Proceedings of the Fifth International Conference on Computing Methodologies and Communication (ICCMC 2021)
IEEE Xplore Part Number: CFP21K25-ART

Human Speech and Text Emotion Analysis

A Survey on Emotion Analysis using Deep Neural Networks

Sweta Singh
singh.sweta254@gmail.com

Abhishek Badoni
badoniabhishek@gmail.com

Mrs. K. Deeba
deebak@srmist.edu.in

Computer Science and Engineering,
SRM Institute of Science and Technology
Tamil Nadu, India

2021 5th International Conference on Computing Methodologies and Communication (ICCMC) | 978-1-6654-0360-3/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICCMC51019.2021.9418474

***Abstract*—Human Emotion Analysis is an extremely wide area of study which can have several implementations for various applications. To work on any kind of project related to Emotion detection it is imperative to collect information on the available models, datasets and performance statistics related to them. This review analysis aims to consolidate a set of unique as well as state-of-the-art approaches that have been proposed and/or carried out by various individuals in this field. The proposed research work has consolidated studies across different datasets majorly focusing on the standard The Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset and Social media datasets. Our survey spans multiple modes of data such as text, speech, speech transcriptions [6] and motion capture [10] to provide a more in-depth analysis of the accuracies and improvement achieved across mentioned modes of data. Models based on Convolution Neural Networks (CNN) [6][8] and Recurrent Neural Networks(RNN)[10] have been reviewed upon among others reflecting the efficiency of each model.**

Keywords— Emotion Analysis, Speech, Text, Machine Learning, Neural Networks

I. *INTRODUCTION*

Human Emotions plays an important role in the world. A major aspect of human communication depends on what emotions are conveyed during a particular conversation. Emotions during these conversations can be detected through vocal attributes such as pitch, timbre, sound pressure level (SPL) and the time gap between words [1]. Any attempt to develop models which could detect emotions need a gold-standard of feature list that has been derived from trained human experts who are consistent in their speech. A similar methodology was adopted during the construction of Valence Aware Dictionary for Sentimental Reasoning (VADER) [2] which is a widely used sentimental lexicon today.

Emotion Analysis has various existing as well as emerging applications. To widen the scope of application it can be considered that the modes of data that can be used for emotion analysis should not be restricted only to either text or speech.

The text provides us with semantics that could be associated with the context however, speech is also important as it provides several low-level speech features that could make a model more robust and efficient [6]. Apart from text and speech, Motion Capture datasets which include facial expressions, hand and head movements can be studied to provide further context to emotion detection methodologies [10].

Multiple modes of data such as speech, text as well as motion can be obtained from the IEMOCAP dataset which has been widely used across this review [5] [6][10]. This dataset was collected at the University Of Southern California (USC) by SAIL (Signal Analysis and Interpretation Laboratory). It is a multi-speaker and multimodal database which contains audiovisual data including speech, text, video as well as motion capture of the face of the actors.

Human Emotion analysis can be used to analyze the performance of a movie based on online movie reviews [9]. Additionally, a wealth of information such as hate tweets [7] as well as bogus statistics [8] can be extracted using emotion detection from social media sites. These can help us in regulating instigation through social media platforms. In medical fields, emotion detection techniques can be extended to help study anxiety as well as depression [3].

II. *LITERATURE SURVEY*

A. Detection and Analysis of Human Emotions through Voice and Speech Pattern Processing (2017)

Authored by: *Poorna Banerjee Dasgupta*

This paper [1] has its base in the deduction that voice and speech analysis can help determine human emotions. Here, the author makes use of the same to propose an algorithmic approach to analyze the various vocal attributes during the speech.

Approach

Human speech is characterized by several vocal attributes, primarily pitch, loudness, timbre and tone. While conversing the tonal quality of the speech changes with a change in emotions which is reflected in one or more of the vocal

attributes. Taking this into consideration pitch, sound pressure level (SPL), timbre and time gaps have been taken into account in this paper as the attributes that would be used to measure and differentiate between different emotional states. Three test cases have been taken into account for examination. The states being: *normal*, *angry* and *panicked* emotional state. Here the normal state is taken as a base for the other two states and two speech samples are taken for the same measuring the pitch, SPL, timbre and time gaps. The panicked and angry speech samples reflect orations in a panicked or overwhelmed state and an angry or agitated state.

On comparing Table I with Table II [1]it can be seen that the pitch and the mean SPL values increase for the angry state. Additionally, time gaps between words decrease which indicates hurried speech which is shrill and loud.

