ISSN: 2229-6956 (ONLINE)
DOI: 10.21917/ijsc.2019.0254

INVESTIGATION ON TEXT CONTENT ORIENTED EMOTION IDENTIFICATION

S.R. Adarsh
Senior Software Engineer, Samsung R&D Institute, Karnataka, India

Abstract

Human-Computer interaction (HCI) researches the use of computer technology mainly focused on the interfaces between human users and computers. Researchers in the domain of HCI observe the ways in which human users interact with computers and propose new technologies that let users interact with computers in narrative ways. Research in HCI is situated at the intersection of various fields like computer science, design, behavioral sciences, media studies and many more areas of study. In the world of digitization HCI is a very powerful and most relevant area of research and needs the digital systems to reproduce the human behavior appropriately. Emotion is one of the important aspects of human behavior and plays an important role in HCI. To exhibit accurately intelligent behavior, this needs to recognize the emotion of human behavior. There are various ways to express the Human emotions like facial expression, written content or texts and speech, nowadays enormous amount of textual data is generated and gathered into the cloud and social media and blogs are among others. This research paper investigates on the overview of emotion recognition from various texts and expresses the emotion detection methodologies. Boundaries of 'emotion detection methodologies' are investigated in this paper, those are addressing the text normalization process of useful data using various handling techniques for both plain text and short messages. This paper combines the two most common approaches to emotion classification in text: the rule based approach and the machine learning approach and compares the performance of various classification algorithms.

Keywords:

Human-Computer Interaction, Facial Expression, Plain Text and Language Processing Tools

1. INTRODUCTION

Emotions are a significant aspect for contact and communications between inhabitants. The exchange of emotions and feelings through various text messages, posts of personal blogs and SMS, Twitters, Facebook etc. are the informal style of writing and are a challenge for researchers. Meaningful extraction of various types of emotions from text can be applied for deciding the HCI which control and direct the communication. Other way of expression of Emotions by human are speech, facial and text based emotion.

Emotions expressed by Text messages are the essential information units of document. Strapparava et al. in [1] has explained the various approaches for classifying the emotions using an algorithm 'Customized Decision Tree Algorithm' it has been explained that this text level emotion detection method mainly depends on the emotion expressed by the human beings in personalized sentences, a manner of the document that consecutively relies on the expressed emotions by the personage words. It has been shown that Emotions could be expressed by speech, text and facial expression also. The research shows that worldwide the emotions are divided into six types which are (i) joy, (ii) love, (iii) anger, (iv) surprise, (v) sadness and (vi) fear. Limited amount of work has been conducted on text based emotion recognition techniques. The short messaging languages like twitter have the ability to interrupt and misrepresent the accepted language processing tasks performed on the text data.

Texts data will recognize some terms that do not belong to plain text. As per search query, following data of the text abbreviations are used by the large numbers of users in their chat messages. They are:

ILY means I Love You, ROFL means Rolling on Floor Laughing, LMK means Let Me Know, STFU means Shut the 'Freak' Up, SMH means Shaking My Head and YOLO means You Only Live Once, etc. Despite the fact that going through these abbreviations in the sentences, the human brain will resolve the short messaging language word to a meaningful word. When the word being referred to as "*cnt*" and its nearest words are "*he or she*" and "*immediately*", human know that it is "*can't*". That's because humans have experience with short messaging language terms in their brain's memory by earlier experience made long back. Similarly, the slang abbreviations are available in a large number, like Q.T. normally it is used in the expression on the Q.T., meaning "*behind closed doors*" or "*in confidence*" cutie - a homophone derived from the articulation of "*QT*", Quality time. Language Processing tools, they are educated and adopted to work properly with natural or plain text. Mapping the short messaging language words to plain text words can be very insightful in some cases. An erroneous mapping can result in alternations of the original meaning or it may wipe out semantics under the applied context. When considering the sub-phrase "*ur a marvel*" as an example, '*ur*' can be considered as '*your or you are*'. Human beings can understand that its 'you're a marvel" and not '*your a marvel*'. But it will be practically difficult to map them from a language tool. Thus it depends on the background of which the particular word is used.

This research paper deals with latest growing importance in the area of emotion recognition using text messages. The speedy growth of the cloud/internet has facilitated improved on-line communication, written contents and blog posts over the clouds/websites and opens the newer and faster possibilities to detect the emotions from those text messages. Mello and Graesser in [2] has observed in their research led to generation of large amounts of online content rich in the human user emotions, opinions and sentiments, although the above requires a detailed computational approaches to analyze this online content, distinguish, and draw useful conclusions and recognition of emotions successfully. Various approaches are available in the direction of recognizing the 'divergence of sentiment'.

The sentiment can be either positive or negative or both. It is required to explore in the area of 'sentiment recognition of types of emotions using the text documents'. In this paper the 'Recognizing emotions' part has been investigated in depth and same has been conveyed by a text which can provide an insight into the research's intent and that can lead to better understanding of the data in text's content. In computational linguistics, this is

becoming critical from a relevance point of view in the recognition of emotions from text data. The examples are affective computing, the tasks of market analysis and opinion mining or accepted language interfaces such as electronic learning environments or educational information and education based games.

There are large numbers of social network platforms such as Google+, Twitter and Facebook provide information for users. Micro-blogging platform is the popular among others in the world. It is also the fastest growing social network platform and has a dominant position in the area of micro-blogging [22]. Researchers also noticed that tweets often convey pertinent information about the user's emotional states [23]. Emotion analysis on Twitter has thus become an important research issue in the micro-blogging area [23]. Public platform based online data to perform emotion analyses significantly reduces the costs, efforts and time needed to administer large-scale public surveys and questionnaires. These data and results present great opportunities for social scientists. Khan et al. presented a paper in [25] which consists of new entity-level sentiment analysis method for Twitter. Lexicon based method perform entity-level sentiment analysis. The method gave highest precision, but low recall. To improve the recall, additional tweets that are identified automatically by exploiting the information in the result of the lexicon-based method has been applied. A classifier is then trained to assign polarities to the entities in the newly identified tweets [24] [26].

