JADAVPUR UNIVERSITY

Study of Human Emotion Using Multimodal Data

by

Soumik Chanda

REGN. NO:154162 of 2020-2021

EXAM ROLL NO: M4CSE22038

under the supervision of

Dr. Sarbani Roy

Professor

Department of Computer Science and Engineering, Jadavpur University

Kolkata, West Bengal, India

Thesis submitted in partial fulfillment of requirements
for the degree of

Master of Computer Science and Engineering

of

JADAVPUR UNIVERSITY

August 16, 2022

Certificate from the Supervisor

This is to certify that the work embodied in this thesis entitled **"Study of Human Emotion Using Multimodal Data"** has been satisfactorily completed by **Soumik Chanda** (Registration Number 154162 of 2020-21; Class Roll No. 002010502038; Examination Roll No. M4CSE22038). It is a bonafide piece of work carried out under my supervision and guidance at Jadavpur University, Kolkata for partial fulfillment of the requirements for the awarding of the **Master of Engineering in Computer Science and Engineering** degree of the Department of ComputerScience and Engineering, Faculty of Engineering and Technology, Jadavpur University, during the academic year 2021-22.

_____

Dr. Sarbani Roy,
Professor,
Department of Computer Science and Engineering,
Jadavpur University.
(Supervisor)

Forwarded By:

_____

Prof. Nandini Mukherjee,
Head,
Department of Computer Science and Engineering,
Jadavpur University.

_____

Prof. Chandan Majumdar,
DEAN,
Faculty of Engineering & Technology,
Jadavpur University.

Department of Computer Science and Engineering Faculty of Engineering And Technology Jadavpur University, Kolkata - 700 032

Certificate of Approval

This is to certify that the thesis entitled **"Study of Human Emotion Using Multimodal Data"** is a bonafide record of work carried out by **Soumik Chanda** (Registration Number 154162 of 2020-21; Class Roll No. 002010502038; Examination Roll No. M4CSE22038) in partial fulfillment of the requirements for the award of the degree of **Master of Engineering in Computer Science and Engineering** in the **Department of Computer Science and Engineering, Jadavpur University**,during the period of June 2021 to June 2022. It is understood that by this approval, the undersigned do not necessarily endorse or approve any statement made, opinion expressed or conclusion drawn therein but approve the thesis only for the purpose of which it has been submitted.

Examiners:

(Signature of the Examiner)

(Signature of the Supervisor)

Department of Computer Science and Engineering Faculty of Engineering And Technology Jadavpur University, Kolkata - 700 032

Declaration of Originality and Compliance of Academic Ethics

I hereby declare that the thesis entitled **"Study of Human Emotion Using Multimodal Data"** contains literature survey and original research work by the undersignedcandidate, as a part of his degree of **Master of Engineering in Computer Science and Engineering** in the **Department of Computer Science and Engineering, Jadavpur University**. All information have been obtained and presented inaccordance with academic rules and ethical conduct.

I also declare that, as required by these rules and conduct, I have fully cited and referenced all materials and results that are not original to this work.

Name: Soumik Chanda

Examination Roll No.: M4CSE22038

Registration No.: 154162 of 2020-21

Thesis Title: Study of Human Emotion Using Multimodal Data

Signature of the Candidate:

ACKNOWLEDGEMENT

I am pleased to express my gratitude and regards towards my Project Guide Dr. Sarbani Roy, Professor, Department of Computer Science and Engineering, Jadavpur University, without whose valuable guidance, inspiration and attention towards me, pursuing my project would have been impossible.

For his recommendations and persistent support, I would like to convey my sincere gratitude to Asif Iqbal Middya, Research Fellow at Jadavpur University in Kolkata.

Last but not the least, I express my regards towards my friends and family for bearing with me and for being a source of constant motivation during the entire term of the work.

Soumik Chanda
MCSE Final Year
Exam Roll No: M4CSE22038
Regn. No: 154162 of 2020-21
Department of Computer Science and Engineering,
Jadavpur University

JADAVPUR UNIVERSITY

Abstract

Faculty of Engineering and Technology
Computer Science and Engineering

Master of Engineering
Study of Human Emotion Using Multimodal Data

by Soumik Chanda

Emotion analysis and classification problem is an emerging as well as challenging task in today's world. It has an immense application on business, healthcare and educational industries. Precise emotion recognition of the user is pretty much difficult as the real set of possible emotional states are fuzzy in nature. Not only that, but also short utterances cannot be properly classified because of their dependency on the context of the dialogue. So, various machine learning and deep learning models can be deployed on multi modal data (text, emoji, audiovisual data etc.) in order to precisely classify and analyze various types of emotion specified by eminent researchers along with various evaluation metrics like confusion matrix, weighted f1 score, accuracy etc. so that predictive models can be compared and assessed. In this work, multi modal multi party dataset has been taken into consideration and this dataset has been accordingly preprocessed so that it can be annotated and features can be extracted from textual and audio data. Then, baseline algorithm like bcLSTM has been deployed on uni modal data and audio along with bimodal data and at last they are analyzed by some evaluation parameters relevant to this domain.

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

1.1 Overview

Emotion analysis problem is pretty much challenging and active field of research in today's world. Precise emotion recognition of the user is pretty much difficult as the real set of possible emotional states are fuzzy in nature. Not only that, but also their instability along the time lapse arouses subsequent difficulties for emotion analysis. Today researchers from various domains (e.g., computer science, healthcare, communication etc.) are attracted towards this domain. In this work, different machine learning (ML) and deep learning (DL) models can be discussed on social media data (text, emoji, audio data etc.) to correctly classify and analysis various types of emotion specified by eminent researchers. Various evaluation metrics like Accuracy, F-score are assessed for various models. Also, performance can be enhanced for various predictive ML and DL models by tuning various model parameters and hyper parameters.

1.2 Motivation

Now-a-days, emotion analysis has huge potential to explore. People become victim of stress, mood swing and various psychological problems which are literally unknown to them very often due to work stress, rat race in daily life, excessive use of social media which make people feel insecure as well as degrading their mental health to some extent. Not only that, but also it has paramount importance in data driven marketing. Some applications of emotion analysis can be as follows—

1) Improvement of customer experience:- In today's competitive world, service providers collect more and more information about their users. Emotion analysis is used for this type of sales and marketing in various domains.

    a) Enhanced website customization:- Generally web page layout, contents etc are created and displayed according to the convenience of user. Adding information about the emotions of users could provide more accurate model for users. Sometimes first impression test is used as a good predictor in web page design. In first impression testing, the most important distinction is to differentiate user's interest from boredom or disgust, which scenario is especially dedicated to web page usability testing.

    b) Advertisement reaction model:- Sometimes the revenue model for some companies is partially based on the on-line advertising. Generally in those cases, viewer's emotion can be tracked using facial recognition software. By analyzing these emotional reaction, a set of information can be collected which can be beneficial in choosing the appropriate type of advertising depending on the target audience.