On comparing Table I with Table III [1] we can note that in a panicked state the mean pitch values, time gaps as well as timbre ascend time increases for the normal state. The increase in time gap indicates longer pauses and the increase in pitch indicates a shrill voice.

Inference

It can be concluded from the analysis of Tables I-III [1]that there is a visible change in the tonal parameters with a change in speech emotions. By studying the deviations from the standard basis, which would be the normal state, it is possible to distinguish human emotions from vocal attributes.

TABLE I.

VOCAL STATISTICS OBTAINED FOR A NORMAL EMOTIONAL STATE

	Pitch(Hz)	**SPL(dB)**	**Timbre Ascend Time(s)**	**Timbre Descend Time(s)**	**Time gaps between words(s)**
Speech Sample I	1248 Hz	Gain -50 dB	0.12 s	0.11s	0.12s
Speech Sample II	1355 Hz	Gain -48 dB	0.06 s	0.05s	0.12s

TABLE II.

VOCAL STATISTICS OBTAINED FOR FOR AN ANGRY EMOTIONAL STATE

	Pitch(Hz)	**SPL(dB)**	**Timbre Ascend Time(s)**	**Timbre Descend Time(s)**	**Time gaps between words(s)**
Speech Sample I	1541 Hz	Gain -30 dB	0.12 s	0.10s	0.09s
Speech Sample II	1652 Hz	Gain -29 dB	0.06 s	0.04s	0.10s

TABLE III.

VOCAL STATISTICS OBTAINED FOR A PANICKED EMOTIONAL STATE

	Pitch(Hz)	**SPL(dB)**	**Timbre Ascend Time(s)**	**Timbre Descend Time(s)**	**Time gaps between words(s)**
Speech Sample I	1433 Hz	Gain -46 dB	0.13 s	0.09s	0.13s
Speech Sample II	1560 Hz	Gain -44 dB	0.07 s	0.04s	0.14s

B. VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text (2015)

Authored by: *C.J Hutto, Eric Gilbert*

VADER also know as Valence Aware Dictionary for sEntimental Reasoning is one of the most popular models that is widely used for simple emotion analysis. This is the paper that presents VADER and compares it to eleven state-of-the-art benchmarks such as LIWC, the General Inquirer, SentiWordNet, ANEW, and techniques based on Maximum Entropy, Support Vector Machine (SVM), and Naive Bayes algorithms.

Approach

The aim of the study to report upon the following:

I. Constructing and Validating a Valence-Aware Sentiment Lexicon: A Human-Centered Approach

Around 9000 features candidates are obtained by analyzing well-established word banks such as ANEW, LIWC and GI and adding lexical features such as emoticon, acronyms and slangs that reflect emotions. Each of these feature candidates is then rated using Wisdom-of-the-Crowd (WotC) approach for sentiment expressions. The rating is carried out on a scale of -4 to 4 where -4 denotes extremely negative, 4 denotes extremely positive and 0 denotes neutral emotion. The rating was done with the help of a micro-labour website, Amazon Mechanical Turk (AMT). 10 screened human raters carried out this task All the lexical features that had a non-zero rating and had a standard deviation of less than 2.5 aggregate was kept and the others were eliminated. This process yielded 7,500 features that had a validated sentiment polarity. The resultant list of features serves as the gold-standard of features that is used in the sentiment lexicon for VADER.

II. Identifying Generalizable Heuristics Humans Use to Assess Sentiment Intensity in Text

Around 10K tweets were pulled from Twitter, among which 400 positives and 400 negative snippets were extracted. These snippets were analyzed using the Pattern sentimental analysis engine. The extracted 800 snippets were then rated independently by two human experts who rated them on a scale of -4 to 4. An approach similar to the Grounded Theory

approach was used to determine the intensity of sentiment in the texts which led to five generalizable heuristics.

a) Punctuations such as the exclamation mark (!) can increase the intensity of a text.

b) Similar to the exclamation mark, capitalization also increase the intensity of the emotion the capitalized keyword conveys.

c) Intensifiers or degree modifiers can increase or decrease the intensity of any given text.

d) Contrast conjunctions such as "but" indicates a shift in the preexisting emotion of the text, with the latter half of the text being more dominating and dictating the overall review

e) On analysis of tri-gram preceding a sentiment reflective text, it has been determined that in around 90% of the text negation flips the polarity of the text.

III.Controlled Experiments to Evaluate Impact of Grammatical and Syntactical Heuristics

Further based on the heuristics that were identified before 30 tweets are pulled and around six to ten variations of the tweets are made by controlling the grammatical and syntactical meanings. This gave around 200 snippets which were added to the 800 snippets extracted before and then rated by human experts on a scale of 1 - 4. The result of the 1000 rated snippets shows that 95% of the data with an exclamation point increase the intensity by 0.261 to 0.322.

