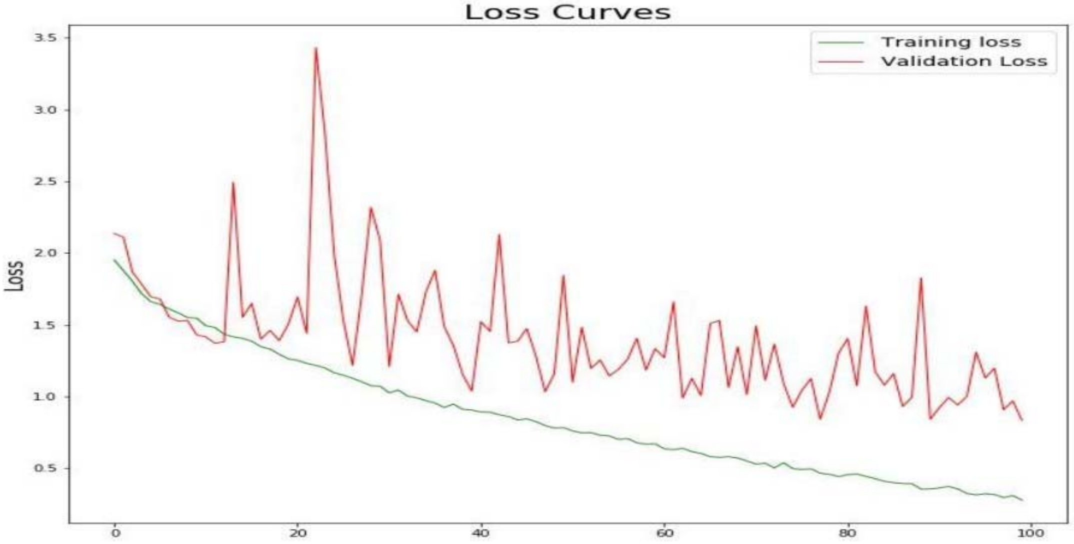


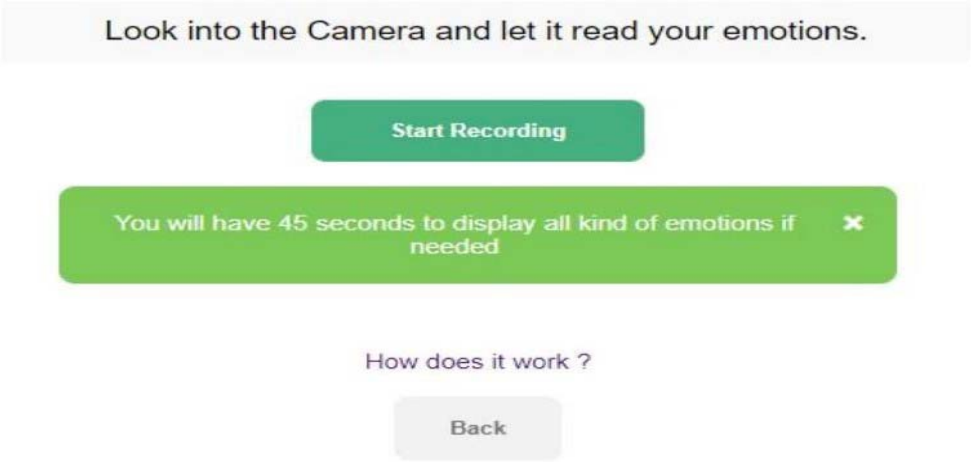
Fig. 12: Audio Sentiment Loss Curve

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

4.3 Video Sentiment Analysis: In this Video Model, 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. 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

© CINEFORUM

Kaggle Challenge dataset for training and testing purposes.