2) Usage in education field:- It has been reported that some of the emotional states support learning process while other suppress them. Automatic emotion recognition can help to explore these phenomena by making assessments of learner emotional states. Sometimes they are very much resourceful to evaluate educational resources, especially those prepared for self-learning. Even emotion analysis has immense potential of building intelligent tutoring system which is deployed for different educational applications.

3) Providing emotion-aware personalized gaming:- Sometimes creators of gaming program use emotion recognition technology in order to deliver a better gaming experience. Facial expression of individual playing the game is captured by a webcam. The collected data is then fed to an emotion recognition tool which infers emotions and the game play is then altered based on the emotional state of the player, therefore providing a personalized experience. Not only that, but also emotion analysis helps in finding accurate insights regarding the difficulty level of the game, factors on why one leave the game. The information is then used to provide players with dynamic changes in order to deliver a smoother experience, increase in gaming time and higher player retention.

4) Making the interview process better for best suited:- Some company is analyzing the emotional expressions of candidates applying for job. As applicants undergo a video interview through computer, tablet or smart phones; an AI algorithm constantly measures the facial expression, emotions or other personality traits to make a detailed report regarding candidate's emotional reaction which helps to shortlist the pool of candidates who are best suited for the job. Moreover, unlike human recruiters who may be the subject of unconscious bias, the analysis method decides purely based on the captured emotions.

5) Improvement in health care industry:- Emotion analysis itself is very crucial in healthcare related field. For instances, some researchers are interested in how diseases of the brain like Alzheimer's disease, Parkinson etc. affect the ability to communicate emotions or language impairments. Some human-computer interaction (HCI) is developing themselves versatile to analyze emotions better day-by-day.

6) Building Chat clients:- Emotion analysis has been used effectively in building several real-time based chat applications which in turn can be used for assessing the critical mental health issues. Generally the input of the system is a text while the output is an expression image which depends on the emotion hidden in sentence as well as the intensity of that emotion.

1.3 Objective

Generally emotion can be defined as a psycho-physiological process which is triggered by conscious or unconscious perception of any object or situation. It is often associated with temperament, mood, motivation etc. It plays a crucial part in human communication and can be expressed either verbally through emotional vocabulary or by expressing nonverbal cues such as facial expression, gestures etc. A lot of people express their emotion, sentiment or feelings in social media via multi modal data (text, emoji, audiovisual data etc.). So it is pretty obvious that multi modal data has to be handled to predict various types of Emotion. One of the main challenges is words with multiple emotion polarity. In case of such words, identification of exact emotion is very challenging. Sometimes audio data, carrying a lot of noise with them, can be a hindrance to emotion analysis. In case of homophones where words seem similar, can not be justified with any single emotion. Last but not the least, people using sarcastic words can be a confusing state for correct emotion analysis as the internal emotion hidden in these words seems to be pretty much different. Also the use of short utterances, which are popular now-a-days for micro-blogging concept, act as a hindrance for proper emotion classification as they mostly depend on statement. So, various models are taken into consideration for better classification and recognition issue.

1.4 Organization

The work is organized as follows: chapter 2 describes some existing work which are related to emotion classification domain for textual, audio or multi modal data. In chapter 3, the problem statement has been discussed along with data collection step as well as with the workflow of the work. Also each part of the workflow is described concisely. Chapter 4 discusses the experimental setup and result related to the deployed model for emotion classification purpose. Finally chapter 5 concludes the work and also discusses the future scope related to it.

CHAPTER 2

Literature Review

In this section, a detailed discussion about emotion variety along with datasets, generally used for analyzing it, is provided. Recent advances in emotion recognition have motivated to create novel databases containing emotional expressions in different modalities. These databases mostly cover speech, text data or audiovisual data. The visual modality includes facial expressions or body gestures. The audio modality covers posed or genuine emotional speech in different languages.

However, the idea of analyzing emotions by computers was first discussed in the paper "Affective Computing" of Picard (1997) [16] where the foundational ideas have been discussed. But multi modal data has to be considered as there are a lot of cues to be captured. Now before analyzing various data, a look into properties as well as efficient ways of representing emotions has to be discussed. Generally there are four key properties of emotion which are (a) Antecedent (situation or cause acting as a trigger for emotion); (b) Signal (Physiological method used for conveying emotion); (c) Response (expected reaction of an emotion); (d) Coherence. If the emotion "sadness" or "grief" is considered, then any unfortunate event occurrence can act as antecedent, weeping may be a signal whereas consoling that person is an act of response. The fact of expressing sadness in a similar way for most humans indicates the coherence for that particular emotion. Till date, various discrete categorizations of emotions have been proposed. In a broader viewpoint, only happiness and sadness can be considered as the binary basic emotion. But specifically, six basic emotions i.e. anger, surprise, disgust, enjoyment, fear and sadness can be considered for classification purpose which is proposed by Dr. Ekman et.al. [6]. Tree structure of emotions is proposed by Parrott et.al. where more than 100 emotions are organized into three levels of emotions (Primary, Secondary, Tertiary). Even dimensional scales of emotion, such as emotion wheel by Plutchik et.al. [7] and valence-arousal scale by Russell et. al. [11] also have been proposed. Russell's valence-arousal scale is generally used to quantitatively describe emotions. In this scale, each emotional state can be placed on a 2D plane with arousal and valence as the horizontal and vertical axes. Sometimes, third dimension dominance can also be added in the model. Arousal can range from inactive (i.e. uninterested, bored) to active (i.e. alert, excited) whereas valence ranges from unpleasant (i.e. sadness, stress, boredom) to pleasant (i.e. happiness, excitement, contentment). Dominance ranges from a helpless and weak feeling (without control) to an empowered feeling (in control of everything). By the way, it has to be understood that arousal is related to the level of control but not with the magnitude of emotion. Plutchik's wheel of emotions has 3 components which are (a) Basic emotion pairs (Emotions are arranged such that the antonymy between emotion pairs are maintained like on the second ring from inside, it is pair-wise joy-sadness, anger-fear, anticipation-surprise and trust-disgust), (b) Emotion lobes (Generally four lobes are present, each related with a basic emotion pair. Also three concentric circles are present

where the innermost circle includes ‘activated’ basic emotions as outermost circle includes ‘deactivated’ basic emotions), (c) Combination of emotions (basic emotions combine to form more complicated emotions which are listed outside each of the lobes on the outermost circle).