Inference

VADER sentiment lexicon on comparison with other seven well-known sentiment lexicons General Inquirer (GI), Linguistic Inquiry Word Count (LIWC), Affective Norms for English Words (ANEW), SenticNet (SCN), Word-Sense Disambiguation (WSD) using WordNet, SentiWordNet (SWN) and the Hu-Liu04 opinion lexicon gives a Pearson Product Moment Correlation Coefficient of 0.88 which is superior to the other lexicons. On adjusting the thresholds it was further concluded that the accuracy for VADER could go up to 0.96 which is better than individual human raters(0.86). Further analysis of VADER against machine learning models proves that its performance is either at par or better than the other machine learning models in their respective trained domain (Table IV)[2].

TABLE IV. RECORDED ACCURACY OF VADER AGAINST OTHER MACHINE LEARNING MODELS TRAINED IN PARTICULAR DOMAINS.

	Tweets	**Movies**	**Amazon**	**NYT**
VADER	0.96	0.61	0.63	0.55
NB(tweets)	0.84	0.53	0.53	0.42
ME(tweets)	0.83	0.56	0.58	0.45
SVM-C(tweets)	0.83	0.56	0.55	0.46
NB(movie)	0.56	0.75	0.49	0.44
ME(movie)	0.56	0.75	0.51	0.45
NB(Amazon)	0.69	0.55	0.61	0.48
ME(Amazon)	0.67	0.55	0.60	0.43
SVM-C(Amazon)	0.64	0.55	0.58	0.42
NB(nyt)	0.59	0.56	0.51	0.49
ME(nyt)	0.58	0.55	0.51	0.50

C. Study of Depression Analysis using Machine Learning Techniques (2019)

***Authored by**: Devakunchari Ramalingam, Vaibhav Sharma, Priyanka Zar*

Human emotion can be perceived as a state of a person and by extension can also be perceived as the state of the human mind in cases. Here[3], depression is a state of mind that needs to be assessed and properly examined. This paper focuses on various social media platforms serving as a dataset for the classification of depression and /or for a person having suicidal thoughts using machine learning techniques.

Approach/Information Gained

Multiple approaches have been highlighted which have been derived from several experiments.

An approach proposed by Melissa N Stolar, Margaret Lech, Shannon J Stolar, Nicholas B Allen[11] uses a model that builds upon optimized spectral roll-off parameters for depression symptom detection from clinical speech datasets. Pre-processing techniques such as Blind Source Separation is used to filter out any unnecessary cross-talk which is present as background noise. Windowing is done which normalizes a clean signal within a range of -1 and 1, based on max amplitude. Voice Activity Detector(VAD) is made to use to categorize the speech into unvoiced or voiced. Techniques such as Spectral Category(S*) and Spectral Category (S) is used for feature extraction and analysis from the speech inputs. A simple Support Vector Machine(SVM) classifier has been employed for further classification of the extracted features. (Gender dependence was seen in this study where it was more effective in males than females.)

Another study on effective content analysis of Online Depression Communities done by Thin Nguyen, Dinh Phung, Bo Dao, Svetha Venkatesh, Michael Berk [12] whose aim was to effectively identify depression stigma on online platforms such as Twitter or Weibo. These platforms served as a dataset for the study. Syntax and semantic analysis was carried out on the dataset to analyze keywords and their context in the paragraph. This is an important step to understand the overall emotion of the paragraph in question using emotion detection systems. To build the model, posts which contain stigma keywords are further classified as stigma and non-stigma based on whether they reflect stigma or not. This is done to avoid depressive disorder. Further, data modelling is improved by extracting linguistic features that could be used for differentiation between stigma and non-stigma. Classification algorithms such as Simple Logistic Regression (SLR), Multilayer Perceptron Neural Networks (MLPNN), Support Vector Machine (SVM), and Random Forest (RF) are then implemented to build classification models based on the linguistic features extracted. Equal sized subgroups were made from the existing dataset and tested

against each of the models. The results of the study showed that about 6.09% of the relevant posts on Weibo reflected depression stigma.

Inference

Vocal features extracted from clinical speech conversation can be used for depression analysis using proper pre-processing techniques. Similarly, linguistic analysis can be helpful in the detection of stigmatizing attitudes on social media. The detection of stigma on 6.09% of relevant posts was significantly higher than the 0.7% of the relevant posts which was a statistic presented by previous studies.
However, it is to be noted that depression is gender-specific and different basis needs to be defined for different genders.