There are two approaches to this problem: the 'Rule Based Approach' (RBA) and the 'Machine Learning Approach' (MLA), the combination of them has been applied in this paper. The classifier of MLA is based on supervised machine learning algorithm, requires labeled data which is used to detect and classify the emotion of a tweet. Experimental result shows that MLA performance is better by around three percentages than RBA, the performance has been improved due to removal of the error data while training the model. The approaches are involved with the concepts of NLP, Artificial Intelligence, and Machine Learning for the development of the system. Our major contributions in this paper are detection of emotion for non-hash-tagged data and the labeled data creation for machine learning approach without manual creation. Architecture of Emotion Detection System, Workflow of Rule based Emotion Detection Approach, Training and Testing approaches and Workflow of Learning Based Emotion Detection methods have been explained elaborately with block diagram. In our rule based system we have classified the complete 1.2 million records; it gives the accuracy over 88.3%. Here we are using non-exhaustive cross-validation called k-fold cross validation technique to validate the result. Since the rule based classification includes a countable number of error data which reduces the performance, such tweets should be avoided. The output of RBA is fed to MLA. With the proposed system we are able to detect and classify the emotion of the tweet. This technology can measure public mood of people in a community, which may help social scientists to understand the quality of life of population. The research paper proposes an Emotion Detection System, a method for classifying Twitter text messages into distinct emotional classes they express. With the proposed system we are able to detect and classify the emotion of the tweet. The system performance has been increased in the machine learning approach when compared to the Rule based approach. This technology can measure public mood of people in a community, which can help social scientists to understand the quality of life of population.

This paper describes about the application areas, survey of different text based emotion recognition methodologies and their limitations and text normalization techniques for resolving the short messaging language.

2. LEARNING TO IDENTIFY EMOTIONS IN TEXTUAL FORM

The application areas of textual emotion detection have been explained by Carlo Strapparava and Rada Mihalcea in [3] in their research article 'Learning to Identify Emotions in Text', where they pointed out four applications of emotions, they are briefed as follows:

2.1 SPEECH GENERATION FROM THE TEXT

It has been stated that the main goal of text to speech creation is to classify the emotional similarity of sentences in the text, for identifying exact communicative representation of text- to-speech synthesis. The text documents are the collections of various sentences which mainly have emotional contents. In the form of verbal communication, readers can express the accurate emotions from that text content by modifying the manner of speech, including pitch of the sound and their intensity. In order to generate expressive speech from the text, it is important to identify the emotions from the text contents. Description of the emotions correctly in the text passage and how to convey the prosodic form in order to communicate the emotions from the given text credentials are the two applications of text to speech creation systems.

2.2 COMPUTER ASSISTED CREATIVITY

From the time of the journey of invention of computers, humans have developed the influence of computer systems in terms of their various working domains mainly increasing speed and minimizing the size with respect to time and aiming the capability to perform various tasks exponentially. Artificial Intelligence is having capability to push the technology towards creating computers/machines that might be one day as intelligent as human beings. Artificial Intelligence is ways of making a computer-controlled robot sense intelligently, like how intelligent humans think. The automated generation of evaluative terminology with a bias on certain divergence orientation is a key component in automatic custom-made advertisement and persuasive communication.

2.3 SENTIMENT ANALYSIS WITH TWITTER DATA

Sentiment Analysis with Twitter data mainly focuses on information retrieval and knowledge unearthing from the text. Using Artificial Intelligence, as a tool, the computer should be able to detect and express the emotions, which is the main target of the analysis of sentiments or opinion mining. Here the customer's opinion about products, services, improvement for finding their preferred choices in the business domain can be influenced effectively. One of the important convincing uses of

sentiment analysis these days are brand awareness, if we can comprehend what people are communicating about an individual in a natural context, one can make efforts towards addressing the key problems and improving the process of business transaction. We know that Algorithmia is an open marketplace for algorithms, facilitating developers to create future's smart applications today. The Algorithmia marketplace constructs it easy to take out the content one need from Twitter and channelize it into the right algorithms for the sentiment analysis. There are limited algorithms on this platform for exploring different information from Twitter and a number for the sentiment analysis.

The Work Flow:

Step 1: Draw together important tweets from Twitter.

Step 2: Preprocessing without the word elimination.

Step 3: Apply the right sentiment analysis algorithm.

Step 4: Analyze the results.

Step 5: Thrash out further improvement and next steps

Defining the keywords, importing the Algorithmia library in python, and calling the right function by collecting the tweets required from Twitter:

L1: #Import the packages we'll need, L2: import pandas as pd,

L3: import Algorithmia, L4: #Define the two companies who's sentiment we want to compare,

Step 1: keyword1 = "tesla",

Step 2: keyword2 = "comcast",

Selecting the sentiment algorithm depends on various factors they are level of detail, rapidity, cost, accuracy etc. The application for the executives of the company is to focus on increasing positive sentiment regarding to the brand image on social platform. Opinion is collected from web forums, discussion groups, blogs, comment boxes, online e-learning systems. Opinion Mining is a very important application of web data. It is used to collect user's opinion and extract the meaningful patterns from it. Both the positive and negative annotation is an active area in the sentiment analysis, during the process of decision formation; it may take best decision on the basis of opinion of others data extracted from the twitter. The emotion footnote may increase the effectiveness of their applications.

2.4 IMPROVES HUMAN COMPUTER INTERACTION

Human Computer Interactions (HCI) researches the design and makes use of computer technology, mainly focused on the interfaces between human and computers. The emotion recognition system should be applied in different kinds of the HCI systems, such as automatic answering systems, discussion systems, and human robots etc. A system that based on the user's emotion makes the HCI coordinated.

