



# Real time sentiment analysis of natural language using multimedia input

Rishit Jain<sup>1</sup> · Revant Singh Rai<sup>1</sup> · Sajal Jain<sup>1</sup> · Ruchir Ahluwalia<sup>1</sup> · Jyoti Gupta<sup>1</sup>

Received: 21 March 2022 / Revised: 19 September 2022 / Accepted: 30 March 2023 /

Published online: 14 April 2023

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2023

## Abstract

Semantics and Sentiments are parts of our daily speech and expressions that helps to convey the message in the tone intended. The accurate interpretation of emotions and actions is prudent as it expresses the true meaning of the message. This interpretation has been studied extensively in the past two decades, where professionals from various disciplines have pondered this question. Every action and expression—whether it's in a speech, in a video or through some written material—helps the recipient understand the intent behind the message. The primary motive in these studies has been to automate the analysis of these sentiments by teaching the computers to do so, using the audio, video and text-based data that has been collected so far. Machine Learning (ML) and Deep Learning (DL) is the discipline that can help us tackle such a problem which requires analysis and recognition of copious amounts of data. Classification based on these multi-media inputs has seen the application of several common and uncommon ML techniques such as Support Vector Machines (SVMs), Bayesian Networks (BNs), Decision Trees (DTs), Convolutional Neural Networks (CNNs) and K-Means Clustering. These techniques, to a certain level of accuracy, can classify a certain part of a message into a different emotion. Through this research, firstly, a comparison is represented between the previously conducted studies and secondly, a system is developed of our own that enables Real Time Sentiment Analysis and helps a user assess his/her day-to-day attitude and get appropriate recommendations for the same.

**Keywords** Sentiment analysis · Chatbot · Audio · Video and text analysis · Machine learning · Natural language processing

## 1 Introduction

In today's day and age, the existence of popular social media platforms such as – Instagram, Twitter or Whatsapp—provide us with a fighting chance against the menace of Stress and Anxiety, by raising awareness and sharing resources to help those in need. Social media

---

✉ Jyoti Gupta  
jyoti.gupta@bharatividyaapeeth.edu

Extended author information available on the last page of the article

directly affects the mental health of a person [12]. Studies have conclusively proved that along with physical fitness, mental fitness is just as crucial for the healthy operation of a human body, hence, equally focused treatment for those who feel weak or mentally distressed is the need of the hour. The COVID-19 pandemic has taken a serious toll on mental health with people forced to be confined in their home, cut off from the world and normal interactions. Thus, there is a growing need to find ways to easily identify and prevent mental health issues along with increasing access to mental health services [24]. First aid for mental health is not very popular and developed as compared to physical health. Providing mental health first aid can help ease the symptoms a person may be experiencing [17].

To facilitate these issues, this project was taken on in order to create a platform that would help people assess their condition and mental health more extensively and take any necessary precautions if warranted. Such a platform would not only provide people with an efficient platform to conduct precursory psychiatric diagnostics, but it would also serve a big role in raising awareness amongst the people. The platform will enable this via sentiment analysis using audio and video. Analysis based on audio or video alone is not sufficient since a human expresses himself not just through words but through his facial expressions and body language. Face often tells a different story than words. By listening to a person without looking at them one can technically understand them, but he cannot gauge their feelings. Hence, in this platform, a person would be required to answer a set of questions and their response would be used to analyse their immediate mood and emotions. The audio would be converted to text and then processed to perform sentiment analysis to categorize the mood throughout the session. Alongside this, OpenCV can be used to detect facial emotions through facial recognition. Combining both the results would give us a report of the person's state of mind which can be used for further diagnosis.

The motivation of this paper is to conduct a thorough research of the different studies conducted for the discipline of sentiment analysis based on audio, video and text input. The premise of this paper was based on research work titled—'Psychiatric Precursory Diagnostic Test'—conducted as part of a college thesis, wherein a cursory test was created from the perspective of a counsellor who can assess the day-to-day mood and attitude of a user through targeted questions, and based on the results produced, the system recommends appropriate activities to help the user. In the interest of covering all bases, this study contains an outlook from a technological and psychological point of view.

This research was predicated on a simple idea to digitise the prognosis conducted by a psychiatric professional and make that technology easily available to users. The stigma around mental health is a major issue preventing people from seeking the help they require [16]. The challenges faced while getting mental health services are alone enough for a person to disregard their health. The aim of this project is to provide a reliable resource which a person feels comfortable using in their day to day life without facing the stigma. While replicating the exact behaviour and capabilities of a professional might not be prudent, the primary objective is to develop a model that can at the very least, perform accurate sentiment analysis and depict a user's mood. Much like the analysis conducted by any psychiatrist who converses with the user and uses his behaviour to draw conclusions, our system uses speech, text-based and video input to identify the user's mood and accordingly suggest activities for the user to indulge. The research stems from this project, wherein the authors have conducted a study of the extensive studies performed over the past few decades and draw out the best and the most effective machine learning models that would facilitate the entire project.