D. Hate Speech Detection from Code-mixed Hindi- English Tweets Using Deep Learning Models (2018)

Authored by: *Satyajit Kamble, Aditya Joshi*

Social media has defiantly brought people closer. But it is undeniable that this has also become a platform for people to spread hate and provoke conflicts for their own/shared interest. These messages before they lead to chaos can be eliminated This project aims to detect hate speech texts in on the social media platform Twitter, focused on tweets written in Hindi or English, or a mixture of both. They aim on doing so by using and comparing three deep learning models on a benchmark dataset for code-mixed Hindi-English tweets.

Approach

*CNN model:*Domain specific embedding correlating to a sentence is fed into a CNN (one dimensional) model. The resultant is a feature map upon which a layer of global Max Pooling is used, having a dropout probability of 50%. To the final result which is in the form of a single feature vector, sigmoid activation function is applied. *LSTM model:*Considering the Sequential nature of the dataset, a comparison is made using the LSTM algorithm. The data passed through the LSTM layers are accumulated at each timestamp and the model is returned to provide a sequence of each time stamp in the process. This then undergoes through a layer of global Max Pooling and finally sigmoid activation function to get a binary prediction. Also, BilSTM model is employed to get a better understanding of the dynamics of the data for two directions and not just in a sequential manner.

Inference

The experiment provided a model with an improvement of 12% in F-Score over the preceding models. It was also observed that 1 dimensional Convolution Neural network produced the best accuracy among the other with an accuracy of 82.62% and the F-score of 80.85%. Both LSTM and BiTSTM had similar resultants but BiLSTM slightly better with an accuracy of 81.48% and an F-Score of 78.36.

E. Speech Emotion Recognition with Heterogeneous Feature Unification of Deep Neural Network (2019)

Authored by: *Wei Jiang, Zheng Wang, Jesse S. Jin, Xianfeng Han and Chunguang Li*

A novel architecture using deep neural networks is proposed which focuses on efficient extraction of features from heterogeneous acoustic feature groups which might contain irrelevant information that may lead to degradation of emotion recognition. The extracted features and then fed into a trained fusion network and further classified using a Support Vector Machine.

Approach

The proposed speech emotion recognition architecture majorly has three modules,
I. *Feature extraction module,* this module extracts audio data from video input. Open source toolkit Opensmile which has proved efficient in extracting low-level acoustic features has been used to extract features such as IS10, MFCCs and eGemaps. VGGish and SoundNet bottleneck feature act as extractors for high-level acoustic features.
II. *Heterogeneous unification module,* this module has five branches corresponding to the five acoustic features that were extracted by the feature extraction module. Due to the high dimensionality and heterogeneous nature of the features extracted, the exploitation of the intrinsic relations among them in a low-level representation becomes difficult leading to decrease in recognition performance. To overcome these hurdles, the unification module comes into play. It uses unsupervised feature learning techniques to convert the heterogeneous distinct distribution spaces of the extracted features into a unified representation space. The autoencoder structure which is a multiple layer feed-forward neural network is employed to take input the representation spaces and yield new linear-transformation at high-level spaces.
III. *Fusion Network Module* is a deep neural network module that consists of four layers, one input layer and three hidden layers. This module deploys a simple fusion strategy to provide enhanced speech emotion recognition performance. It extracts the association between the unified joint feature representation from the previous module, resulting in a 1024-dimensional feature vector which acts as the final representation for the acoustic features
The proposed architecture is tested against the Interactive Emotional Dyadic Motion Capture Database (IEMOCAP) dataset where different classifiers such as K-Nearest Neighbour (KNN), Random Forest (RF) and SVM among others have been employed.

On analysis of TABLE V[5], it can be seen that SVM performs the best among the classifiers that have been used. There is a 9% gap between the performance of SVM and KNN, due to which SVM has been used as the final classifier.

TABLE V. COMPARISON OF DIFFERENT CLASSIFIERS BASED ON CLASSIFICATION RESULTS

Classifier	KNN	LR	RF	SVM
Angry	0.56	0.64	0.66	0.65
Happy	0.71	0.73	0.77	0.79
Sad	0.38	0.39	0.46	0.45
Neutral	0.58	0.62	0.64	0.69
Total	0.55	0.59	0.63	0.64

Inference

The proposed architectures yield the best accuracy of 64% which outperforms the previous accuracy measures achieved by Lakomkin E., Wermter, S., Weber C., Magg S.. [13] and Gu et al. [14] which is 58% and 62% respectively. The statistical analysis clearly shows the difference between the existing state of art architecture and proposed architecture showcasing that this is a novel approach benefitting greatly from the unification and fusion model.