3. TEXT BASED EMOTION RECOGNITION METHODS

Significant research in affective computing has been done to detect facial, acoustic and gestural emotions. They would stand as important indications for advanced HCIs. Emotion detection from texts is gaining more attention these days. However, text-based participation is still the most common way for humans to relate with computers, and thus emotion recognition from texts should be refocused as an imperative research concern in affective computing.

In this paper, we have surveyed accessible research of emotion detection and investigated the limitations to improve recognition capabilities, describing a suggestion of incorporated system architecture. These enhancements include identification of recently-evolved vocabularies, organized emotion ontology based on OCC model developed by Ortony et al. in [4] have explained about the methodology and detected various emotions in the form of case-based reasoning. The above model distinguishes more than 20 emotion types differentiating by the psychological situations. Emotions involving a central point on events from those focused on events and those focused on objects.

The identified emotions would stand as important indications for advanced HCI. There are four text based emotion recognition procedures: Keyword spotting method, which is having three subdivisions based on document, line and words. The Lexical Affinity Method is based on vocabulary of a person or language. The learning based approach is based on SVM. In machine learning, the support vector machines (SVMs) are supervised learning models with connected learning algorithms that investigate data used for the classification and regression analysis. The last emotion recognition method is the hybrid methods. Binal and Wu in [5] has elaborated the various Computational Approaches for the detection of Emotion in Text format.

3.1 KEY WORD SPOTTING (KWS) METHODOLOGY

Performing method by KWS on textual databases is comparatively simple. The text data are perused for a given directory of words and the location wise position of the words which is tagged contained by the text. Translating this method for bringing into play in speech databases is a 2-stage process. First, a state-of-the-art LVCSR (Large Vocabulary Continuous Speech Recognition) Engine is engaged in to transforming the entire speech signal into text. The LVCSR engine performs the exploration for the most probable string of words based on algorithm known as Viterbi search algorithm. This employs mainly three components (i) language model, (ii) acoustic models, and (iii) a large lexicon of words. In the next stage, the KWS method utilizes well recognized text-based search techniques to locate the keywords contained by the text. The indexing phase can be incorporated on the consequential text in order to increase speed of the search response time.

This approach is straightforward to implement and instinctive because it involves identifying words to explore for in text as investigated by Liu et al. in [6]. The keyword pattern matching difficulty can be described as the predicament of finding occurrences of keywords from a known set as substrings in a specified string. These words are classified into various categories such as anger, fear, disgust, sadness, happy, surprise and others. The above observations have been supported by Wen et al. in [7].

The KWS has well defined sequences from the Text documentation to tokenization of text data. The next step is emotion keyword detection followed by the analysis of the intensity. Negations are being checked and finally emotion class is identified. The above process can be represented in Fig.1.

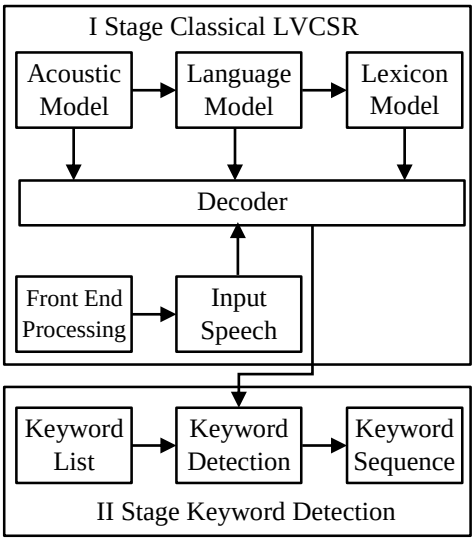

Fig.1. LVSCR keyword spotting system

3.1.1 Acoustic KWS:

The 'Acoustic KWS' is accepted method of the Key Way System. In the Acoustic KWS method the engine does not attempt to write down the entire stream of the speech. It employs the Viterbi search engine, employs a speech identification engine on the speech, which is intended to cover up all potentially verbal words. A smaller set of selected keywords is used as the identified vocabulary as stated by Thambiratnam in [8] and Moyal et al. [9]. They developed a general model as part of the acoustical models known as Phonetic Search Model for Large Speech Databases as given below in the Fig.2.

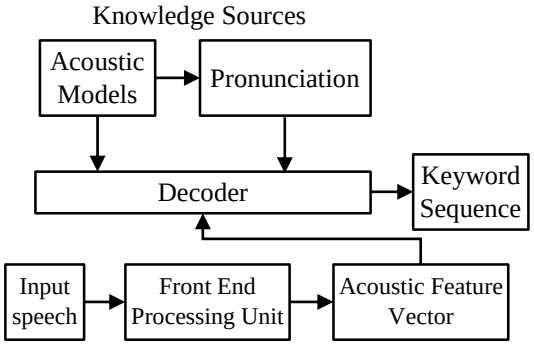

Fig.2. Acoustic keyword spotting system

3.1.2 Phonetic Search KWS:

Phonetic search KWS make use of a phonetic search engine. There are two stages of this system. The first stage employs phoneme decoder. It transforms the speech input into a sequence of textual form. Here the decoder transforms the speech signal into a lattice of phonemes as investigated by Garofolo in [10] and Beth Logan et al. in [11].

Next stage, phonetic search engine employs a distance calculation to compute the textual distance between the phoneme sequences. This sequence corresponds to the keyword vocabulary. The Fig.3 shows the block diagram of the phonetic search engine. It uses 2 types of input data: a list of keywords, here individual word is represented by a sequence of phonemes, and the phonetic data is processed through a phoneme decoder to produce a string of recognized phonemes.