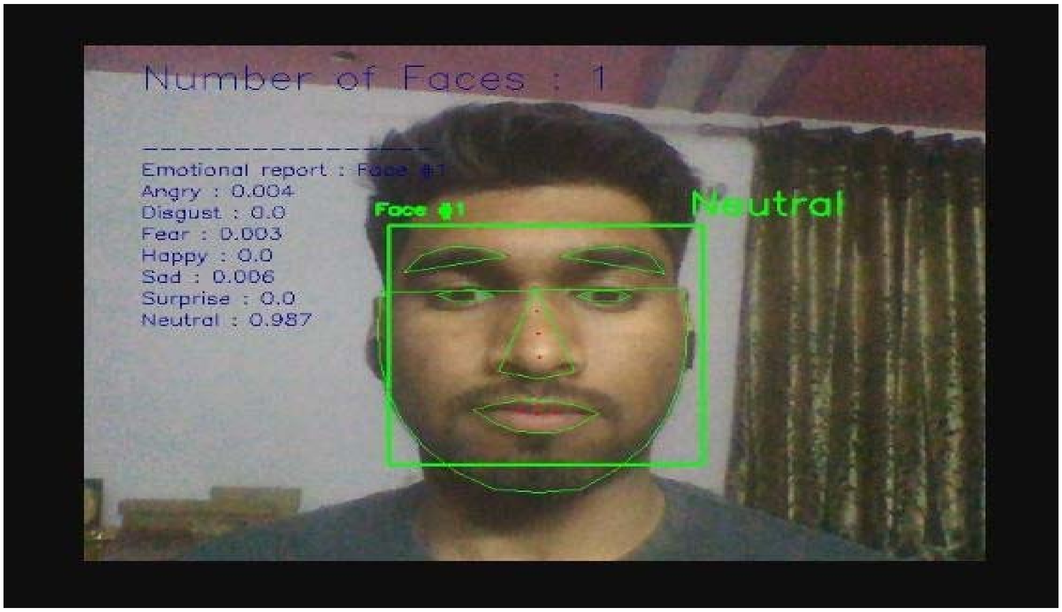
Look into the Camera and let it read your emotions.

You will have 45 seconds to display all kind of emotions if needed

Fig. 13: Video Sentiment Home-Page

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

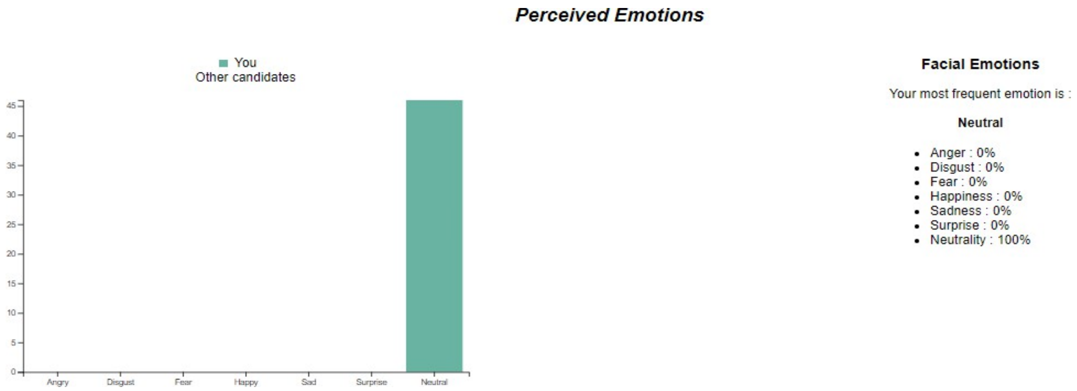


```
Number of Faces : 1

----------------
Emotional report : Face #1
Angry : 0.004
Disgust : 0.0
Fear : 0.003
Happy : 0.0
Sad : 0.006
Surprise : 0.0
Neutral : 0.987
```

Fig. 14: Emotion Detected(Neutral)

© CINEFORUM



Facial Emotions

Your most frequent emotion is :

Neutral

- Anger : 0%
- Disgust : 0%
- Fear : 0%
- Happiness : 0%
- Sadness : 0%
- Surprise : 0%
- Neutrality : 100%

Fig. 15: Probability Bar Plot for Our Input Live Video

The figure 14 and 15 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.