F. Deep Learning based Emotion Recognition System Using Speech Features and Transcriptions (2019)

Authored by: *Suraj Tripathi, Abhay Kumar, Abhiram Ramesh, Chirag Singh, Promod Yenigalla*

The novelty of this study lies in the fact that a combined approach using speech as well as the text has been proposed here. Spectrogram and MFCC speech features are taken into account to incorporate the emotion-related low-level characteristics and text is taken into consideration to capture the semantic meaning.

Approach

The proposed CNN model consists of speech features as well as text transcriptions. This model can be broken down into two modules:

I. *Text-based CNN Model:* Text transcription model takes in text sequence derived from the IEMOCAP dataset. The sequence is word embedded which maps the text to real numbers or vectors. Word embedding here is an ideal representation for DNN as it can predict surrounding words in a sentence and can be efficient in representing the context in which the words are distributed. The embedded vectors are convoluted with kernels of different sizes. The max-pool layer extracts features from each of the convolution layers and feeds them to the Fully Connected (FC) Layer. Finally, the classification is carried out by the softmax layer which gives output as probabilities. Batch normalization has been used in this model to prevent over-fitting and sensitivity to initial weights.

II. *Speech-based CNN Model*: Both Spectrogram, in which speech is represented over time and frequency as well as Mel Frequency Cepstrum(MFC), which represents speech as a Short-Term Power Spectrum of sound has been tested alternatively for speech input to the 2-D CNN layers. Input is feed to a set of 4 parallel 2D convolutions having 200-kernels for each parallel convolution step. The kernel sizes when spectrogram is used are 12 x 16, 18 x 24, 24 x 32, and 30 x 40, whereas the sizes when MFC is used are 4 x 6, 6 x 8, 8 x 10 and 10 x 12. The features generated from these layers are then fed into max-pools which further picks out a set of 4 features each and feeds them further into Fully Connected (FC) layers. This model implements two FC layers with batch normalization having sizes of 400 and 200. Rectified Linear Unit (ReLu) activation function has been used in the convolution layers as well as the first FC layer. The output from the final FC layer feeds the input to the softmax layer which produces the final classification.

TABLE VI. COMPARISON OF ACCURACIES

Input	Overall Accuracy	Class Accuracy
Spectrogram (Lee[17])	62.8	63.9
Spectrogram (Satt[15])	68.8	59.4
Text (proposed)	64.4	47.9
Spectrogram(proposed)	71.2	61.9
MFCC(proposed)	71.3	59.9
Spectrogram & MFCC (proposed)	73.6	62.9
Text & Spectrogram (proposed)	75.1	69.5
Text & MFCC (proposed)	76.1	69.5

Inference

An analysis between the models containing spectrogram and MFCC with and without text shows that the maximum overall accuracy is obtained for the Model containing both *Text and MFCC* which is 76.1%. Class accuracies proved to be the same for both the models containing *Text and MFCC*; and *Text and Spectrogram* which is 69.5%(Table VI)[6]. The best overall frequency here is around 7% higher than the state-of-art architecture accuracies recorded by Satt et al[15] which is 68.8%. This novel technique of using Text transcriptions as well as speech incorporates several low-level features which helps increase the accuracy of emotion detection.

G. Emotional Analysis of Bogus Statistics in Social Media(2020)

Authored by: *Dr Wang Haoxiang*

This is a unique study that dives into the world of Social Media and analyzes the emotions that are reflected behind bogus posts. Bogus statistics or fake information are widespread on social media these days and have a major impact on the user who ends up reading them. This study aims to detect the underlying emotion in these bogus posts to differentiate fake information from the truthful ones using recurrent neural network and comparing it with existing neural network techniques (Table VII)[8].

TABLE VII. THE ACCURACY COMPARISON OF THE PROPOSED MODEL WITH ANN AND DNN

Dataset	ANN	DNN	Proposed RNN
Keras	89.96	92.55	95.22
Tensorflow	88.62	92.42	94.88
SS-TDS	89.66	91.28	94.83
Average Accuracy(%)	89.41	92.08	94.97

Approach

The proposed architecture makes use of Recurrent Neural Network (RNN) as it is capable of evaluating sequences with the hidden layers helping in learning data from the previous layers. RNN makes use of the whole sequence to train the model and predict the corresponding results instead of classifying over a single perceptron. It can also process dynamic input of various lengths without affecting the model size which provides better performance on previous information. The weight function is time function based which significantly improves the computation time.
The input to the RNN is encoded using a unique integer and fed to the initial *embedding layer* which learns all the words after initializing the randomly assigned weights. The input node is connected to the output node using the *dense layer* which classifies the connected neural network layer. Random nodes in the network are activated using the *dropout layer* to avoid data overfitting.
The back-propagation algorithm is used to train the proposed RNN where repetitive connections are used to measure measures of the error functions. The chain rule is used in the place of the Time Step function to define the error. The fixed hidden layer vectors are used by long-term dependencies in a text to process the complete information. The output is filtered by the hidden states followed by the global average pooling layer which extracts the features resulting in either positive or negative output.