This paper is structured in sections so as to give us an ordered manner of information. Section 1 informs us about the dataset inculcated to train the Sentiment Analysis model

and the chatbot model. Succeeding the datasets is Sect. 2 comprising of the diligent Literature Review done by various authors in the field of Sentiment Analysis and their contrasts in work have been presented. Following that is Sect. 3 that encompasses Methodology of the model. It encapsulates all the specific details about the methods, functions and libraries used for the different models used in the project. Following that is Sect. 4 of the paper that presents us with the various findings, results and observations gathered through this project. Section 5 finally concludes our project and the research conducted for it. Penultimately, the last section of the paper contains all the references and citations to previous studies.

## 2 Datasets

Through this study, a myriad of different techniques have been explored in order to learn and identify the techniques that best cater to the domain of sentiment and semantics analysis. In this pursuit, the authors developed certain unique own models that yielded remarkable accuracies based on multimedia inputs. To train these models, certain datasets were used that could ensure a plethora of different options for the holistic coverage of most bases in the model. For textual sentiment analysis that takes the input from a user in the form of text and takes the speech to text input, the Amazon Reviews dataset was employed. Meanwhile for the chatbot model, a custom dataset was created in the.json format consisting of several relevant tags and responses.

### 2.1 Text sentiment analysis

Amazon Customer Reviews is one of Amazon's rich databases containing over hundreds of reviews over 2 decades with customers expressing their views on products. This makes Amazon Customer Reviews a rich source of information for academic researchers in the fields of Natural Language Processing (NLP), Information Retrieval (IR), and Machine Learning (ML), amongst others.

The dataset contains the customer review text with accompanying metadata, consisting of three major components:

- Reviews intended into the properties (and the evolution) of customer reviews potentially including how people evaluate and express their experiences with respect to products at scale.
- Reviews in multiple languages to analyse perception of the same product across different languages and countries.
- Non-compliant reviews with respect to Amazon policies helping flag unwanted and biased reviews.

### 2.2 Chatbot – custom dataset

A sophisticated chatbot was developed which is capable of carrying out intelligent conversations with a user. To achieve this ability, we created a custom dataset using a.json file. It contained 9 primary tags ("Welcome", "Exit", "Thanks", "Incorrectinput", "Inspiration", "Support", "Sad", "Happy", "Joke") which enabled the grouping of distinct types of inputs given by the user to which the bot could appropriately reply. The input given by the users

were defined as patterns and the response given by our bot was defined as responses. With this dataset, chatbot was trained appropriately to our customizations, in order to give our users an interactive and satisfied experience.

### 3 Literature review

The 21<sup>st</sup> Century marked the advent of the digital age that has caught an unparalleled pace in the first two decades, wherein advancements in technology have been made that cater to eradicate most of our problems. People around the globe hold phones in the palm of their hands that facilitate communication. Nowadays, we have access to the world wide web that answers all of our questions, we have the fastest vehicles at our disposal to facilitate travel. Machines are growing smarter by the day in order to cater to us humans, and in fact make our lives easier. The field of teaching computers to perform certain tasks using previously created data, is known as Machine Learning. One major sub-discipline of this field is that of Sentiment Analysis, wherein a machine is taught to study and recognise the different human emotions. This task has been achieved through proper analysis of multimedia inputs such as – Text, Audio or Video.

Over the past few decades, meticulous analysis and feature extraction has been conducted on multimedia inputs to classify them on the basis of the sentiment or emotion exhibited. A cogent and sound analysis of these studies would present a distinguished comparison between the different techniques that have been used by professionals across the globe. One such comparison is projected in Table 1 drawn below, where different models employed for video, audio, and text-based sentiment analysis were examined.

In the extensive study conducted by Rao et al. [22], techniques such as Support Vector Machines (SVMs), Decision Trees and OpenCV were employed for Text, Audio, and Video based inputs, respectively. Through this study, they obtained an accuracy of about 70% in identifying a total of 6 emotions. Secondly, in the paper by Li et al. [14], an SVM model was used for sentiment analysis through video-based input, on the MOUD dataset [21] and CMU-MOSI [29] dataset. For the datasets in consideration, they attained accuracies of 63.9% and 71.1% respectively. In the paper written by Schmidt et al. [25] sentiment analysis is conducted on the textual and audio version (audiobook) of the historic German plays, where Emilia Galotti by G.E. Lessing was taken as the base case. This study employed the Naïve Lexicon method and the free Vokatari tool for Text and audio-based analysis, respectively and has presented a substantial accuracy for both models.

Gaikwad et al. [9] employed models such as Support Vector Machines, Naïve Bayes and K-Nearest Neighbours on text-based inputs and achieved excellent levels of accuracy, wherein the SVM model proved to be the most accurate, with an accuracy of 82%. Raza et al. [23] used a total of six different models for sentiment analysis based on textual data, wherein they obtained remarkable accuracies. Singh et al. [26] used models such as Naïve Bayes, J48, BFTree and OneR for sentiment analysis on the Amazon Reviews and IMDB movie reviews datasets, wherein the OneR model gave the best accuracy. In their paper, Stappen et al. [27] used video transcriptions of the MuSe-CAR dataset for sentiment analysis and topic recognition using a support vector machine model, which gave an accuracy of 66.16%. Bhuiyan et al. [4] used NLP techniques to analyse YouTube User comments to retrieve relevant videos and attained an accuracy of about 75%. Agarwal et al. [1] used Recurrent Neural Networks with variants like LSTM and GRUs to perform multimodal sentiment analysis on the CMU-MUSI dataset.