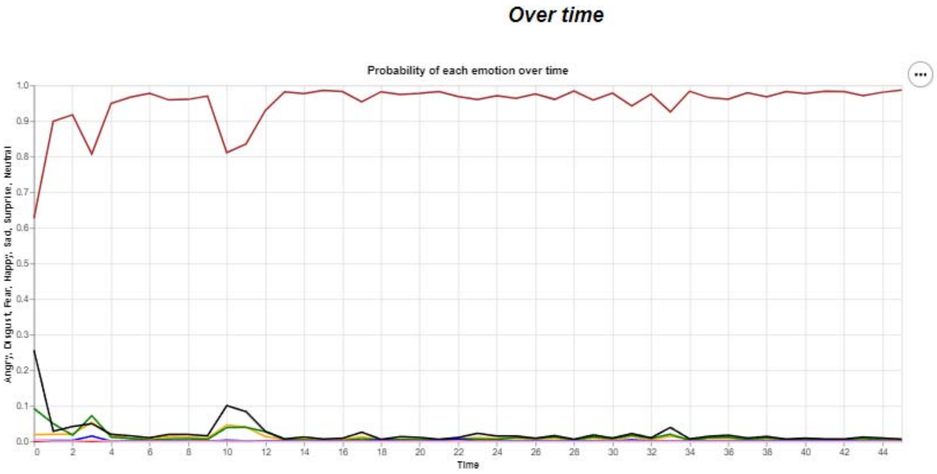
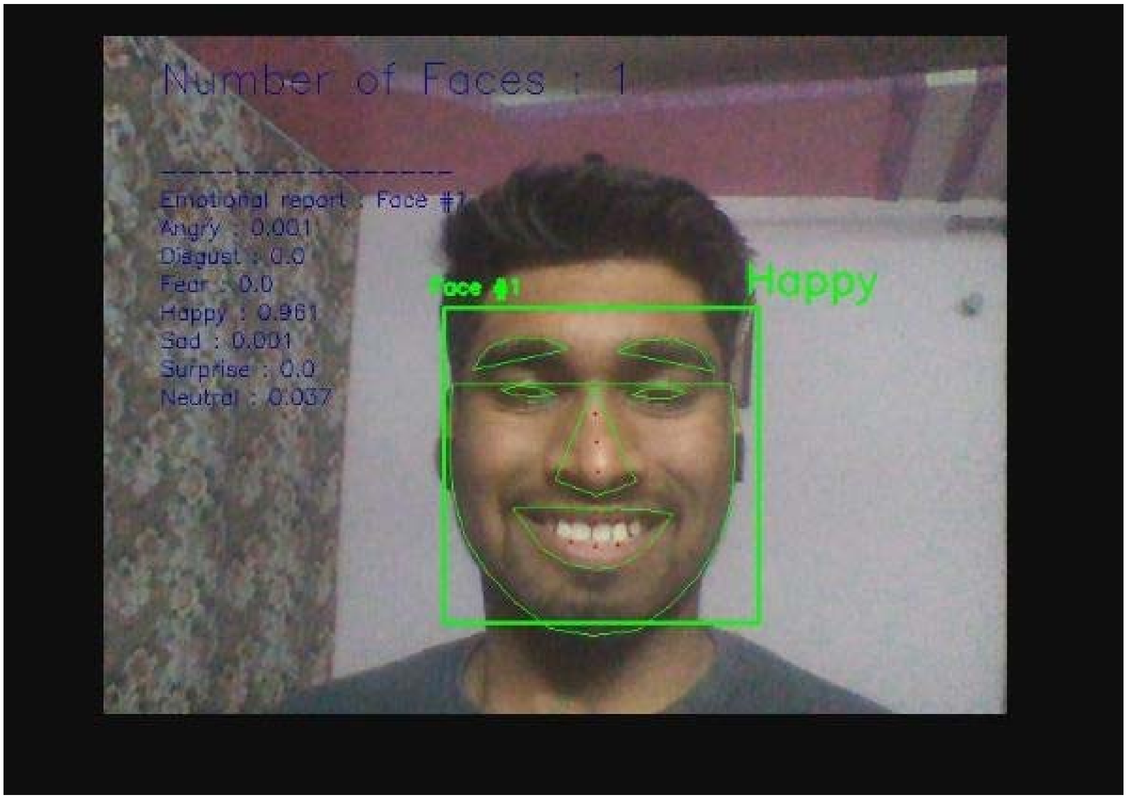


Fig. 16: Line Chart for Varying Emotions

Fig 16 shows us how our emotions vary concerning the time using a line chart that can be used eventually to get the mean Sentiment.