Inference

Comparative analysis of the proposed RNN architecture with ANN and DNN carried out on 3 different datasets show that RNN performs better than both ANN and DNN with an average accuracy of 94.7% which is 5% more than the accuracy provided by ANN and 2% more than DNN.

H. Multi-Modal Emotion recognition on IEMOCAP Dataset using Deep Learning(2018)

Authored by: *Samarth Tripathi, Homayoon Beigi*

Works done previously on emotion analysis using the IEMOCAP dataset has focused only on either or speech or speech transcriptions, however here analysis has been carried out using the different modes of data provided by the dataset including speech, text, motion capture from facial expressions as well as hand movements. This multi-modal approach has been used with the aim of more robust emotion recognition.

Approach

Models using speech, text, Motion Capture(MoCap) as well as a combined model making use of all the above have been trained and studied upon.
I. Speech-based Emotion Detection: The neural model proposed here follows the preprocessing techniques of [16]. Fourier frequencies and MFCCs are used to extract a total of 34 features and generate corresponding feature vectors. The flattened input speech vectors are trained using cross-entropy loss. Adadelta was used as an optimizer for LSTM models. 2 bidirectional LSTM having 128 units each have been used where the second BiLSTM has Attention implementation as well. The LSTM layers are followed by a Dense layer of 512 units which have Rectified Linear Unit (ReLU) activation function implementation. This FC model with three hidden layers stated above used 4 output neurons implemented in the final softmax layer to give resulting in an accuracy of 55.65% which is an improvement over accuracy present by Chernykh et al. [16] which is 54%.

II. Text-Based Emotion Detection: Preprocessing of the text input is done using Glove embedding in the model proposed for text-based analysis generating a 500,300 vector for each utterance. The vector is fed through 4 1D Convolution filters having a Dropout of 0.2 with ReLu activation function. Each of these convolution filters has 3 kernels which filter sizes of 256,128,64 and 32. The filters are followed by a 256 unit Dense layer having ReLu activation function with Dropout 0.2 as well. The final output softmax layer used 4 output neurons as the speech model. Adam is used as an optimizer in this model. The above model for text results in an accuracy of 64.78%.

III. MoCap Based Emotion Detection: The MoCap data has been processed to produce a vector of 200,189 dimensions. The MoCap based model makes use of 4, 2D Convolution Filters and 2 Dense Layers each having a dropout of 0.2. All of the convolution filters have kernel size 3 with filters of sizes 32,64,64,128 and Stride 2. The Dense Layers following the filters have 1024 and 256 units each. ReLu activation function has been used throughout the model and Adam has been used as an optimizer for the filters. The MoCap model illustrated when tested with hand, face and head movements combined gives an accuracy of 51.11%.

IV.A Combined Approach: A model which used all of the above-specified models of speech, text and MoCap was build with few alterations to provide the best results. The proposed combined model with alterations uses LSTM instead of convolution filters for text emotion detection. Recurrent Dropout has been tried out in the text as well as speech detection for better-combined results. The LSTM architecture for the text emotion detection part of the combined model contains 256 units each. The accuracies noted with different alterations lie between 67% and 71%.

Inference

This study approaches the concept of emotion detection in a multi-modal approach giving accuracies of 54% for text, 64.78% for speech and 51.11% for MoCap based emotion detection. A significant improvement is observed when all the individual methods are combined amounting to an accuracy of 71%. This concludes that multi-modal data could be used to train a more robust model

III. RESULT AND DISCUSSIONS

Social media platforms have been proven to be one of the most widely used source for solely text-based emotion analysis[3][7][8]. Models based on 1D-CNN and RNN performed well when utilized for text datasets giving an accuracy of 82.62%[7] and 94.7%[8] respectively.