**Table 1** Studies for sentiment analysis on video, audio and text-based inputs

Authors	Year of publication	Type of media input	Technique used	AUC / Accuracy
Rao et al. [22]	2021	Video, Audio and Text	Text—SVM Audio—Decision Tree Video—OpenCV	70%
Li et al. [14]	2019	Video	bi-LSTM	71.1%
Schmidt et al. [25]	2019	Text and Audio	Text – Naive Lexicon Audio – Vokutari tool	Text – 56% Audio – 32%
Gaikwad et al. [9]	2016	Text	SVM, K-NN, NB	SVM – 82% KNN – 73% NB – 64%
Raza et al., [23]	2019	Text	Naive Bayes, SVM, Logistic Regression, K-NN, Random Forests, Decision Trees	SVM: 88% K-NN: 86% DT: 87% NB: 87% RF: 88% LR: 86%
Singh et al. [26]	2017	Text	Naive Bayes, J48, BFTree, OneR	91.3%
Stappen et al. [27]	2021	Video	Support Vector Machines	66.16%
Bhuiyan et al. [4]	2017	Video	NLP	75.435%
Agarwal et al. [1]	2019	Text	Recurrent Neural Networks, Long Short-term Memory and Gated Recurrent Units	78.05%
Kaushik et al. [13]	2013	Audio	Maximum Entropy (ME)	82%
Chu et al. [6]	2017	Audio, Video	Audio—CNN Video—CNN	0.652
Periera et al. [19]	2016	Video, Audio and Text	Text—CNN Audio—SVM Video—NLP	75%
Luo et al. [15]	2019	Audio	CNN and LSTM	69.64
Parveen et al. [18]	2016	Text	Naive Bayes Algorithm	56%

**Table 1** (continued)

Authors	Year of publication	Type of media input	Technique used	AUC / Accuracy
Ezzat et al. [8]	2012	Audio, Speech-to-Text	Naive Bayes, SVM, DT, K-Means, Fuzzy C-means	SVM: 94.4% K-M: 91.7% DT: 86.1% NB: 88.9%
Al-Azami et al. [3]	2020	Video	SVM and Logistic Regression	SVM: 88.17% LR: 85.69%
Rosas et al. [20]	2013	Video, Audio and Text	Text—SVM Audio—SVM Video—SVM	64.86%
Poria et al. [5]	2017	Video	Long Short-Term Memory	MOSI: 80.30% MOUD: 68.11% IEMOCAP: 78%
Gautam et al. [10]	2014	Text	Naive Bayes, Maximum Entropy, SVM, WordNet	NB: 88.2% ME: 83.8% SVM: 85.5% WordNet: 89.9%

Kaushik et al. [13] employed a Maximum Entropy (ME) model to develop a text-based emotion detection model, which uses decoded speech transcripts of YouTube videos.

Chu et al. [6] employed an audio-visual approach to sentiment analysis by using sophisticated models on the Spotify dataset and a vast collection of movie clips, wherein an AUC of 0.652 was obtained. Sentiment analysis of News Videos was conducted by Pereira et al. [19] based on the audio, visual and textual features of these videos, using a myriad of ML techniques, achieving an accuracy of 75%. Luo et al. [15] used a parallel combination of an LSTM and CNN based network to conduct audio-based sentiment detection on the MOSI dataset. Naïve Bayes algorithm was used on the Twitter dataset by Parveen et al. [18] for sentiment analysis, which yielded an accuracy of 57%. In the research conducted by Ezzat et al. [8], text-based classification was conducted using Speech to Text conversions on a set of Call Centre Audio Conversations. A plethora of techniques were used for this research wherein, the SVM model yielded the highest accuracy of 94.4%.

In this paper, Al-Azani et al. [3] fused textual, auditory and visual data for sentiment analysis on the MOSI, MOUD and IEMOCAP datasets by developing SVM and Logistic Regression based classification models. The paper by Rosas et al. [20] explores multimodal sentiment analysis on Spanish videos available online using a support vector machines model that yielded an overall accuracy of 64.86%. Poria et al. [5] conducted multimodal emotion analysis using an LSTM based model on user-generated videos and on MOUD, MOSI and IEMOCAP datasets, where remarkable accuracies were obtained for each dataset. Lastly, in the study conducted by Gautam et al. [10], twitter data was used for sentiment analysis using models based on Naïve Bayes algorithm, SVM and Maximum Entropy, and WordNet was employed for semantic analysis. Through this study, it was found out that Naïve Bayes model gave the highest accuracy for sentiment analysis, meanwhile, WordNet gave an accuracy of 89.9% for semantics analysis.

## 4 Methodology

The core principle behind the research work is sentiment analysis using audio and video. For better and reliable results multiple inputs are considered. The audio input would be converted to text and then processed to perform sentiment analysis to categorize the mood throughout the session. Alongside this, OpenCV will be used to detect facial emotions through facial recognition. Combining the results obtained from both the inputs would give us a report of the person's state of mind which can be used for further diagnosis.