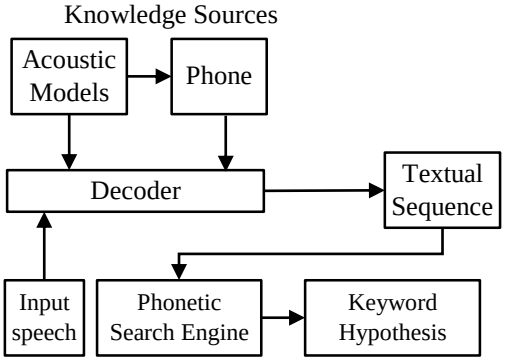

Fig.3. A phonetic search KWS system

Frinken in [12] suggested the Keyword spotting method in real time chat system. In this system three Emotion model has been presented, they are: (i) Social Information Processing Model (Happiness and Sadness), (ii) Five Factor Model (FFM) developed by Ortony, Clore and Collins (OCC) and Ekman Model.

The main feature of the first model has been considered as Linguistic Enquiry and Word count Program (LIWC) having granularity. The greater the granularity will provide deeper level of detailed input data. Granularity is usually used to characterize the level of detail in a set of data. The task description of laboratory controlled online chat enacting both the sadness and happiness and communicating with the strategies that human being can employ to express the emotion in textual form.

The second model of FFM is having the task description based on emotion detection incorporating the personality factor in chatting system to enhance the result quality. The features considered here is Open Mind Common Sense (OMCS) knowledge based concept with granularity in the form of sentence. The last model uses the database word-net. Emotional weight of the OMCS knowledge base is having the task description of emotion detection in chat system and displaying emotion using a sentence as granularity. The keyword spotting technique use diverse methods like Word-based Keyword Spotting (WKS), Line-based Keyword Spotting (LKS) and Document-based Keyword Spotting (DKS).

3.2 FRAMEWORK OF LEXICAL AFFINITY MODELS

Lexical Affinity method is an addition of keyword spotting method. Lexical Affinity method gives most probabilistic similarity for a particular emotion to subjective words rather than detecting predefined emotional keywords from the text. The possibilities which are assigned by this method are element of linguistic corpora. There are some limitations like pre-notion toward specific variety of texts and does not recognize the emotions from the text that does not exists. Consider an example, "I met my enemy by mistake". In this observations the word 'by mistake' is reflects the good probability of a pessimistic emotion. The exact condition in this sentence is that 'by chance' word is not giving the negative emotional consideration.

A framework for the effective and fast computation of lexical affinity models has been applied in the framework of Latent Semantic Analysis by Landauer and Dumais in [13], where about 65% of the questions were responded correctly. Two researcher Terra and Clarke in [14] used different measures and arithmetical estimates to answer the various questions, achieving more than 73% and 81% correct responses respectively. In another similar research Jarmasz and Szpakowicz in [15] used a thesaurus to work out the distance between the alternatives and the target word, answering around 79% correctly. Terra and Clarke in [14] trained a system to answer the various questions with an approach derived from combined components, including a module for LSA, thesaurus and a number of heuristics based on the patterns of synonyms. The combined approach has answered around 98% of the questions accurately after being trained over 350 examples as an alternative of using arbitrary windows to compute word similarity measures, we model lexical affinity by applying the total observed distance distribution along with parametric and independence models for this allocation.

Results indicates that, with minimal endeavor to adapt the models, it is possible to achieve good results by applying the simple natural language tasks, such as GRE fill-in-the-blanks tests and TOEFL synonym questions in this framework. This framework permits the use of terabyte scaled corpora by providing an algorithm to extract couples of co-occurrence for the models, thus enabling the use of precise estimators.

3.3 LEXICON-BASED APPROACH

Shiyang et al. in [7] has elaborated about Lexicon-based approaches based on an emotion lexicon. They constructed the emotion lexicon from three resources:

i. Utilize the emotion lexicon
ii. Collect and use slang words
iii. Gather a list of emoticons using the micro-blog website to enhance the lexicon.

This approach applies the segmentation tool to segment the micro-blog text into words. Based on the developed emotion lexicon, it calculates the number of emotion words taking place in a text for each and every emotion type, and then the emotion tag of the text is determined as the emotion type with number of emotion oriented words appearing in the text. The text is labeled as “none” when it does not contain any emotion words. This process also applied on the sentences to get the ‘sentence-level emotion label’.

3.4 LEARNING BASED METHOD

Learning based methods undertake to recognize emotions based on a preceding trained results, which mapped with various machine knowledge based classifiers such as support specific statistic learning methods, vector machines, and judgment trees to detect which emotion class should the input data of the text belong. There are difficulties in this approach. These methods classify the text sentences into only two categories. The reason behind it is the inadequate features excluding emotion keywords, which can be either the negative or positive. Dung and Cao investigated in [16] about the application of the idea that emotions are co-related to human mental states that are caused by a number of emotional measures. They explained that the individual mind starts with preliminary mental state and moves to a different state upon the happening of a certain event. The same thought had been implemented with Markov’s Model in which each text sentence consists of numerous sub-ideas and every one of them is treated as an experience that causes an alteration to a certain state. This change of sequence of events in the sentence is followed by the method which determines the emotion of the text. This system could achieve the F-score of about 35% when tested on the International Survey on Emotion Antecedents and Reactions dataset as pointed out by Klaus et al. in [17], where the precision achieved was more than 12% with respect to above method. In another method developed by Wen et al. in [7], known as SVM-Based Approach, where the learning-based approaches uses the LIBSVM toolkit. Three kinds of text-based features are employed at both the document-level and the sentence-level emotion classification.

i. *Emotion Lexicon appearance*: This takes a large number of texts in word form of each emotion type occurring in a sentence as feature.
ii. *Word characteristics*: All the words appearing in this approach as a micro-blog text or sentence are used as feature.

3.5 FUSION BASED APPROACH

This approach is based on an arrangement of two methods, the keyword based method and the learning based method. The major advantages of this method are that it can relinquish higher accuracy results from preparation and adding knowledge-rich linguistic information from the thesauri and dictionaries. Another advantage of this method is to balance the higher cost involved for the retrieval task as well as minimization of the information tasks difficulties. Yang et al. in [18] developed a hybrid model for the emotion categorization that includes (i) CRF based emotion cue identification, (ii) Lexicon-keyword spotting, and (iii) Machine-learning-based emotion classification applying Max Entropy, SVM, and Naïve Bayesian. The results generated from the above three methods are integrated by means of a vote-based system. It is a dataset of text notes where it gains an F-score of about 60% with precision. This technique achieved comparatively good results.