Now, a vivid discussion about emotive potential of text can be done. Generally, intonation and accentuation from linguistic study of speech are used for emotion analysis. If they are absent, then there are two modes generally analyzed for emotion analysis from text, which are (a) Affective items (Normally expletives and interjections like ugh, yeah etc. can act as emotion indicators) and (b) Emotive vocabulary (words used from emotive vocabulary such as love, good etc. can directly refer emotion). Now, if a discussion about data set, which is used for comparative study of emotion analysis through text, has to be made; then obviously it can be divided into long text and short text in a much broader sense. Many works prefer short text in this context as many a time long text can be difficult to properly detect emotion whereas short text is pretty much pointed towards emotion expression. Some research papers like analysis of news headlines by Strapparava and Mihalcea (2007) [12] and Bellegarda (2010) [13] along with microblog analysis, generally hashtag-based supervision, by Aman and Szpakowicz (2007a) [14] and Chaffar and Inkpen (2011) [15] can be mentioned in this domain. However, also in the domain long text based emotion analysis, some research papers like email analysis by Liu et al. (2003) [17] and Alm (2008) [18] work on children’s stories has been done but in the latter work, large data sample of a story is broken down into sentences and the annotation is done separately for each sentence.

Now, obviously a discussion about emotion lexicons has to be made in this context. Generally, emotion lexicon acts as a knowledge repository containing textual units annotated with emotion labels or it may exist as a set of word lists and they are used as a knowledge base to understand emotion in a text like words ‘happy’, ‘excited’, ‘pleased’ etc. can refer to the emotion "happiness". It has a use in a simple rule-based emotion analysis system. There are some popular lexicons for emotion analysis which are

LIWC (Linguistic inquiry and word count) system, a popular text processing platform, often works as a baseline result as it provides software to understand sentiment/emotion etc. in a given piece of text. LIWC consists of 4500 words and word stems organized in four categories. The first three categories are non-sentiment categories which are (a) Linguistic processes (generally pronouns, prepositions, conjunctions etc.), (b) Speaking processes (generally speech fillers), (c) Personal concerns whereas the last category belongs to sentiment category which consists of words of psychological processes and it is also divided into two subcategories which are (a) Cognitive processes (words expressing cognition such as possibility or certainty signifying probability of truth value of a statement and inhibition signifying reversal of emotion expressed in that statement) and (b) Affective processes (words relating emotions like anxiety, anger, sadness, positive or negative emotions). LIWC consists of a total of 713 words corresponding to cognitive processes and 915 words corresponding to affective processes. For creating manually operated LIWC, at first hierarchy of categories for lexicon has been created and they are populated with words manually annotated by three judges.

ANEW (A new english wordlist) is another lexicon, consisting of 1000 words, each of which is annotated in form of three tuple namely pleasure, arousal and activation. But here, attribute 'pleasure' depicts positive/negative emotion, 'arousal' points towards intensity of the 'pleasure' where 'activation' indicates about the word giving any signal of being controlled or not. ANEW is also created manually corresponding with involvement of 25 annotators who use ScanSAM sheet for marking of the tuples. This lexicon also has a spanish version where a correlation between these two is shown to establish the quality of ANEW-spanish.

Emo-Lexicon is also another type of lexicon which is generally built by crowd-sourcing taking the leverage of online crowd-sourcing platform like amazon mechanical turk. This structure, consisting of 10000 terms, allows a word to belong to more than one emotion category showing fuzziness. But it is vulnerable to quality flaws as annotators here may not aware of the given word or not enough serious to do the work with attention. So, to deal with that problem, words are collected at first from a thesauras to compare with General Inquirer and Wordnet selecting some word-sense pairs. Then annotators are checked by giving a target word along with four words and asked to select one closing that target word. If it is correctly done, then they are given the task of annotating the word for various emotion category else word-sense pair is discarded. This is generally regarded as quality control step.

Another very popular lexicon is Wordnet-Affect, an annotated version of Wordnet, which acts as a linked lexical repository using semantic concept. Wordnet superficially resembles a thesauras which groups words based on their meanings. In Wordnet, english words are generally grouped into sets of cognitive synonyms known as 'synset' which express distinct concept. Synsets are interlinked using semantic and lexical relations such as hyponymy and antonymy. Hyponymy expresses a relation between two concepts where one concept is a type of other concept like 'beef' acts as hyponym of 'meat' whereas antonymy defines lexica l relation known as opposites. Wordnet also includes several similarity measures like Wu-Palmer similarity etc. Wordnet-Affect consists of 2874 synsets annotated with a-label, which is considered as an emotionspecific label (happy, sad etc) or a set of cognitionspecific label (certainty, impossibility or possibility etc) and this is created in a semi-automatic way. At first, manually a set of core synsets has been created where each synset is assigned to at most one a-label determined by annotators. Now, along with Wordnet graph structure, an emotion known synset label is projected to other synsets using relations of Wordnet. At last, the resultant lexicon is manually evaluated. Similarly, Chinese Emotion Lexicon has been built in a much similar approach to Wordnet-Affect. But here the core set of words is expanded using similarity matrix replacing the graphical structure and relations used in Wordnet. The similarity value in the matrix is updated using syntagmatic similarity, paradigmatic similarity and textbtlinguistic peculiarity.

Another approach to emotion analysis, for any social media text corpus, introduces some POS specific prior polarity features. After manual annotation of corpus and introducing emoticon and acronym library for preprocessing, prior polarity of words has been done with the help Dictionary of Affect in Languages (DAL) and wordnet. Also, annotation can be done with the help of various standard libraries (AFINN, Bing, NRC etc.) for text corpus. Tree-Kernel method has been introduced to obviate the need for feature engineering. Actually,

classification task has been done by analyzing the features of various models like unigram model, senti-features model, tree kernel model, kernel plus senti-features model and unigram plus senti-features model. Some researchers describe another method where at first, to classify a phrase as emotion phrase or not; Cohen's kappa model is used to classify them into known mutually exclusive categories. It is also used for pair wise agreement between the annotators on emotion/non-emotion labeling of the sentences in the corpus. Agreement on emotion intensities can also be measured using kappa as there are distinct categories- high, medium and low. There are also emotion indicators which are words or strings of words selected by annotators as marking emotion in a sentence. MASI method is chosen for measuring agreement on co-reference annotation and in the evaluation of automatic summarization. The General Inquirer (GI) as well as Wordnet-Affect can be used for feature extraction purpose. At last, Naive Bayes and Support Vector Machines (SVM) are used for emotion classification.