Through this section, the implementation of the elegant system predicated on the premise of sentiment analysis can be seen, which was developed after conducting a thorough research of the different ML techniques that have been employed over the past few decades for real time sentiment analysis. In the following subsections, the 4 primary points of focus will be covered for this project and the expected outcomes for the same. The 4 primary points of focus are –

1. Facial Emotion Recognition – Video Based Analysis: MTCNN Model
2. ChatBot Model – NLP model
3. Sentiment Analysis – Text input analysis: Support Vector Machines model

#### 4. Speech to Text model – Audio conversion to text for Sentiment analysis: Google API

Using these 4 aspects, the proposed project was implemented in the form of a user-friendly platform that caters directly to the user by asking targeted questions and provides him with a quick fix solution.

This section encapsulates all the specific details about the methods, functions and libraries used for the different models used in the project.

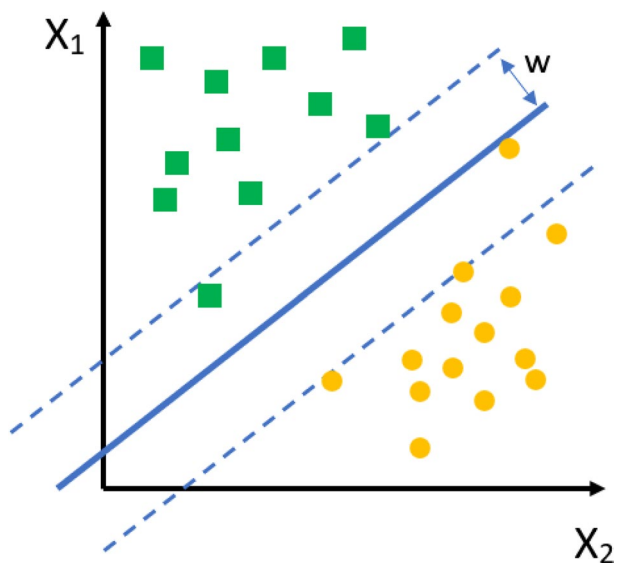
### 4.1 Sentiment analysis

Being a sub-discipline of Natural Language Processing (NLP), sentiment analysis refers to the examination and identification of natural language user inputs in order to predict the positive, negative or neutral inclination of the input [2]. This helps the users find out the true sentiment which in-turn helps them comprehend the real meaning of the given text. It has become prominent in the past few years as increasingly more companies are exploring this domain to identify their user's responses to their products, allowing them to better suit and cater their products to the needs of the customers. It can also be used to gauge the general reaction of the netizens on certain topics or certain new stories whether the outcome has a positive or negative emotion or does it barely affect anyone.

The main aim of every sentiment analysis is to find whether the given data is positive, negative, or neutral. But as we delve deeper into studying the underlying emotions of a human being using machine learning they are also focusing on the emotions like whether the data represents if the user is happy, cheerful, sad, sorry, etc. Using lexicon is an efficient way of determining these range of emotions with the help of neural networks. Lexicon is a list containing various emotions corresponding to certain words.

For this research, after a thorough study, a supervised machine learning algorithm – Support Vector Machines (SVMs) [7] – was employed, that is one of the most commonly used classification and regression techniques. SVM is a sophisticated model which is

**Fig. 1** Two-Dimensional Hyper-plane - Linear Classification using SVM





predicated on creating a hyperplane plotted in an N-dimensional space that segregates the underlying data into respective classes. Figure 1 represents a two-dimensional representation of an SVM model which has a hyperplane separating the data points into two classes. The tuning parameter for an SVM model is called Kernels. There are several different types of kernels, where RBF is mostly used for Non-Linear problems, while linear kernels are used for Linear Classification problems.

Using the sklearn library, Asophisticated SVM model using the previously vectorised data is developed. The model yielded an F1-score of 91%, a recall of 0.91 and a 0.91 precision, proving the highly efficient nature of the technique in cursorily classifying the data into positive, negative or neutral. The F1-score is a statistical measure that helps one analyse the performance of a model and it is given by the harmonic mean of the precision and recall, as in the formula below –

$$F1 - Score = 2 \times \left( \frac{precision \times recall}{precision + recall} \right)$$

The model was initially trained and tested using the Amazon Reviews dataset which comprises of thousands of user inputs of different natures. The model was also saved to be called for future instances using the Pickle library. On testing the model with custom user inputs, an accuracy of over 91% was obtained. Considering the forthcoming employment of this model for the overall sentiment analysis using the chatbot inputs, the model worked ideally, and it can be effectively used for several instances.

## 4.2 NLTK

Natural language processing (NLP) is a field that focuses on making natural human language usable by computer programs. NLTK, or Natural Language Toolkit, is a Python package that we can use for NLP.

A lot of the data that could be analysed is unstructured data and contains human-readable text. Therefore, before programmatical analysis of the data, it first needs to be pre-processed. Following are the steps involved in pre-processing of the data that allows us to feed meaningful and efficient data into the Model.