The major approaches for text based emotion detection have been presented here and it has been shown that how the syntactic and semantic information can be advantageous for emotion detection. But current methods are lacking in deep semantic analysis for detecting secreted phrase patterns and supplementary investigations are needed to be identified, built and incorporated knowledge rich linguistic resources that have a focus on detecting emotions.

4. SCOPE FOR FURTHER DEVELOPMENT

Four limitations have been pointed out of the above methods that are as follows:

4.1 INDISTINCTNESS IN KEYWORD DEFINITIONS

Using emotion keywords is a simple and clear-cut way to detect connected emotions, the meanings of keywords could be

numerous and vague, as the majority of words could change their meanings according to dissimilar usages and contexts. Even the minimum set of emotion labels could have unlike emotions in some extreme cases such as cynical sentences.

4.2 INABILITY OF RECOGNIZING SENTENCES, LACKING EMOTIONAL KEYWORDS

Keyword-based move is totally based on the set of emotion. Therefore, sentences without any emotional keyword would mean that they do not hold any emotion at all, which is obviously erroneous.

For example, '*I have cleared the GRE exam*' and '*Hooray! I passed the GRE exam*'.

Should denote the same emotion (happiness), but the former without "*hooray*" could remain hidden if '*hooray*' is the only keyword to become aware of this emotion. The syntax structures and semantics have enough influences on the expressed emotions, which is shown in the example as, '*I laughed at her*' and "*She laughed at me*'. In this example the different emotions from the first person's perspective is more significant, ignoring the linguistic information. This emotion poses a problem to keyword-based methods.

4.3 DIFFICULTIES IN FORMATIVE EMOTION INDICATORS

Learning-based techniques can automatically determine the possibilities between features and emotions and these are still requiring keywords, but it is in the form of features. The most intuitive features could be emotive view which can be seen as marginal notes within the texts. The cascading problems would be still those that are persistent in keyword-based approaches.

5. TEXT NORMALIZATION TECHNIQUES AND SHORT MESSAGING LANGUAGE

There are various techniques to normalize the text messages. The major approaches in this regard are being represented by (i) Spell Checker technique, and (ii) Encoder Decoder technique. The above techniques are explained as follows:

5.1 SPELL CHECKER APPROACH

The spell checker approach focuses on determining the short messaging language words to simple text words not only by looking at the characters, but considering contexts of those words too. This approach has been elaborated by Damerau in [19]. The researcher used to edit the distance to find out words confusions in this approach, for checking the suitability of a particular word for replacing it, context based spell checker approach has been applied. Lowest amount of editing the distance is used in identifying the difference between two words in this technique. Finally, in the process of transforming one string to another string, least number of edit operations is requisite in this method.

5.2 ENCODER-DECODER TECHNIQUE

The summarization of the text is a challenge in the field of natural language processing for generating an accurate, short, and smooth summary of a source document. The Encoder-Decoder recurrent neural network architecture which is developed for machine translation has proven itself effective when applied to the problem of text summarization. It can be tricky to apply this architecture in the 'deep learning library'; here some of the flexibility is sacrificed to make the library simple, clean, and easy to be used. The Encoder-Decoder architecture is a method of organizing recurrent neural networks (RNN) for sequence prediction problems. The RNNs have a variable number of either the inputs or the outputs; it can be both inputs and outputs. The encoder reads the total input sequence and encodes them into an internal representation, normally a fixed-length vector named the context vector. The decoder studies the encoded input sequence from the encoder and creates the output sequence. The decoder is required to generate each word of the output sequence and provide to the two sources of information.

Context Vector: This is the encoded illustration of the source document provided by the encoder. It may be a fixed-length encoding as in the simple architecture. The generated sequence is provided with modest groundwork, like distributed representation of each created word through a word embedding process.

Generated Sequence: The sequence of words previously generated as a summary.

This model aims to capture the notion of two levels of significance using two bi-directional RNNs on the source input side, first one at the word level and another one at the sentence level. The operations of the attention mechanism manage at both levels simultaneously.

Basically Insertion, Deletion and Substitution are the edit operations. During the process of the determining the value the above approach, the representation for each edit operation is used. In spell checking application, the least amount of edit distance of short messaging language words is calculated and using those values, the correct word of short messaging language word is predicted as the word with lowest amount of edit distance value as stated in the research being developed by Norvig in [20]. In this approach the minimum edit distance has been applied to determine the short messaging language words which is not accurate that's why it cannot be acceptable. With an example we can understand the Text contains plain, text words and short messaging language words. We can consider a word 'good morning', the most frequently used small messaging language word of it is 'gdmrng', but there are versions of word 'gmrng' such as 'gmng', 'gdmrng'. We can predict the most possible correct word of Good Morning, using minimum edit distance.

In this technique, an experiment has been performed to find the accuracy. Peter Norvig's simple spell correction algorithm has been applied in [20] and in the book of Jurafsky and Martin in [21]. The algorithm is customized for testing the accuracy of this. Different derivations of short messaging language terms were tagged with their short messaging language words. Subsequently the words were processed to the spell checker and evaluated the results with more than 60 % accuracy.

6. PROPOSED SYSTEM

This paper combines the rule based approach along with the machine learning approach to get better results. Before we propose the final system, its vital to understand the difference between the rule based emotion classification and the machine

learning based emotion classification. The constituents of the sentence play an important role in a rule based system where any sentence is classified based on the maximum occurrence of words from a particular class. In contrast, the machine learning approach focuses on both the syntax and semantics making it better than the rule based approach. This makes the machine learning models perform better the rule based approach in case of unseen data.