In this context, some approaches can be discussed which is helpful to extract features from text corpus. The most naive as well as popular approach is keyword spotting which is a method of spotting keywords from any emotive vocabulary. Here, based upon several unambiguous emotion words such as 'happy' or 'enraged' etc, any text corpus is classified into some affect categories. Lexical resources like Wordnet-Affect or Roget's Thesauras (Roget, 2008) [19] is used as emotive vocabularies here. Also some linguistic modifications (such as negations) can be applied in order to analyze emotion. Another approach, slightly sophisticated than previous one, is lexical affinity where an arbitrary word is given a probabilistic affinity for a particular emotion instead of merely labeling them as just some obvious affect words. These probabilities are usually obtained from linguistic corpora. It generally outperforms pure keyword spotting but also is vulnerable to two problems; one of which is that it can easily be tricked by negations and different word senses as it operates solely on the word-level and the second is that lexical affinity probabilities are often biased toward text of a particular genre, thus making it difficult to develop a domain-independent model. Also statistical approach is used for emotion analysis from text where a large training corpus of annotated texts is fed to a machine learning or deep learning algorithm to make it learn the relationship among lexical entities as well as paradigmatic features like punctuation. Latent semantic analysis (LSA) as a statistical method is popular for affect classification of texts. But statistical methods are generally proven as semantically weak, which means several lexical or co-occurrence elements in this model have little predictive value individually though there is an exception of some obvious affect keywords. Some traditional machine learning algorithms, like Naive Bayes and Support Vector Machines, seem ubiquitous for emotion classification by this method though some papers show that SVM are better than Naive Bayes systems where in a paper, it is also told that sequence labelers can outperform traditional SVM in the task of emotion analysis. These approaches broadly extract features like part of speech counts for different POS, frequency of different emoticons, frequency of punctuation and counts regarding to words such as longest and shortest word length etc. as internal features as well as affective keywords labeled as external features.

Also in recent days, also a lot of work has been done on audio or speech emotion analysis domain. A model, designed by Fayek et al. [20], uses DNN for emotion recognition based on two types of database namely eNTERFACE05 and SAVEE and MFCC features of audio has been extracted. The result is the classification into 6 types of emotion where it gives an average accuracy for both the databases. But if the numbers of neurons are increased in DNN, then computation time increases along with degradation of performance. Audio clip from movie, news channel etc. has been used as dataset, along with MFCC for feature extraction, by Kagalkar et al. [21] where SVM and GMM classifiers are used for emotion prediction. Another model, using real time CNN to detect emotions, is proposed by Fung et al. [22] which can classify emotions into three categories namely angry, happy and sad. But here numbers of detected emotions are very less. Another system, analyzing speech content through text classification, is proposed by Ezzat et al. [23] where the focus is on text mining by converting audio into text to detect emotion. But this model could not attain good classification results on thoroughly correct transcription by unsupervised learning techniques. A new model, proposed by Huang et al. [24], uses five layers DBNs technique for feature extraction to classify emotion in four types of emotion upon 'BUAA-DB' dataset by using non-linear SVM as classifier but the drawback is that the time cost of DBNs model is much longer than other feature extraction techniques. Another interesting approach, proposed by Chavhan et al. [25], uses MFCC and MEDC feature extraction on 'Berlin-DB' dataset with the help of SVM classifier to classify emotion in four categories. Here the model is done with three types of input speech signal (gender independent, male and female) giving 100% accuracy for female speech. Also a model, proposed by Gayar et al [26], tries to analyze sentiments from speech by using text classification where audio is converted into text to use extracted keywords for distinguishing positive and negative polarity but with lower accuracy. This paper's importance lies in the fact that it is performed on artificially generated dataset and there is also the chance of lower accuracy when the same model will perform on some natural dataset because of unsupervised speech recognition system. Another approach of using acoustic and lexical features on 'USC-IEMOCAP' dataset is discussed by Chen et al. [27] in his paper. Last but not the least, a useful approach of using PCA-DCNNsSER technique on 'IEMOCAP' database is proposed by Zheng et al. [28] which is showing better result than the SVM for emotion classification. But calculated accuracy is less due to improper distribution of emotion based data in the mentioned dataset.

In this context, some audio features can be discussed. Generally, at first a high pass filter suffices the need of holding high frequency baud by increasing its amplitude and decreasing the amplitude of lower frequency as typically higher frequency holds more important information to extract while noise can be mingled along with lower frequency. But this can be replaced by channel normalization method. Then the audio signal is sliced into short speech sequences, called frames and the Hamming window is often chosen for windowing as it softens the edges which are created due to framing. Normally audio features can be categorized into four groups, namely continuous, qualitative, spectral and TEO-based features. But the spectral features are mostly used due to their better prediction accuracy. Generally, Linear Predictive Coding (LPC) works for encoding an analog signal into digital format. It generally predicts the value of next sample of the signal based on the information

of preceding samples, thus forming a linear pattern along with the objective of obtaining a set of predictor coefficients minimizing the mean squared error. It is generally used for speech reformation as well as for training high quality audio at low bit rate but it is often combined with other feature extraction methods to get better outcome. Linear Prediction Cepstral Coefficient (LPCC) is also used as feature extraction method which generally uses auto-correlation technique. It is successfully used into characterizing vowels but also comes with a disadvantage of significant sensitivity towards quantization error. Another non-linear based feature extraction method, Teager Energy Operator (TEO), is used for suitably detecting stress levels of emotion. Last but not the least, one of the most popular feature extraction methods is Mel Frequency Cepstral Coefficient (MFCC) which is better for N-way classification. It includes various steps and among them one is mel frequency scale utilization which is normally tuned to the human's ear frequency response and so it has an immense application in this domain. Generally a feature called cepstrum, a windowed short term signal, is derived from the FFT of the speech signal and then with the help of a log based transform, the signal fits in the frequency axis of the mel frequency scale and finally decorrelated with the help of modified Discrete Cosine Transform (DCT). But it is somewhat susceptible towards noise.