### 4.2.1 Tokenizing

Tokenizing process allows us a comfortable way of splitting our text data into smaller processable data. It makes it easier to crunch, allowing us to work with more modest bits of text that are still moderately reasonable and significant even outside of the context of the remainder of the text. It is the first step in the pipeline which converts the enormous unstructured data into easily processable and algorithm friendly structured data (Table 2).

**Table 2** Tokenized words

Before tokenizing	After tokenizing
“I dropped out of Reed College after the first 6 months, but then stayed around as a drop-in for another 18 months or so before I really quit.”	['I', 'dropped', 'out', 'of', 'Reed', 'College', 'after', 'the', 'first', '6', 'months', ',', 'but', 'then', 'stayed', 'around', 'as', 'a', 'drop-in', 'for', 'another', '18', 'months', 'or', 'so', 'before', 'I', 'really', 'quit', '.']

## 4.2.2 Stop words

Removing stop words is a crucial step in our pipeline which allows us to filter out the unwanted words which are not essential to process while doing our analysis. This streamlines our process and reduces the number of words that need to be processed making our process fast and efficient. Frequently used words like 'i', 'am', 'to' which do not really contribute to finding out the emotion of the message are some examples of stop words which are scrubbed out in the pipeline (Table 3).

## 4.3 Stemming

Stemming process is one of the important processes which allows users to extract meaningful data from the given text values. It uses tokenized words to extract root or base words from different variants of a similar word. This allows us to clean the majority of similar words having the same meaning and further making our training process much faster and efficient. It is one of the main parts of the pipelining process. The main obstacle in using this technique is to look out for under stemming and over stemming (Table 4).

- Under stemming usually occurs when related words are not condensed to the same stem even if they are similar resulting in a false negative.
- Over-stemming is encountered when related words are condensed to the same stem even if they are not remotely related, resulting in a false positive.

### 4.3.1 Lemmatization

Lemmatization is another process in the pipeline where grouping of words takes place where the words are crumpled and are then processed as a single item. Although it looks like stemming, it is a crucial step as it brings context to words by linking words which have similar meaning.

### 4.3.2 Quantifying emotions

An elaborate dataset was created which contains copious number of words and the emotion attached to them. After the extraction of the meaningful words from our text, the text is compared with the text in the database, which allows us to find the emotion hidden behind the text. After successfully extracting the words and its emotion, the text was run through

**Table 3** Before and After Stop Words

Before removing stop words	After removing stop words
['I', 'dropped', 'out', 'of', 'Reed', 'College', 'after', 'the', 'first', '6', 'months', ',', 'but', 'then', 'stayed', 'around', 'as', 'a', 'drop-in', 'for', 'another', '18', 'months', 'or', 'so', 'before', 'I', 'really', 'quit', '.']	['dropped', 'Reed', 'College', 'first', 'months', 'stayed', 'around', 'drop-in', 'another', '18', 'months', 'really', 'quit']

**Table 4** Before and after—stemming words.

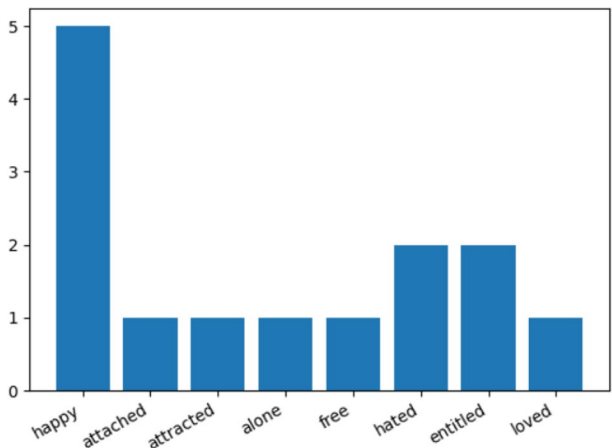
Before stemming	After stemming
(a) ['friend', 'friendship', 'friendzone', 'friendly', 'friendlily', 'friendlier', 'friendlier']	Friend: friend Friendship: friend friendzone: friendzon friendly: friend friendlily: friend friendlier: friend friendlier: friend
(b) ['dropped', 'Reed', 'College', 'first', 'months', 'stayed', 'around', 'drop-in', 'another', '18', 'months', 'really', 'quit']	Dropped: drop Reed: reed College: college First: first Months: month Stayed: stay Around: around drop-in: drop-in another: another 18: 18 Months: month Really: real Quit: quit

a Counter which allows us to quantify the emotions present in the words. Figure 2 shows a plot of the magnitude of emotions detected in a sample video fed into the classifier.

### 4.4 Facial emotion recognition

Emotions are the biggest driving force in any decision making. It is often said that a person might be saying something, but their face might be saying something else. Facial expressions provide a lot of insight into a person’s mood or emotions. Two models were

**Fig. 2** Emotions



considered for facial emotion recognition (FER), Multi cascade convolutional network (MTCNN) and Haar Cascade classifier.

The MTCNN is a neural network model used to detect faces. It consists of 3 convolutional networks cascaded together, hence the name. It requires face features such as eyes, mouth and nose. Haar Cascade is a well-known model for detecting faces. It uses edge or line detection to detect facial features. In recent years, Convolutional Neural Networks (CNNs) have increased the accuracy of face detection manifolds [28].