The research paper deals with the proposal related to the machine learning approach and knowledge based approach for detecting the emotion/mood of the text based tweet.

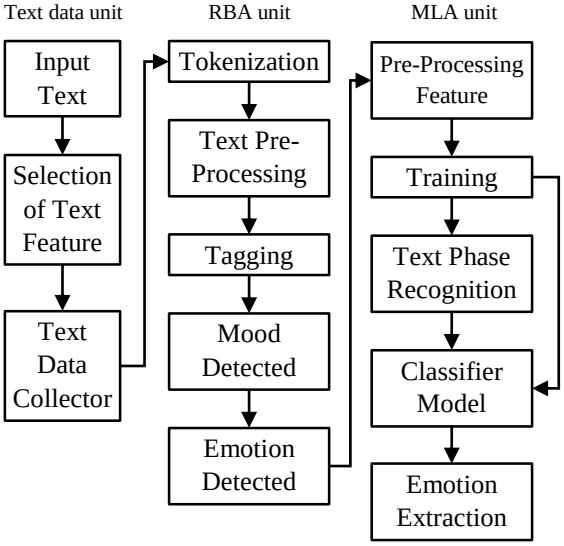

Fig.4. Emotion Detection System Architecture

The Fig.4 describes the architecture of the emotion detection system which includes text data collection, input text data, tagging, text feature selection unit, tokenization unit, pre-processing, knowledge base preparation, knowledge validation and classifier model. The system is based on the two basic approaches:

- Rule Based Approach and
- Machine Learning Approach

6.1 DATA COLLECTION PROCESS

"The row text data is obtained from WordNet Statistics and sentiwordnet databases [27], [28]. Total 12,48676 tweets were collected. The dataset is having 6 attributes like polarity, identification, date, query, username and tweet. Here last attribute, the tweet, has been focused. In the dataset their purpose is to identify the level of emotion.

6.1.1 Distinctiveness of Tweets:

- *Length*: The highest length of a twitter message is having 140 characters, but average length of a tweet is 16 words or 81 characters. This is dissimilar from the previous sentiment classification research work that was mainly focused on classifying the longer bodies of work.
- *Language model*: Twitter messages are posted from many different media, mainly including their cell phones. The frequency of slang and misspellings in tweets is higher than in other domains.
- *Domain*: Twitter posts short messages are having a variety of topics. This is different from the past research work done in this area.

6.2 RULE BASED APPROACH

RBA systems are used to store and maneuver knowledge to interpret information in a constructive way. It is based on the rules generated for the problem solving methods. A rule is a set of specific conditions and its actions. The rules are being represented by knowledge. A rule is in form of '*Condition and Effect*'.

Below are the different tasks performed in Rule Based Approach:

6.2.1 Pre-processing and Tokenization:

Tokenization: The tweets are stored in a file with six attributes stated earlier separated by comma; the sixth attribute is the tweet. We have used NLTK tokenizer package [33]. It uses Punkt Sentence Tokenizer (PST); which divides the text into a list of words and sentences for dividing a sentence into words. Straightforward PST divides the strings into sub-strings using a particular delimiter string split method directly, because of its effectiveness. The easy tokenizer is not available as separate function; instead, it can use the split method directly. Using split method with space as delimiter, the text can be expressed into tokens. Finally, the whole sentence is made into tokens.

Example:

Input: "@Friends I was too much excited to see you!

Output: „@Friends, „I, „was, too, much, excited, to, see, you!.

Pre-processing: Analyzing the text data that has not been cautiously screened may give the wrong impression about the results. If there is much unrelated information the unreliable data reduces the performance. The seven processes are given as follows:

1. Tweets contain usernames which start with the @ symbol before the username
2. Numerous tweets can contain url link. These links are removed by using regular expression".
3. Some tweets may contain Arabic numbers; integers are removed using the regular expression to deal with text data.
4. The tweets are converted to lower case.
5. Tweets can contain special characters; those can be stripped of using strip function.
6. Several tweets may contain short forms; those short forms tweets are replaced with full forms using a predefined list.
7. The text is standardized, from the shortcuts. The necessary texts of the social media are standardized into full form.

6.2.2 Tagging:

Part of Speech (POS) tagging plays major role in feature extraction process as POS goes in semi semantic way. The matching word may have different meanings depending on its convention. Here the POS tag is applied for the word based on its previous and following word. Tagging is based on context. Here each word is tagged properly.

6.2.3 Feature Selection:

The features are extracted using Term Frequency - Inverse Data Frequency (TFIDF) method, but this method goes with syntax whereas POS tagging method goes with the semantics. Each and every tags are not the emotional one, to know the emotion carrying words, the Russell's Circumflex Model (RCM) is applied. The Circumflex modal proposes that emotions are distributed in a two-dimensional spherical space, containing pleasure and activation dimensions [32]. It has been observed that the emotion carrying words are mostly nouns, adjectives. In the process of characteristic selection, we have considered noun phrases, adjective phrases, few verbs, adjectives, and adverbs. All the specified tags may not be useful, to know the mainly specific emotional words, we have used the more frequently repeated words in the dataset as features. The feature space does no longer take account of all the words, but instead it only contains the emotional words from the Knowledge base [202]. This method reduces the dimension of feature space noticeably, without losing informative terms.

6.2.4 Negation:

The words with preceded by negative feeling word can be called as negated features like “unhappy” or “cannot walk” etc. The negation features have the capability to completely show the opposite side of a tweet. It changes the semantic of the sentence, so these negated features are very important.