Vocal attributes such as pitch, timbre, SPL and time gaps were tested for angry and panicked emotions with normal as the basis. Significant deviations for each emotion prove these attributes could be employed for further detection along with other vocal features. Speech features can also be extracted using MFCC and Spectrogram [6]. MFCC representation of speech has been observed to be used more effectively, setting it as a standard speech representation for speech emotion detection [4][5][6][10]. On comparison of MFCC with Spectrogram along with text integration, MFCC with Text (76.1%) outperforms Spectrogram with Text(75.1%)[6]. Models incorporating CNN and BiLSTM provide the best results when dealing with speech features [6][10] where CNN gives an accuracy of about 71% [6] and BiLSTM gives an accuracy of 55.65% [10]. On comparing several classifiers SVM, yields better results than its counterparts [5] making it an optimal classifier for implementation. Consolidation of results in Table VIII.

TABLE VIII. LITERATURE SURVEY

Paper Name	***Proposed Year***	***Approach***	***Application and/or Achievements***
A. *Detection and Analysis of Human Emotions through Voice and Speech Pattern Processing*[1]	2017	An algorithmic approach is taken by analyzing vocal attributes such as pitch, sound pressure level (SPL), timber and time gaps to detect emotions that might be reflected in the said attributes.	An increase in pitch was observed when in an angry or agitated state. A decrease in the time gap when angry and an increase when panicked was also observed. A visible deflection in values was noticed where the normal state metrics were taken as a base which proved emotion detection possible through vocal attributes.
B. *VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text* [2]	2015	The construction and validation of Vader along with its evaluation is carried out. Generalizable Heuristics used by Humans to assess emotions are studied.	A gold-standard of the feature list is constructed. Five generalizable heuristics are defined which are used to assess sentiment intensity. VADER sentimental lexicon is compared with seven other lexicons and gives the highest Pearson Product Moment Correlation Coefficient of 0.88 which could increase to 0.96 at certain thresholds.
C. *Study of Depression Analysis using Machine Learning Techniques* [3]	2019	Several methodologies have been highlighted here including a novel one aiming at identifying depression stigma on online platforms.	Results show 6.09% of relevant posts on Weibo online platform reflect attitudes related to depression stigma.
D. *Hate Speech Detection from Code-mixed Hindi- English Tweets Using Deep Learning Models* [7]	2018	Three models based on CNN-1D, LSTM and BiLSTM have been employed for detection of hate speech.	CNN-1D model gave the maximum accuracy among the three models with 82.62%. LSTM model and BiLSTM models gave an accuracy of 80.21% and 81.48% respectively.
E. *Speech Emotion Recognition with Heterogeneous Feature Unification of Deep Neural Network* [5]	2019	A three-module architecture for speech emotion detection has been proposed using MFCC, neural networks and tested against several classifiers.	The architecture that has been proposed yields an accuracy of 64% with SVM chosen as the final classifier.
F. *Deep Learning based Emotion Recognition System Using Speech Features and Transcriptions*[6]	2019	CNN based approach has been applied to both speech as well as speech transcriptions. MFCC as well as Spectrogram speech features have been tested and analyzed. A combined model using both speech as well as text transcription has been proposed and analyzed.	Class accuracy of 69.5% has been recorded for both Text with MFCC model and Text with Spectrogram Model. Overall accuracy achieved of Text with MFCC model is 76.1% which is higher than Text with Spectrogram model which is 75.1%.
G. *Emotional Analysis of Bogus Statistics in Social Media*[8]	2020	Recurrent Neural Networks have been implemented to detect the underlying emotions in bogus statistics or fake information. This emotion analysis aims to help distinguish fake news from true ones and regulate social media posts.	The proposed RNN architecture results in an accuracy of 94.7%.
H. *Multi-Modal Emotion recognition on IEMOCAP Dataset using Deep Learning*[10]	2018	Emotion Analysis has been carried out on speech, text, facial expressions as well as hand movements. Convolution and LSTM based architectures have been tested out for building the individual and combines model.	For individual model and accuracy of 54% for text-based emotion detection, 64.78% for speech-based emotion detection and 51.11% for Motion Capture based emotion detection has been recorded. The combined model for Text, Speech and MoCap yields an accuracy of 71.04%.

IV. CONCLUSION

It can be seen from Table IX that this survey was carried out to consolidate the various approaches proposed by individuals in the field of Emotion Detection as well as certain benchmarks that are most widely used such as the IEMOCAP dataset. This dataset containing 12 hours of multimodal audiovisual data contains text, speech, video as well as motion capture.