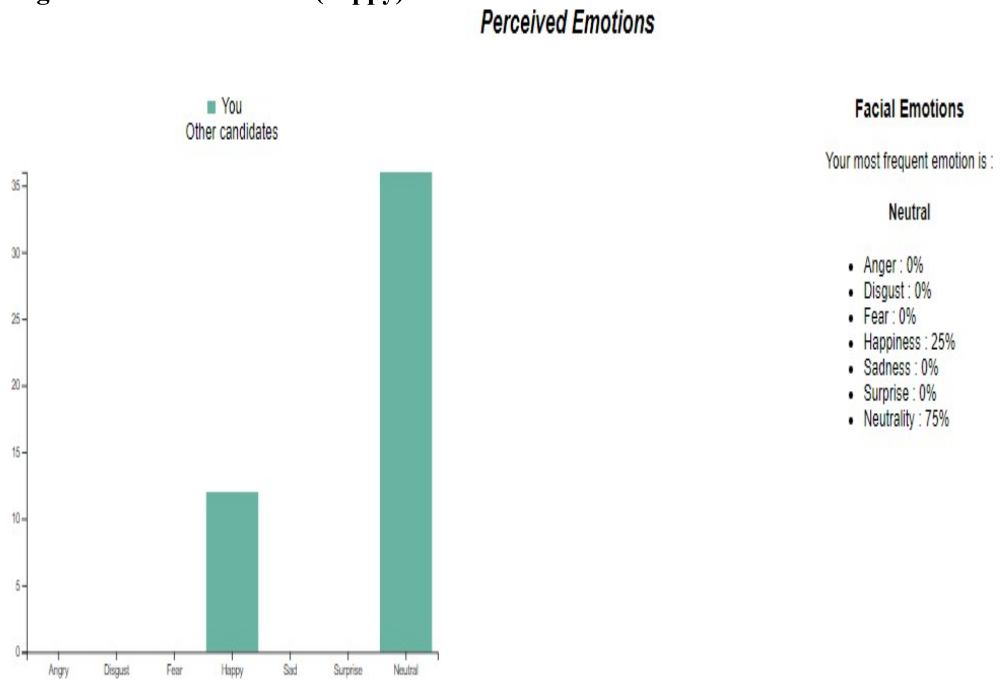
© CINEFORUM



```
Number of Faces : 1

-----------------
Emotional report : Face #1
Angry : 0.001
Disgust : 0.0
Fear : 0.0
Happy : 0.961
Sad : 0.001
Surprise : 0.0
Neutral : 0.037
```

Fig. 17: Emotion Detected(happy)



Perceived Emotions

Facial Emotions

Your most frequent emotion is :

Neutral

- Anger : 0%
- Disgust : 0%
- Fear : 0%
- Happiness : 25%
- Sadness : 0%
- Surprise : 0%
- Neutrality : 75%