The RBA is used to classify the tweets under four categories. This applies certain rules to get a winding up. Shaheen et al. provided the seed list in [29] which consists of different emotional words those are nothing but extracted frequent features. The tweet is interpreted line by line along with its tag. A regulation is written in the form of ‘if’ and ‘then’. Here the tweet is read with its tag in order to keep away from the computational cost, if all the words are made to read then the time required to read 1.2 million text words will be more. So to avoid the unnecessary complex computations the specified tags like noun, verbs, adverbs, adjectives and combinations of them carrying the emotions are taken into account. Then the extracted word is confirmed against the seed list. If the word is there in the seed list, then the word is stockpiled on to a dictionary as suggested by Povoda et al. in [31]. Same process is repeated for each tweet. The emotional words in the tweet will be considered in this category. Finally, if a large number of words in a tweet come under any one of the category then that class label is allocated to the tweet.

Here four different classes, are considered they are:

- C1 – Happy-Active Class;
- C2 – Happy-Inactive Class;
- C3 – Unhappy-Active Class;
- C4 – Unhappy-Inactive Class.

C1: The happy-active class specifies additional happy emotions like so-happy, over-excited, overconfident, overjoyed, etc.

C2: The happy-inactive class specifies comparatively less happy emotions like calm, silent, influenced, pleased, etc.

C3: The unhappy-active class specifies a lesser amount of sad emotions like apprehensive, nervousness, fear, angry, furious, etc.

C4: The unhappy-inactive specifies supplementary sad emotions like miserable, very depressing, dissatisfied etc.

The Fig.5 provides the Workflow of Rule based Emotion Detection approach. The Fig.6 provides the block diagram of supervised training process of data collection.

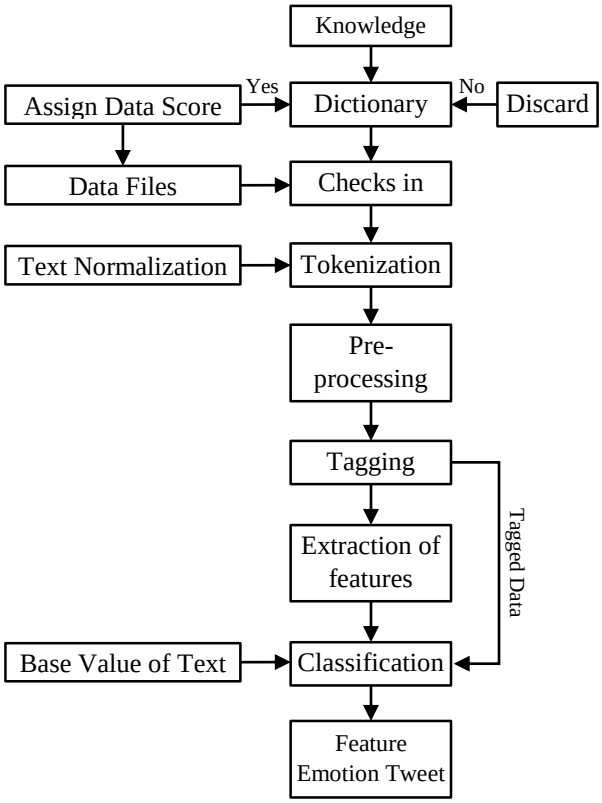

Fig.5. Workflow of Rule Based Emotion Detection Approach

6.3 MACHINE LEARNING APPROACH

A number of statistical classification methods are available that have been applied to text tagging. Bayesian classifier (Naive Bayes classification) is applied in this research paper. It consists of two phases: Phase I: Training Phase and Phase II: The testing Phase. Bayesian classifiers model is based on the word features in different classes, texts are classified based on subsequent probabilities generated based on the existence of different classes of words in text. Naive Bayes has been extensively used for classifying text because it is plain and efficient functions which are based on Bayes rule. The aim of this method is to select the class with a maximum probability.

6.3.1 Training Data Creation:

The required training data is being already created through rule based classification and this approach has been applied by Srinivasu et al. in [23]. The text data free from the error tweets are fed to the training algorithm.

6.3.2 Training Phase:

The output of the statistical classification methods stated above is used as input for the next phase (the training phase). It is used so as to minimize the time and cost of the whole process. The training of text data can be created physically but for that we need an expertise which may be expensive. Physical creation of training data consumes more time. Rule based method is simple to create when compared with physical creation of training of text data. Training the text data is created in a semi-automatic way

applying the rule based system as far as this approach is concerned.

In the training phase the text tweets are supplied with a class label. Each and every tweet is associated with a class label. In this paper the work we have applied are of four class labels, so each tweet will fall under one of the specified class categories. The text should be converted to Arabic numeric, since the MLA works with numeric data. The Naive Bayes conditional assumption is:

$$P\left(\frac{doc}{C_j}\right) = \prod_{i=1}^{len(doc)} P\left(a_i = \frac{W_d}{C_j}\right) \quad (1)$$

where, C_j = class, W_d = Word.

The above assumption pronounces that the probability of the tweet of assured class is going through the length of the text tweet; the probability of every one word is classified as Cj or class. The values are the number of times a text word occurred in the given tweet follows the following formulation in the Eq.(1) and Eq.(2).

$$P(W_d/Class) = (N_v + 1)/(n + \text{Vocabulary}) \quad (2)$$

where, n = total no. of words with specified class and N_v = no. of times word occurred with the specified class Vocabulary equals size.

Procedural Steps:

Step 1: Calculate N_v i.e., no. of times the text word occurred with class

Step 2: Calculate n i.e., total no. of text words for given class

Step 3: Calculate $p(w_d/C_j) = N_{v/n}$ i.e. the probability of word for the class provided.

Step 4: Calculate the probability of class

Step 5: Calculate the total no. of unique words (vocabulary)

Step 6: Repeat the whole process.