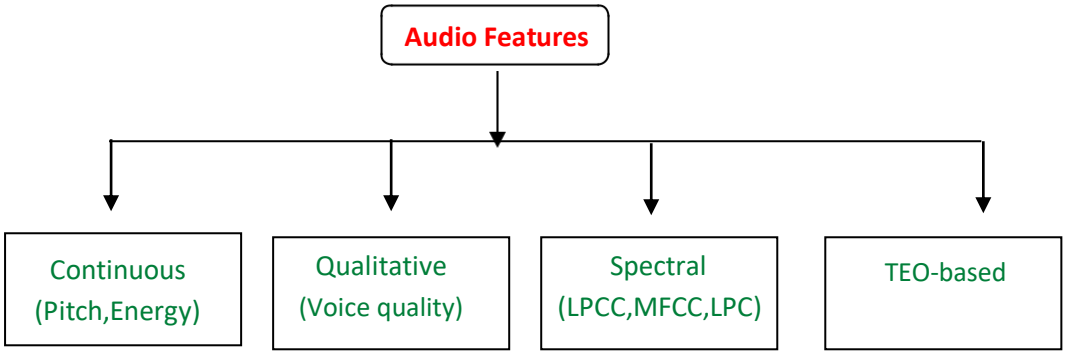

Fig.2.1 Various types of audio features

Also, a little bit of discussion has been required about the classifiers which are generally used for audio data analysis. Some popular classifiers in this case may be GMM, HMM, DNN which is an extended version of ANN. The Hidden Markov Model consists of the first order markov chain whose states are hidden from the observer. It is generally a sequential generating probabilistic model which is valid for speech signal frames but problems such as computational complexity, need of model parameters' initialization in a proper way before training are also present. The Gaussian Mixture Model can also be used but it is far more suited for global features. Deep Neural Network can also be used as it is sometimes more useful in learning high level invariant features from the data. Also, Vector Quantization (VQ) , Support Vector Machine (SVM) can also be used as classifiers.

Not only those, but also some efforts have been made toward implicit affective tagging of multimedia content. A method, proposed by Kierkels et. al. [8], discusses a way of personalized affective tagging of multimedia content from peripheral physiological signals using linear regression. Also "personalized content delivery", as a tool for affective indexing, proposed by Hanjalic and Xu [9] as well as a scene affective characterization using a Bayesian framework, proposed by Soleymani et.al. [10] are some important addition in this

context. Also a multimodal database related approach, in which the database of the electroencephalogram (EEG) and peripheral physiological signals of 32 participants are recorded, can be discussed in this scenario. Here, each of the participants watches 40 one minute long excerpts of music videos and also rates each of the video in terms of the levels of arousal, valence, dominance, like/dislike with the help of valence-arousal scale of Russell. A novel method for stimuli selection is proposed using the retrieval of affective tags from the last.fm website along with video highlight detection along with an online assessment tool, normally the well known self-assessment manikins (SAM). An analysis of the participants' ratings during the experiment is presented where the correlation between the EEG signal frequencies and the participants' ratings are judged. Some methods and results are presented for single trial classification of arousal, valence, like/dislike ratings using the modalities of EEG, peripheral physiological signals and multimedia content analysis. Finally decision fusion of the classification results from different modalities is performed and a more robust decision is made.

Table 2.1 Existing works

Prior Work	Domain of Dataset	Approach Summary	Features Used/Lexicons
Zhang et. al. (2018)	EMO-DB	SVM	MFCC
	RML		
	eNTERFACE05		
	BAUM-1s		
Ahmad et. al. (2017)	Berlin dataset	CNN Model	MFCC,LPC,LPCC
Dario Bertero, Pascale Fung (2017)	Data collected through an ongoing annotation project	CNN	FFT
Kang and Ren (2016)	Chinese blogs	Hierarchical Bayesian Models	
Bertero et. al. (2016)	TED-LIUM-DB	CNN	
Fayek et. al. (2016)	eNTERFACE05	DNN	MFCC
	SAVEE		
Shivaji J. Chaudhari, Ramesh M. Kagalkar (2015)	Audio clip from movie, news channel, some dialog, and recorded Audio clip	SVM and GMM	MFCC

Prior Work	Domain of Dataset	Approach Summary	Features Used/Lexicons
Chen et. al. (2015)	USC-IEMOCAP	SVM	Acoustic and lexical features
Zheng et. al. (2015)	IEMOCAP	PCA-DCNNsSER	Acoustic features
Huang et. al. (2014)	BUAA-DB	SVM	DBNs
Gayar et. al. (2012)	36 recorded audio files	Decision tree Naïve bayes SVM K-nearest	MFCC
Chaffar and Inkpen (2011)	News blogs health diary posts	Supervised Classification (Naive Bayes, SVM, Decision Trees)	Bag of words features; Word Net Affect
Ezzat et. al. (2010)	Call centre database of 36 audio files	HC KNN SVM Naïve Bayes	MFCC
Hassan et. al. (2010)	500 utterances spoken by actors	J48 CART Naïve Bayes K MLP	MFCC,LPC,LPCC
Bellegarda (2010)	News	Supervised (NB, SVM, Decision Trees); New approach based on LSM	Bag of words features
Ghazi et. al. (2010)	Blogs	Hierarchical Classifiers	Bag of words features

Prior Work	Domain of dataset	Approach Summary	Features Used/Lexicons
Strapparava and Mihalcea (2008)	News	Supervised (Naive Bayes) & Unsupervised (LSA) Classification	Lexical features, Word-Net Affect
Alm (2008)	Stories	Supervised Classification	Bag of words features; Roget's thesaurus
Aman and Szpakowicz (2007)	Blogs	Supervised classification (Naive Bayes, Support Vector Machines)	Word-Net Affect, General Inquirer, Features such as presence of punctuations (?,! etc.)
Yang et. al. (2007)	Blogs	Supervised classification (SVM, CRF)	Custom-made lexicon
Strapparava and Mihalcea (2007)	News headlines	Rule-based; Unsupervised Classification	Word-Net Affect
Chul Min Lee,Srikanth S. Narayanan (2005)	7200 utterances of calls	LDC KNN	BASE F10 F15 PCA
Liu et. al. (2003)	Emails	Real-world knowledge concept models (Commom Sense Affect)	
Feng Yu et. al. (2001)	Chinese Teleplays-DB	SVM	Pitch Statistic
Olveres et. al. (1998)		Rule Based (Natural Language Parsing)	

CHAPTER 3

Methodology

This chapter basically presents and describes the overall process of the thesis work. The problem statement of the work is presented which mainly is based upon emotion analysis of multimodal dataset. This chapter also describes how the dataset can be presented and how the pre-processing of the dataset can be done to fit into the model obviously after doing feature extraction.