The process for facial emotion recognition is the same for images and videos except for an additional step in the case of videos. Theoretically, video is a series of images in motion. The video is split into separate images and each image is passed through the emotion recognition algorithm to detect the various emotions in the frame and their magnitude on a scale from 0 to 1. Post this an output video is generated which consists of a box around the face in the input video with the live emotion detection along with their magnitude. Figure 3 accurately represents the processing of a video input by splitting it into frames and then further passing it to the classifier for sentiment analysis.

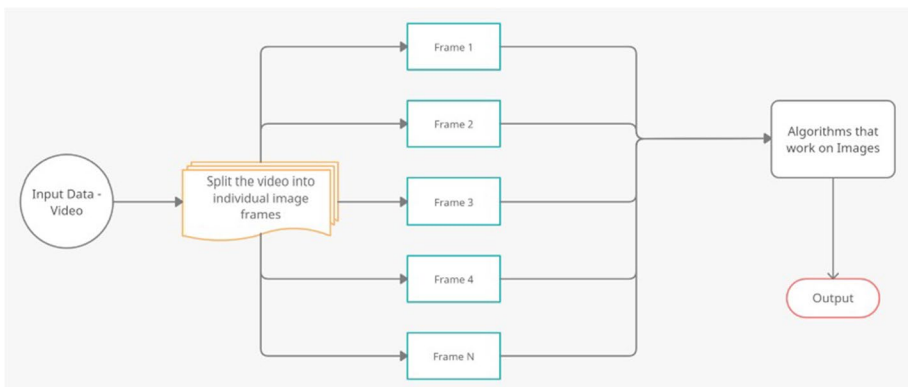
Pandas library is used to analyse the result of all frames and their output is plotted against time using matplotlib for the complete duration of the video. Further analysis is drawn by taking the sum of all the emotions detected in the video and forming a table with emotions and their values.

The text, speech and video input models together constitute the sentiment analysis model which is capable of producing reliable and cogently justified outputs after thorough analysis of the user's input. Figure 4 represents the flowchart of the entire project which explains the process of sentiment analysis and input processing in a simple manner.

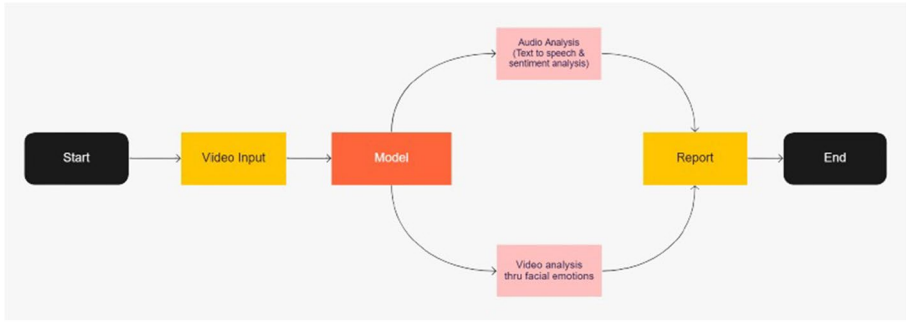
## 5 Result

The model for textual sentiment analysis yielded an F1-score of 91%, a recall of 0.91 and a 0.91 precision, proving the highly efficient nature of the technique in cursorily classifying the data into positive, negative or neutral.

The result of the video analysis is obtained in the form of a graph consisting of emotions plotted against time. The magnitude of emotions is plotted on a scale of 0 to 1. The



**Fig. 3** Pictorial representation of video input processing for sentiment analysis

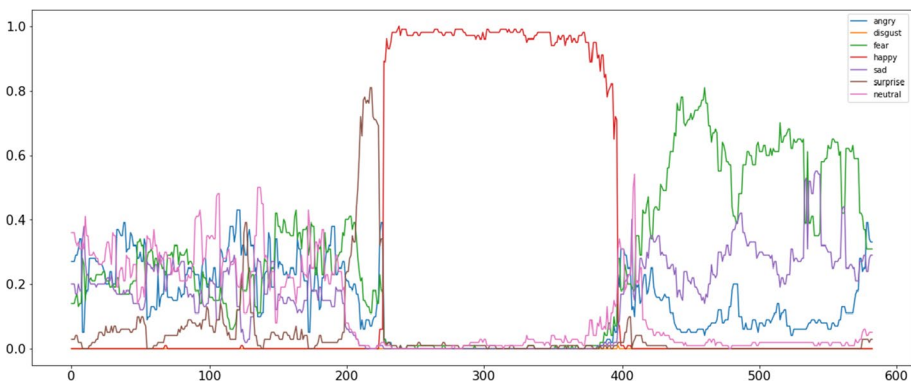


**Fig. 4** Simple flowchart representation of the sentiment analysis model developed by the authors

X-axis of the plot represents the timespan of the video while the Y-axis represents magnitude of emotion. Assorted colours have been used to signify different emotions. It can be seen that until around the 200 s mark, a mix of emotions can be detected in the video while after that till the 400 s mark a strong signal of “Happy” has been detected which could be cross verified from the video. After the 400 s mark again a mix of emotions is observed. This information can be corroborated from Fig. 5 given below, in which lines of different colours represent the different magnitudes of the emotions in the sample video fed to the classifier. Moving on, from Table 5, thus obtained, it can be seen that the most dominant emotion in the test video is ‘happy’ which is followed up by ‘fear’ which was visible in the video as well and thus, it can conclude that an accurate analysis has been made.