Taking different types of data into perspective, text data provides important syntactical meaning and context to the keywords in question however it lacks certain low-level features which prevent it from being as robust as speech emotion detection. Speech features contain requires low-level features and when paired with 2D-CNN models can give an overall accuracy of 71% which is higher than 64% yielded by text[6]using CNN models. Combined architecture for text and speech were proposed that gave a 75%[6] accuracy which proves that the best of text and speech features can be combined for a more robust architecture. Similarly combined architectures using Motion Capture, Speech and Text yielded an accuracy of around 71% which was 17% higher than the accuracy obtained by just text[10].

RNN based model gave the highest result for text(94.7%) [8] whereas for speech 2D-CNN gave the highest at 71%[6]. SVM as a classifier worked best giving a high of 79% which was 2% high that Random Forest and 9% higher than KNN[5].

V. REFERENCES

[1] B. Dasgupta Poorna, Detection and Analysis of Human Emotions through Voice and Speech Pattern Processing, International Journal of Computer Trends and Technology (IJCTT)– Volume 52 Number 1 October 2017.

[2] C. Hutto, E. Gilbert "2.3 VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text (2015)" published in The International AAAI Conference on Web and Social Media - 2014

[3] D. Ramalingam, V. Sharma, P. Zar, "Study of Depression Analysis using Machine Learning Techniques " in International Journal of Innovative Technology and Exploring Engineering (IJITEE)ISSN: 2278-3075, Volume-8, Issue-7C2, May 2019.

[4] T. Pandurangan , " Emotion Analysis based on Real Time Human Voice Tones " on 2017/06/07

[5] Jiang ,Wang , Jin , X. Han and C. Li , " Speech Emotion Recognition with Heterogeneous Feature Unification of Deep Neural Network " in Sensors 2019, 19, 2730; doi:10.3390/s19122730

[6] S. Tripathi, A. Kumar, A. Ramesh, C. Singh, P. Yenigalla , " Deep Learning based Emotion Recognition System Using Speech Features and Transcriptions" accepted in CICLing:International Conference on Computer Linguistics and Intelligent Text Processing, 2019

[7] S. Kamble, A. Joshi, " Hate Speech Detection from Code-mixed Hindi-English Tweets Using Deep Learning Models" presented in the 15th International Conference on Natural Language Processing (ICON-2018)

[8] Haoxiang, Wang. "Emotional Analysis of Bogus Statistics in Social Media." Journal of Ubiquitous Computing and Communication Technologies (UCCT) 2, no. 03 (2020): 178-186.

[9] Mitra, Ayushi. "Sentiment Analysis Using Machine Learning Approaches (Lexicon based on movie review dataset)." Journal of Ubiquitous Computing and Communication Technologies (UCCT) 2, no. 03 (2020): 145-152.

[10] S. Tripathi, H. Beigi,S. Tripathi, " Multi-Modal Emotion recognition on IEMOCAP Dataset using Deep Learning" arXiv:1804.05788v3;2018

[11] Melissa N Stolar, Margaret Lech, Shannon J Stolar, Nicholas B Allen; "Detection of Adolescent Depression from Speech Using Optimised Spectral Roll-Off Parameters"; Biomedical Journal of Scientific & Technical Research; 2018.

[12] Thin Nguyen, Dinh Phung, Bo Dao, SvethaVenkatesh, Michael Berk; "Affective and Content Analysis of Online Depression Communities"; IEEE Transactions on Affective Computing; Volume 5; pp. 217-226; 2014.

[13] Lakomkin, E.; Weber, C.; Magg, S.; Wermter, S. Reusing Neural Speech Representations for Auditory Emotion Recognition. In Proceedings of the Eighth International Joint Conference on Natural Language Processing, Taipei, Taiwan, 27 November–1 December 2017; pp. 423–430. Computational Linguistics, Melbourne, Australia, 15–20 July 2018; Volume 2018, p. 2225.

[14] Gu, Y.; Yang, K.; Fu, S.; Chen, S.; Li, X.; Marsic, I. Multimodal Affective Analysis Using Hierarchical Attention Strategy with Word-Level Alignment. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Melbourne, Australia, 15–20 July 2018; Volume 2018, p. 2225.

[15] Satt, A., Rozenberg, S., Hoory, R.: Efficient Emotion Recognition from Speech Using Deep Learning on Spectrograms. In: INTERSPEECH, Stockholm (2017).

[16] Chernykh, Vladimir, Grigoriy Sterling, and Pavel Prihodko. "Emotion Recognition From Speech With Recurrent Neural Networks." arXiv preprint arXiv:1701.08071 (2017).

[17] Lee, J., Tashev, I.: High-level feature representation using recurrent neural network for speech emotion recognition. In: INTERSPEECH (2015).