3.1 Problem Statement

The main goal of this work is to predict various emotion categories which is generally done according to six basic types of emotion according to Ekman namely anger, disgust, fear, joy, sadness, surprise along with neutral type of emotion. Generally, here the input is any multimodal dataset consisting of corpus of texts, audio, video, audiovisual, image etc. Here MELD, a multimodal multi party database, has been selected as the working dataset which itself is an efficient expansion of EmotionLines dataset. Here, for this dataset, the number of modalities present in training, development and test corpus is three, which can be expressed as {a,v,t} where a, v, t normally represent audio, video and text data respectively.

After that, various different types of methods, like 300 dimensional GloVe vectors as well as popular toolkit openSMILE etc. are deployed to extract various textual and audio features. Now at the next part, baseline model such as bcLSTM are applied along with previously extracted features to distinguish the input dataset into previously mentioned seven types of emotion categories along with other resultant statistics (confusion matrix, weighted average, f1-score etc.)

3.2 A Workflow

In this section, the overall architecture of the work is shown and the task of every module is discussed. It also shows the flowchart of the work i.e. how the output of one module can be used as input to another module. A figure is shown to describe the overall work process in a short. It can be divided into some modules namely data collection, data preprocessing, data annotation, feature extraction, model deployment and model accuracy analysis.

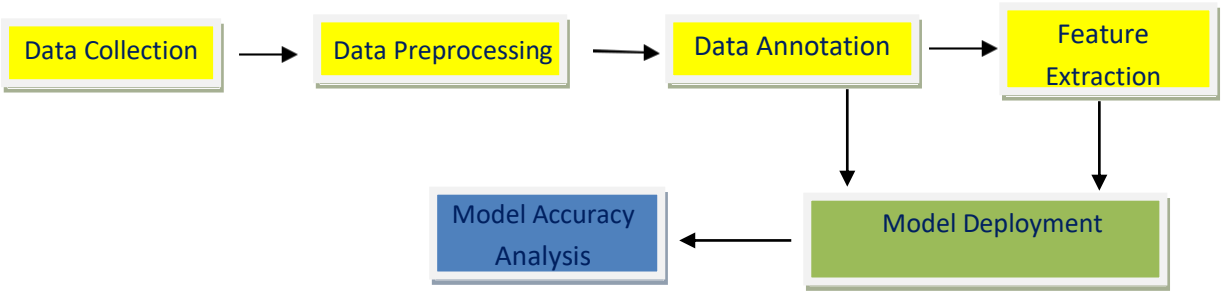

Fig. 3.1 Workflow Chart

- **Data Collection:-** MELD (Multimodal Emotion Lines Dataset) is an extended as well as improved version of 'EmotionLines' dataset which is a dataset containing dialogues of more than two speakers and for that, 'Friends", a popular TV-series is considered from where multimodal data has been collected. There are more than 13,000 utterances in MELD along with their corresponding visual and audio counterparts.
- **Data Preprocessing:-** In this part, the starting and ending timestamps of all utterances from every dialogue in the 'EmotionLines' dataset can be extracted by heuristically extracting the respective timestamp from the subtitles of the episodes. Also information like season ID, episode ID etc. Are also taken into consideration. Generally, two criterion are enforced for this process which are- 1) utterances in a dialogue must belong to the same episode and 2) an increasing order must be maintained for timestamp collection of the utterances in a dialogue. After that timestamp collection of each utterance is done, audiovisual clips for respective timestamp can be collected. Also audio part of those clips can be extracted and they can be formatted as 16-bit PCM WAV files for further processing. Thus, a dataset can be built where all textual, audio and video modalities are present. Not only that, but also a transcription alignment tool 'Gentle' is deployed to find the accurate timestamp for each utterance. This transcription alignment tool automatically aligns a transcript with the audio and for that, word-level timestamps from the audio are extracted.
- **Data Annotation:-** Generally the annotation is done according to the Ekman's six universal emotions (Joy, Sadness, Fear, Anger, Surprise and Disgust) along with Neutral emotion label. Normally, three or five annotators are employed to label each utterance with an emotion label. Then, a majority voting scheme can be applied to select a final emotion label for each utterance and for that, some utterances along with corresponding dialogues are removed where all the annotators are labeling different i.e. majority voting scheme cannot be applied.
- **Feature Extraction:-** In this part, some feature extraction approaches can be discussed as it is needed to select significant and non-redundant attributes for better classification accuracy. For textual features, each token, initialized with pre-trained 300-dimensional GloVe vectors, is fed to a 1D-CNN so that 100 dimensional textual features can be extracted. For audio features, openSMILE can be used which has a great role in extracting 6373 dimensional features which constitute of several low-level descriptors and various statistical functions of varied vocal features. L2-based feature selection with sparse estimators (like SVM) can be used to get dense representation of audio features. Bimodal features are obtained by concatenating audio and textual features. In this case, the visual features is not used as it is sometimes difficult for processing to extract such features which can be used for better classification accuracy than bimodal features.
- **Model Deployment:-** After creation of multimodal dataset is done along with annotation and feature extraction; several baseline algorithm models like text-CNN, bcLSTM, DialogueRNN can be deployed on unimodal text, unimodal audio and bimodal data separately. Concatenation has been used here as feature fusion approach for bimodal variant creation from the unimodal variants.
- **Model accuracy analysis:-** Sometimes the parameter as well as hyper-parameters tuning of the model should be done to get the better performance and more accuracy.

In this case, weighted f1-score, confusion matrix etc. have been taken into consideration for the purpose of analyzing the result given by the model.

3.3 Multi modal Data Collection

MELD (Multimodal Emotion Lines Dataset) can be considered as an extended as well as improved version of 'EmotionLines' dataset. It contains dialogues of more than two speakers and for that, 'Friends", a popular TV-series is considered from where each dialogue containing utterances from multiple speakers have been collected and it is grouped based on the number of utterances. Four groups are created by no of utterances [5, 9], [10, 14], [15, 19], and [20, 24]. There are more than 13,000 utterances in MELD along with their corresponding visual and audio counterparts. Information like season ID, episode ID, timestamp etc. have been collected from the subtitles of TV series 'Friends' which plays an important role into MELD dataset creation. Some utterances along with their counterparts are dropped at the time of annotation as majority voting system fails to label them into a certain category of emotion. Several key statistics of the MELD dataset are also considered such as the average utterance length or number of words in an utterance which is nearly the same across train, development, and test splits; number of unique words; number of dialogues; number of utterances; average number of utterances per dialogue; average number of emotions per dialogue which is three; the average duration of an utterance which is 3.59 seconds and also the number of emotion shifts in successive utterances which is very frequent (4003, 427, and 1003 in train, development and test splits respectively).