MTCNN [11] which uses CNN provides better results than the Haar Cascade classifier. Another advantage of MTCNN is that it creates an arbitrary rectangle around the face for better detection as compared to Haar Cascade which creates a square.

Having created this elegant system of our own, we were able to notice the performance of our model and how it was able to produce reliable results based on live input from the user. After conducting a thorough research of the previous studies conducted in this domain, we were able to notice significant improvements in the outputs and the overall accuracy/F1-score of the models. While the papers focussing on NLP only worked with



**Fig. 5** Emotions plotted against time

**Table 5** Magnitude of the emotions in quantifiable values obtained as output from the MTCNN classifier

	Human emotions	Emotion value from the video
0	Angry	76.78
1	Disgust	0.01
2	Fear	156.73
3	Happy	163.22
4	Sad	89.77
5	Surprise	28.33
6	Neutral	67.22

pre-existing datasets, our model was able to produce accurate responses and predictions based on a user's natural language text input. Furthermore, the live video input from a user was broken into frames for a cogent analysis and complete processing of each frame to identify a sentiment over a certain period of time. The results yielded by this part of the model, on top of the text and speech analysis, were testaments of the excellent performance of the aforesaid classifier. Thus, it can be stated confidently that the extensive, manifold model developed by the authors is an improvement on several past studies as it constitutes multimedia input (text, audio and video) from a user and processes the same in little to no time and predicts to a high level of accuracy, the sentiments being expressed by the user.

## 6 Conclusion

The discipline of Machine Learning and Deep Learning has found prolific applications in the field of semantics and sentiment analysis. Interpretation of emotions and responses through computers helps not just developers, but it helps professionals across various domains. Such recognition can allow counsellors and in fact, the users themselves to identify and keep track of their daily moods.

Through this paper, we have successfully conducted a thorough research of the different ML techniques that have been employed over the past few decades for real time sentiment analysis. Moreover, this paper was predicated on a project developed in order to allow a user to track his/her day-to-day moods and get appropriate recommendations for how they can alleviate their stressful or anxious state of being. This manifold project comprises of the following parts which encapsulate about 3 different ML models, namely –

1. Facial Emotion Recognition – Video Based Analysis: MTCNN Model
2. ChatBot Model – NLP model
3. Sentiment Analysis – Text input analysis: Support Vector Machines model
4. Speech to Text model – Audio conversion to text for Sentiment analysis: Google API

Using these models, we created a user-friendly platform that caters directly to the user by asking targeted questions and provides him with a quick fix solution. Such an application has never-ending applications and can be adapted for several different purposes such as – Communication through Text, Audio or Video, or even several applications in the field of medicine and psychology. In the near future, the aim is to develop a scalable mobile application for Android and iOS users, so that they can assess and keep track of their moods and

reactions on a regular basis and even use this model to interpret the tone and mood of the person they are communicating with.