Fig. 18: Probability Bar Plot for Our Input Live Video

© CINEFORUM

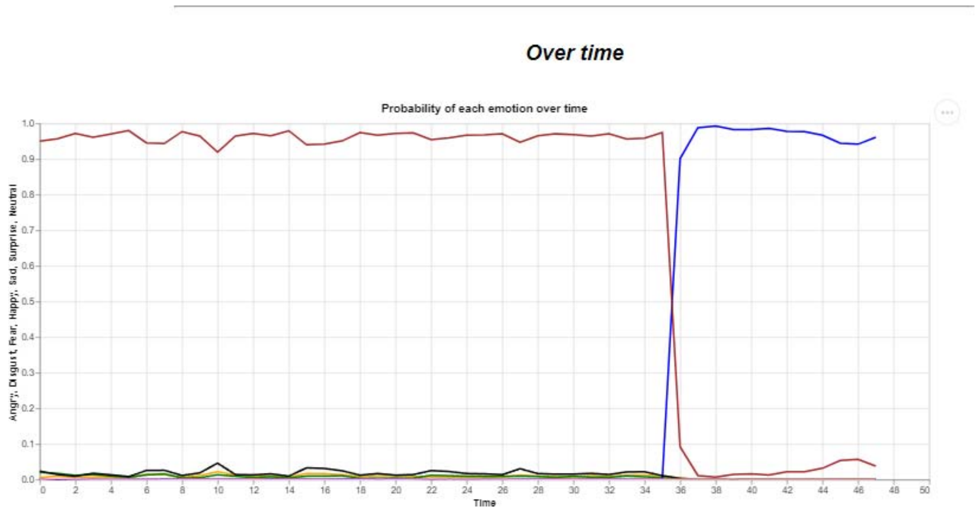


Fig. 19: Line Chart for Varying Emotions

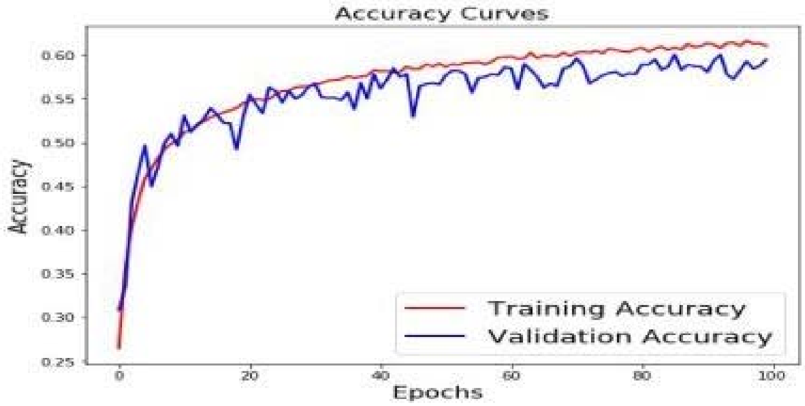


Fig. 21: Xception Accuracy Graph

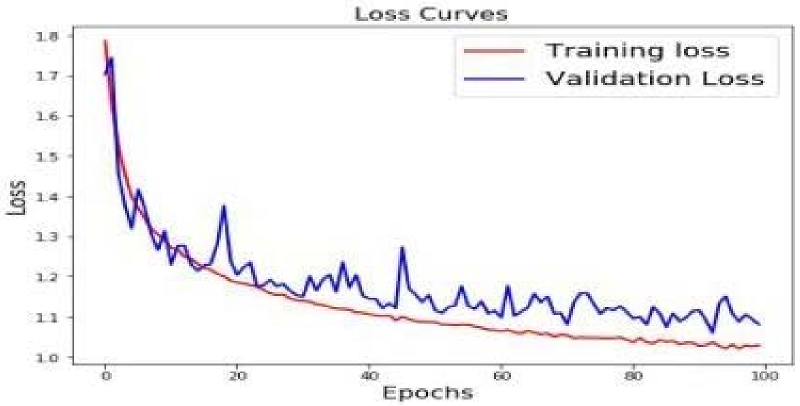


Fig. 21: Xception Loss Graph

© CINEFORUM

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 different lengths of videos like 15sec, 30sec, 40sec, but there was no significant impact on the accuracy, so we only implemented it on 45sec.

5. CONCLUSION

This paper examines current developments in multimodal sentiment analysis. We go over the most widely used feature extraction techniques and datasets in the industry. The summaries and categories of recently published and referenced publications are based on the method of fusion. The paper goes on to address potential applications as well as the difficulties that current approaches confront. Our study of the body of research demon-states that multimodal sentiment analysis, which frequently outperforms unimodal techniques, is a potential strategy for using complementary information channels for sentiment analysis. It may also improve other instruments like subjectivity analysis and entity recognition that now gain from unimodal sentiment analysis. It is our hope that this study will stimulate more multidisciplinary work in this field.

6. FUTURE SCOPE

Future research should go on the topic of understanding emotion in dialogue. One person's expression of emotion influences the others in a discourse. Related research has shown that discourse context is useful for com- prehending human language, and multimodal sentiment research will advance significantly if multimodal systems can replicate human sentiment dependencies. For models to generalise to any language in prediction tasks, further work must be done to make them language independent.

REFERENCES

[1] S. S. Date, K. V. Sonkamble, and S. N. Deshmukh, "Sentiment Analysis Using Computer-Assisted Text Analysis Tools," 2023, pp. 671–679. doi: 10.2991/978- 94-6463-136-4_58.

[2] R. Zhang, C. Xue, Q. Qi, L. Lin, J. Zhang, and L. Zhang, "Bimodal Fusion Network with Multi-Head Attention for Multimodal Sentiment Analysis," Applied Sciences (Switzerland), vol. 13, no. 3, Feb. 2023, doi: 10.3390/app13031915.

[3] P. He, H. Qi, S. Wang, and J. Cang, "Cross-Modal Sentiment Analysis of Text and Video Based on Bi-GRU Cyclic Network and Correlation Enhancement," Applied Sciences (Switzerland), vol. 13, no. 13, Jul. 2023, doi: 10.3390/app13137489.

[4] J. Li, Z. Li, X. Ma, Q. Zhao, C. Zhang, and G. Yu, "Sentiment Analysis on Online Videos by Time-Sync Comments," Entropy, vol. 25, no. 7, p. 1016, Jul. 2023, doi: 10.3390/e25071016.

[5] C. Mi, M. Li, and A. F. Wulandari, "Predicting video views of web series based on comment sentiment analysis and improved stacking ensemble model," Electronic Commerce Research, 2022, doi: 10.1007/s10660-022-09642- 9.