Table 3.1 Some statistics of dataset

Statistics of Dataset	Train	Dev	Test
Number of unique words	10643	2384	4361
Number of Dialogues	1039	114	280
Number of utterances	9989	1109	2610
Number of speakers	260	47	100
Average duration of an utterance	3.59 sec	3.59 sec	3.58 sec

3.4 Discussion on Approach

1) bcLSTM is generally a baseline model which uses a bi-directional RNN approach for context representation. It does not distinguish among different speakers and models a conversation as a single sequence.

2) Normally a two-step hierarchical process is followed into it which at first models unimodal context features and then bi-modal context features.

3) For unimodal text, this model extracts contextual representations for each utterance taking the GloVe embeddings as input which is then fed to 1D-CNN model.

4) For unimodal audio, an LSTM model gets audio representations for each audio utterance feature vector which is extracted using openSMILE and L2-based feature selection method.

5) Finally, the contextual representations from the unimodal variants are used for the bimodal model for classification. Here concatenation has been used as feature fusion approach for bimodal variant creation from the unimodal variants.

CHAPTER 4

Result and Analysis

This chapter describes the experimental setup i.e. in which configuration of the machine, the working model is run. Generally high end computer setup is more preferable for better accuracy for classification, but here normal setup has been used. This chapter also analyses the result which is derived at the time of dealing the emotion classification model using some baseline algorithm model.

4.1 Experimental Setup

The experiment for emotion classification using multimodal data is done on Windows 10 Home 64 bit version with 8th Gen Intel® Core™ i5-8300H @ 2.30GHz × 8 CPU with 8GB RAM machine.Python language and libraries like numpy, pandas, keras, sklearn etc. have been used. Numpy is useful for applying complex mathematical function, linear algebra operation etc. to the data. From keras, various metrics like adam optimizer, LSTM, Conv2D etc. are taken whereas from sklearn, some metrics like confusion_matrix, precision_recall_fscore_support, accuracy_score etc. have been taken.

4.2 Result

Generally after taking multimodal data into consideration, it can be seen that Fleiss' kappa score after data annotation is greater which indicates the utility of using multimodal data. Normally, there are some columns related to dataset in the csv files which can be mentioned like- Sr No. (Serial numbers used to point utterances in case of multiple copies with various subsets), Utterance, Speaker, Emotion, Dialogue_ID (starting with zero), Utterance_ID (index of utterance of any particular dialogue starting with zero), Season (season number of the show for any particular utterance), Episode (episode number of the show for any particular utterance), StartTime and EndTime referring starting timestamp and ending timestamp of the utterance respectively. Generally, utterances along with emotion labels in the training set, development set and test set are contained in the 'train_sent_emo.csv', 'dev_sent_emo.csv' and 'test_sent_emo.csv' respectively. There are some pickle files related with it and mainly they are 'text_glove_average_emotion.pkl' related with 300 dimensional textual feature vectors of each utterance, 'audio_embeddings_feature_selection_emotion.pkl' related with 1611 or 1422 dimensional audio feature vectors of each utterance which are extracted by openSMILE following L2-based feature selection using SVM. Some others are also present

Like ‘text_glove_CNN_emotion.pkl’ related with 100 dimensional textual features which are obtained after training on a CNN, ‘text_emotion.pkl’ representing textual feature vector for unimodal bcLSTM model, ‘audio_emotion.pkl’ representing audio feature vector for unimodal bcLSTM model etc. Various types of emotion, represented with one-hot encoding, are labeled like 'neutral' as 0, 'surprise' as 1, 'fear' as 2, 'sadness’ as 3, 'joy' as 4, 'disgust' as 5 and 'anger' as 6.
Confusion Matrix at the time of testing bimodal data for emotion classification:

Table 4.1 Confusion Matrix

	Neutral	Surprise	Fear	Sadness	Joy	Disgust	Anger
Neutral	982	53	0	65	99	0	57
Surprise	64	128	0	4	40	0	45
Fear	22	2	0	6	6	0	14
Sadness	90	9	0	48	14	0	47
Joy	92	27	0	10	214	0	59
Disgust	28	8	0	5	5	0	22
Anger	75	32	0	17	53	0	168

CHAPTER 5

Conclusion and Future Work

In general, an overview of the work and its prospects can be discussed in this part along with some limitations. One of the limitations of emotion analysis model is that it cannot always precisely classify the actual feelings of a user as they are fuzzy in nature as well as emotion shifts happen around in a short time lapse which is sometimes difficult to manage. Sometimes, the quality of the training data and the way emotion labels are assigned, they can strongly influence the results of the training algorithm. Also the accuracy of various models implemented for doing it does not fulfill the requirements to implement it at a large scale. Last but not the least, classification of short utterances like “yeah”, “okay”, “no” is very difficult as various emotions can be expressed by this type of utterance as they mostly depend on the context of the dialogue.

5.1 Conclusion

In this work, multimodal dataset MELD has been taken into consideration for emotion classification task. MELD is an extended version of ‘EmotionLines’ dataset which includes textual, audio and video data for better and accurate prediction of emotion according to Ekman model of emotion classification. For that, after data preprocessing phase like collection of utterance with other modalities according to timestamp, data annotation is done according to majority voting scheme. After that, various methods are deployed for feature extraction of textual and audio data and bcLSTM algorithm can be used for the purpose of emotion classification. This work can also be extended under other algorithms in the search of better classification.

5.2 Future Work

In future, a try to improve accuracy as well as f1-score for this classification will be made. Even identifying emotion shifts as well as extracting features from visual data will be very much promising. Also, an efficient improvement into the feature extraction method of audio data can be done so that the performance of emotion classification using multi modal data can be improved more. Not only those, but also better or advanced fusion methods instead of using concatenation as method of fusion for the creation of bimodal variant from the unimodal variants can be deployed. In future, this application can have a lot of use for real time personal assistants (Alexa, Google Assistant etc.) which can elevate the user's perception.

Bibliography

1. Koelstra, Sander & Mühl, Christian & Soleymani, Mohammad & Lee, Jong-Seok & Yazdani, Ashkan & Ebrahimi, Touradj & Pun, Thierry & Nijholt, Anton & Patras, Ioannis. (2011). DEAP: A Database for Emotion Analysis Using Physiological Signals. IEEE Transactions on Affective Computing. 3. 18-31. 10.1109/T-AFFC.2011.15.