**Data availability** The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

## Declarations

**Conflicts of interests** The authors declare that they have no conflicts of interests.

## References

1. Agarwal A, Yadav A, Vishwakarma DK (2019) Multimodal Sentiment Analysis via RNN variants. Proc - 2019 IEEE/ACIS 4th Int Conf Big Data, Cloud Comput Data Sci BCD 2019:19–23
2. Ahmad M, Aftab S, Muhammad S, Ahmad S (2017) Machine Learning Techniques for Sentiment Analysis: A Review. Int J Multidiscip Sci Eng 8(3):27–32
3. Al-Azani S, El-Alfy ESM (2020) Enhanced Video Analytics for Sentiment Analysis Based on Fusing Textual, Auditory and Visual Information. IEEE Access 8:136843–136857
4. Bhuiyan H, Ara J, Bardhan R, Islam MR (2017) Retrieving YouTube video by sentiment analysis on user comment. Proc 2017 IEEE Int Conf Signal Image Process Appl ICSIPA 2017(1):474–478
5. Poria S, Cambria E, Hazarika D, Majumder N, Zadeh A, Morency LP (2017) Context-Dependent Sentiment Analysis in User-Generated Videos. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 873–883, Vancouver, Canada. Association for Computational Linguistics
6. Chu E, Roy D (2017) Audio-visual sentiment analysis for learning emotional arcs in movies. Proc - IEEE Int Conf Data Mining ICDM 2017(November):829–834
7. Cortes C, Vapnik V (1995) Support-vector networks. Mach Learn 20(3):273–297
8. Ezzat S, Gayar NE, Ghanem MM (2010) Investigating analysis of speech content through text classification. Int Conf Soft Comput Pattern Recog Cergy-Pontoise, France, 2010: 105–110. <https://doi.org/10.1109/SOCPAR.2010.5686000>
9. Gaikwad G, Joshi DJ (2016) Lexicon Dictionary and Machine Learning. Int Conf Inven Comput Technol 2(1):1–6
10. Gautam J, Atrey M, Nitima M, Balyan A, Shaw R, Ghosh A (2021) Twitter Data Sentiment Analysis Using Naive Bayes Classifier and Generation of Heat Map for Analyzing Intensity Geographically. [https://doi.org/10.1007/978-981-33-6919-1\\_10](https://doi.org/10.1007/978-981-33-6919-1_10)
11. Ghofrani A, Toroghi RM, Ghanbari S (2019) Realtime face-detection and emotion recognition using mtccn and minishufflenet v2. In 2019 5th Conference on Knowledge Based Engineering and Innovation (KBEI) Tehran, Iran, 28 February–1 March 2019; pp. 817–821
12. Kaur H, Ahsaan SU, Alankar B, Chang V (2021) A proposed sentiment analysis deep learning algorithm for analyzing COVID-19 tweets. Inf Syst Front 23(6):1417–1429
13. Kaushik L, Abhijeet S, Hansen JHL (2013) Sentiment extraction from natural audio streams. 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE
14. Li C, Zhan G, Li Z (2018) News Text Classification Based on Improved Bi-LSTM-CNN 890–893. <https://doi.org/10.1109/ITME.2018.00199>
15. Luo Z, Xu H, Chen F (2019) Audio sentiment analysis by heterogeneous signal features learned from utterance-based parallel neural network. In: AffCon@AAAI
16. Maulik PK, Devarapalli S, Kallakuri S, Tewari A, Chilappagari S, Koschorke M, Thornicroft G (2017) Evaluation of an anti-stigma campaign related to common mental disorders in rural India: a mixed methods approach. Psychol Med 47(3):565–575
17. Mechili EA, Saliq A, Kamberi F, Girvalaki C, Peto E, Patelarou AE, Bucaj J, Patelarou E (2021) Is the mental health of young students and their family members affected during the quarantine period? Evidence from the COVID-19 pandemic in Albania. J PsychiatrMent Health Nurs 28: 317–325
18. Parveen H, Pandey S (2017) Sentiment analysis on Twitter Data-set using Naive Bayes algorithm. Proc 2016 2nd Int Conf Appl Theor Comput Commun Technol iCATecT 2016:416–419

19. Pereira MHR, Pádua FLC, Pereira ACM, Benevenuto F, Dalip DH (2016) Fusing audio, textual, and visual features for sentiment analysis of news videos. *Proc 10th Int Conf Web Soc Media, ICWSM 2016*, no. Icwsm, pp. 659–662
20. Perez Rosas V, Mihalcea R, Morency LP (2013) Multimodal sentiment analysis of Spanish online videos. *IEEE Intell Syst* 28(3):38–45
21. Pérez-Rosas V, Mihalcea R, Morency LP (2013) Utterance-level multimodal sentiment analysis. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (Vol. 1)*, pp. 973–982
22. Rao A, Ahuja A, Kansara S, Patel V (2021) Sentiment analysis on user-generated video, audio and text. *Proc. - IEEE 2021 Int. Conf. Comput. Commun. Intell. Syst. ICCCI 2021*, pp. 24–28
23. Raza H, Faizan M, Hamza A, Mushtaq A, Akhtar N (2019) Scientific text sentiment analysis using machine learning techniques. *Int J Adv Comput Sci Appl* 10(12):157–165
24. Samji H, Wu J, Ladak A, Vossen C, Stewart E, Dove N, ..., Snell G (2022) Mental health impacts of the COVID-19 pandemic on children and youth—a systematic review. *Child Adoles Ment Health* 27(2):173–189
25. Schmidt T, Burghardt M, Wolff C (2019) Toward multimodal sentiment analysis of historic plays: A case study with text and audio for Lessing’s *Emilia Galotti*. *CEUR Workshop Proc* 2364:405–414
26. Singh J, Singh G, Singh R (2017) Optimization of sentiment analysis using machine learning classifiers. *Human-Centric Comput Inf Sci* 7(1)
27. Stappen L, Baird A, Cambria E, Schuller BW, Cambria E (2021) Sentiment Analysis and Topic Recognition in Video Transcriptions. *IEEE Intell Syst* 36(2):88–95
28. Xiang J, Zhu G (2017) Joint face detection and facial expression recognition with MTCNN. In *2017 4th international conference on information science and control engineering (ICISCE)* (pp. 424–427). IEEE
29. Zadeh A, Zellers R, Pincus E, Morency LP (2016) Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intell Syst* 31(6):82–88

**Publisher’s note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

## Authors and Affiliations

Rishit Jain<sup>1</sup>  · Revant Singh Rai<sup>1</sup> · Sajal Jain<sup>1</sup> · Ruchir Ahluwalia<sup>1</sup> · Jyoti Gupta<sup>1</sup>

Rishit Jain  
rishitjainn@gmail.com

Revant Singh Rai  
revanrai@gmail.com

Sajal Jain  
sjain30sept@gmail.com

Ruchir Ahluwalia  
ruchirwalia1@gmail.com

<sup>1</sup> Department of Electronics and Communication Engineering, Bharati Vidyapeeth’s College of Engineering, New Delhi 110063, India