[6] A. Ullah, S. N. Khan, and N. M. Nawi, "Review on sentiment analysis for text classification techniques from 2010 to 2021," Multimed Tools Appl, vol. 82, no. 6, pp. 8137–8193, Mar. 2023, doi: 10.1007/s11042-022-14112-
3.

[7] J. Yadav, "Sentiment Analysis on Social Media," Qeios, Jan. 2023, doi: 10.32388/yf9x04.

[8] I. Nadhirah Joharee, N. Nur Wahidah Nik Hashim, and N. Syahirah Mohd Shah, "Sentiment

© CINEFORUM

Analysis and Text Classification for Depression Detection," Journal of Integrated and Advanced Engineering (JIAE), vol. 3, no. 1,
pp. 65–78, 2023, doi: 10.51662/jiae.v3i2.86.

[9] A. Baqach and A. Battou, "Text-Based Sentiment Analysis," in Lecture Notes in Networks and Systems, Springer Science and Business Media Deutschland GmbH, 2023, pp. 106–121. doi: 10.1007/978-3-031-26384- 2_10.

[10] Z. Li, R. Li, and G. Jin, "Sentiment analysis of danmaku videos based on naïve bayes and sentiment dictionary," IEEE Access, vol. 8, pp. 75073–75084, 2020, doi: 10.1109/ACCESS.2020.2986582.

[11] G. K. Nufus, M. Mustafid, and dan R. Gernowo, "Sentiment Analysis for Video on Demand Application User Satisfaction with Long Short Term Memory Model," in E3S Web of Conferences, EDP Sciences, Nov. 2021. doi: 10.1051/e3sconf/202131705031.

[12] A. Prescott, S. Callahan, M. Harper, and J. Throne, "A Multi-modal Fusion-based Sentiment Analysis Model for Short Videos," 2023, doi: 10.21203/rs.3.rs2997353/v1.

[13] M. A. R. Refat, B. C. Singh, and M. M. Rahman, "SentiNet: A Nonverbal Facial Sentiment Analysis Using Convolutional Neural Network," Intern J Pattern Recognit Artif Intell, vol. 36, no. 4, Mar. 2022, doi: 10.1142/S0218001422560079.

[14] E. Chu and D. Roy, "Audio-Visual Sentiment Analysis for Learning Emotional Arcs in Movies," Dec. 2017, [Online]. Available: http://arxiv.org/abs/1712.02896

[15] A. Ghosh, B. C. Dhara, C. Pero, and S. Umer, "A multimodal sentiment analysis system for recognizing person aggressiveness in pain based on textual and visual information," J Ambient Intell Humanize Comput, vol. 14, no. 4, pp. 4489–4501, Apr. 2023, doi: 10.1007/s12652-023-04567-z. 26.

[16] S. Lai, X. Hu, H. Xu, Z. Ren, and Z. Liu, "Multimodal Sentiment Analysis: A Survey," May 2023, [Online]. Available: http://arxiv.org/abs/2305.07611

[17] H. Abburi, R. Prasath, M. Shrivastava, and S. V. Ganga Shetty, "Multimodal sentiment analysis using deep neural networks," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Springer Verlag, 2017, pp. 58–65. doi: 10.1007/978-3- 319- 58130-9_6.

[18] Deng, Yuan Fei. (2023). Research on sentiment analysis methods for text-oriented data. Frontiers in Computing and Intelligent Systems. 3. 42-47. 10.54097/fcis. v3i1.6022 DOI:10.54097/fcis.v3i1.6022

[19] Si, Hongying & Wei, Xianyong. (2023). Sentiment Analysis of Social Network Comment Text Based on LSTM and Bert. Journal of Circuits, Systems and Computers.10.1142/S0218126623502924. https://doi.org/10.1142/S0218126623502924.

[20] Jin, Yuxin & Cheng, Kui & Wang, Xinjie & Cai, Lecai+++++++++. (2023). A Review of Text Sentiment Analysis Methods and Applications. Frontiers in Business, Economics and Management. 10. 58-64. 10.54097/fbem.v10i1.10171.

[21] Chen, Rongfei & Zhou, Wenju &li, yang & Zhou, Huiyu. (2022). Video-Based Cross-Modal Auxiliary Network for Multimodal Sentiment Analysis. IEEE Transactions on Circuits and Systems for Video Technology

© CINEFORUM