2. Aman, Saima & Szpakowicz, Stan. (2007). Identifying Expressions of Emotion in Text. 196-205. 10.1007/978-3-540-74628-7_27.

3. Chaffar, Soumaya & Inkpen, Diana. (2011). Using a Heterogeneous Dataset for Emotion Analysis in Text. 62-67. 10.1007/978-3-642-21043-3_8.

4. Agarwal, Apoorv & Xie, Boyi & Vovsha, Ilia & Rambow, Owen & Passonneau, Rebecca. (2011). Sentiment Analysis of Twitter Data. Proceedings of the Workshop on Languages in Social Media.

5. Kołakowska, Agata & Landowska, Agnieszka & Szwoch, Mariusz & Szwoch, Wioleta & Wróbel, Michał. (2014). Emotion Recognition and Its Applications. Advances in Intelligent Systems and Computing. 300. 51-62. 10.1007/978-3-319-08491-6_5.

6. P. Ekman, W.V. Friesen, M. O'Sullivan, A. Chan, I. Diacoyanni-Tarlatzis, K. Heider, R. Krause, W.A. LeCompte, T. Pitcairn, and P.E. Ricci-Bitti, "Universals and Cultural Differences in the Judgments of Facial Expressions of Emotion," J. Personality and Social Psychology, vol. 53, no. 4, pp. 712-717, Oct. 1987.

7. R. Plutchik, "The Nature of Emotions," Am. Scientist, vol. 89, p. 344, 2001.

8. J. Kierkels, M. Soleymani, and T. Pun, "Queries and Tags in Affect-Based Multimedia Retrieval," Proc. Int'l Conf. Multimedia and Expo, pp. 1436-1439, June 2009.

9. A. Hanjalic and L.-Q. Xu, "Affective Video Content Representation and Modeling," IEEE Trans. Multimedia, vol. 7, no. 1, pp. 143-154, Feb. 2005.

10. M. Soleymani, J. Kierkels, G. Chanel, and T. Pun, "A Bayesian Framework for Video Affective Representation," Proc. Int'l Conf. Affective Computing and Intelligent Interaction, pp. 1-7, Sept. 2009.

11. J.A. Russell, "A Circumplex Model of Affect," J. Personality and Social Psychology, vol. 39, no. 6, pp. 1161-1178, 1980.

12. Strapparava, Carlo, and Rada Mihalcea. "Semeval-2007 task 14: Affective text." Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007). 2007.

13. Jerome R Bellegarda. 2010. Emotion analysis using latent affective folding and embedding. In Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text, pages 1–9. Association for Computational Linguistics.

14. Saima Aman and Stan Szpakowicz, 2007a. Identifying Expressions of Emotion in Text, pages 196–205. Springer Berlin Heidelberg, Berlin, Heidelberg.

15. Soumaya Chaffar and Diana Inkpen, 2011. Using a Heterogeneous Dataset for Emotion Analysis in Text, pages 62–67. Springer Berlin Heidelberg, Berlin, Heidelberg.

16. R. Picard. 1997. Affective Computing. The MIT Press.

17. Hugo Liu, Henry Lieberman, and Ted Selker. 2003. A model of textual affect sensing using real-world knowledge. In Proceedings of the 8th international conference on Intelligent user interfaces, pages 125–132. ACM.

18. Ebba Cecilia Ovesdotter Alm. 2008. Affect in Text and Speech. Ph.D. thesis, University of Illinois at Urbana-Champaign.

19. Peter Mark Roget. 2008. Roget'S International Thesaurus, 3/E**. Oxford and IBH Publishing.

20. Fayek, Haytham M., Margaret Lech, and Lawrence Cavedon. "Towards realtime speech emotion recognition using deep neural networks." In 2015 9th international conference on signal processing and communication systems (ICSPCS), IEEE (2015).

21. Chaudhari, Shivaji J., and Ramesh M. Kagalkar. "Automatic speaker age estimation and gender dependent emotion recognition." International Journal of Computer Applications 117, no. 17 (2015).

22. Bertero, Dario, and Pascale Fung. "A first look into a convolutional neural network for speech emotion detection." In 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP), IEEE (2017).

23. Souraya Ezzat, Neamat El Ghayar and Moustafa M. Ghanem. "Investigating analysis of speech content through Text Classification". International Conference of Soft Computing and Pattern Recognition (2010).

24. Huang, Chenchen, Wei Gong, Wenlong Fu, and Dongyu Feng. "A research of speech emotion recognition based on deep belief network and SVM." Mathematical Problems in Engineering 2014 (2014).

25. Chavhan, Yashpalsing, M. L. Dhore, and Pallavi Yesaware. "Speech emotion recognition using support vector machine." International Journal of Computer Applications 1, no. 20 (2010).

26. Souraya Ezzat, Neamat El Gayar, and Moustafa M. Ghanem. “Sentiment Analysis of Call Centre Audio Conversations using Text Classification”. International Jornal of Computer Information Systems and Industrial Management Applications, volume 4 (2012).

27. Jin, Qin, Chengxin Li, Shizhe Chen, and Huimin Wu. "Speech emotion recognition with acoustic and lexical features." In 2015 IEEE international conference on acoustics, speech and signal processing (ICASSP), IEEE (2015).

28. Zheng, W. Q., J. S. Yu, and Y. X. Zou. "An experimental study of speech emotion recognition based on deep convolutional neural networks." In 2015 international conference on affective computing and intelligent interaction (ACII), IEEE (2015).

29. Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E. and Mihalcea, R., 2018. Meld: A multimodal multi-party dataset for emotion recognition in conversations. arXiv preprint arXiv:1810.02508.

30. Tripathi, V., Joshi, A. and Bhattacharyya, P., 2016. Emotion analysis from text: A survey. Center for Indian Language Technology Surveys, 11(8), pp.66-69.

31. Tripathi, A., Singh, U., Bansal, G., Gupta, R. and Singh, A.K., 2020, May. A review on emotion detection and classification using speech. In Proceedings of the International Conference on Innovative Computing & Communications (ICICC).

32. Gunawan, T.S., Alghifari, M.F., Morshidi, M.A. and Kartiwi, M., 2018. A review on emotion recognition algorithms using speech analysis. Indonesian Journal of Electrical Engineering and Informatics (IJEEI), 6(1), pp.12-20.