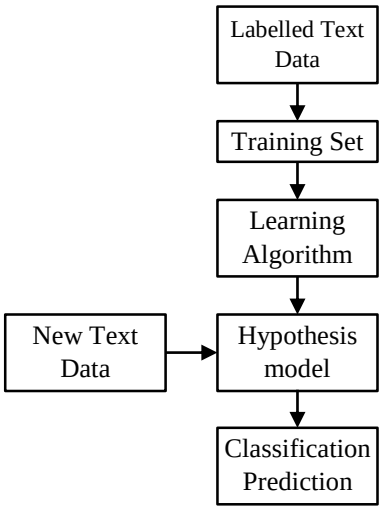

Fig.6. Block Diagram of Supervised Training Process

6.3.3 Testing Phase:

In the testing phases, the text tweet without any class label is provided; it means that the tweet is unnoticed before. The Bayesian algorithm in the training phase, the machine is categorizing the text tweet into one of the class categories provided. The input in the testing phase is a new tweet and the output is category label. The naive Bayes Classifier combines them with a decision rule, which picks up the hypothesis with greatest probability. In simple words, we pick the class which has highest value.

This can be formalized in following equation:

$$V_{bn} = \text{argmax } P(C_j) \Pi P(W/C_j) \quad (3)$$

where, $C_j \in V$, $W \in$ words, V_j represents the class from all V classes and W represents the word.

The (C_j) is the prior probability of the class V_j and $(W|C_j)$ is the probability of word and W for the given class C_j.

Procedure Steps:

Step 1: Read the text file line by line.

Step 2: Read each word of the tweet

Step 3: Calculate the probability of each tweeted word in the tweet against each and every class

Step 4: Check for the highest probability for each tweet word against all the classes, the word with highest probability class is considered.

Step 5: Repeat the process till the all the words are processed.

Step 6: Check the class which is having highest number of tweet words in the given tweet.

Step 7: Repeat the process till the all the words are processed.

The Fig.7 shows the block diagram of testing of the model using the text data from the tweets or features.

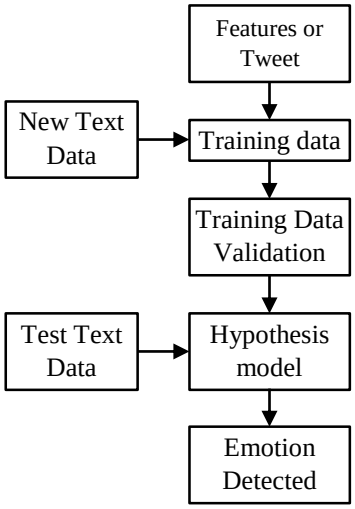

Fig.7. Block Diagram of Testing the Model

7. EXPERIMENTAL RESULTS

The system consists of four classes of emotions as follows:

a) (C1) Happy-Active,

b) (C2) Happy-Inactive,

c) (C3) Unhappy-Active and,

d) (C4) Unhappy-Inactive.

The "Russel's Circumplex Model" has been used here to know the emotion carrying words. Russel's Circumplex Model [33] suggests that the emotions are distributed in a 2-dimensional circular space, containing activation and pleasure and dimensions. The activation dimension measures the mood state. The pleasure dimension measures the feelings of positive or negative state of a person. In our model we have classified the complete 1.2 million

text records, giving the accuracy over 88%. Non-exhaustive cross-validation technique are applied to validate the result. Since the rule based classification includes a fixed countable number of error data which reduces the system's performance, such text tweets should be evaded. Finally, the output of RBA is fed to MLA.

7.1 EVALUATION OF THE PERFORMANCE

For the MLA's operation we are using random number generation tweets. We have generated 9000 tweets randomly, from this 82% tweets are used for training and 18% are used for testing processes. The Table.1 provides the machine learning Naive Bayes Tweet Count for the machine learning algorithm. Test phase tweets have been distributed in the way as shown below to assisting the confusion matrix.

Table.1. Machine learning Naïve Bayes Tweet Count

No. of tweets for training	No. of tweets for testing
7380	1620

Table.2. Confusion Matrix for MLA

***N* = 1600**	**C1 (Happy-Active)**	**C2 (Happy-Inactive)**	**C3 (Unhappy-Active)**	**C4 (Unhappy-Inactive)**
C1	351	34	101	1
C2	33	396	16	2
C3	6	4	348	5
C4	1	2	6	314

Emotion detection system is then measured based on F-measure score recall and Precision obtained as follows:

7.1.1 Precision:

Precision is referred to as positive predictive value.

$$\text{Precision} = TP/(TP + FP) \quad (4)$$

where, *TP* and *FP* are the number of true +ve and false +ve predictions for the considered class.

7.1.2 Recall:

Recall is referred to as sensitivity, corresponds to the true +ve rate of the considered class.

$$\text{Recall} = TP/(TP+FN) \quad (5)$$

The Eq.(4) and Eq.(5) are shown in the Table.3 and Table.4.

Table.3. Precision and Recall for MLA

Category	Precision	Recall
C1	**0.73**	**0.88**
C2	**0.87**	**0.89**
C3	**0.96**	**0.94**
C4	**0.97**	**0.95**

7.1.3 F-Measure:

The F-score is a measure of the accuracy of the tests. F1 score is interpreted as a weighted average of the precision and the recall, where F1 score approaches its best value at One and worst at Zero. The F-measure of the system is 0.883.

Table.4. F-measure for MLA Validation

Category	F-measure
C1	0.79
C2	0.92
C3	0.91
C4	0.94

The Naive Bayes is validated by applying the cross validation technique. To check the performance against other MLA's, training is done for SVM, KNN and Decision trees. The validation results are given in the below table. Randomly generated text tweets are fed to the model and checked for its category validation.

Table.5. Accuracy of RBA and various MLA's

Approach	Accuracy
RBA	85.2%
MLA (Naïve Bayes)	88.3%
MLA (SVM)	88.8%
MLA (Decision Trees)	89.5%
MLA (KNN)	90%

The Table.5 provides the precision of the earlier approach and MLA. The rule based system accuracy is over 85% and the accuracy of maximum of the MLA's is 90%. Compared with the earlier approach the accuracy of MLA is more because some error data has been removed while training the MLA algorithm.

8. CONCLUSIONS

We have proposed an Emotion Detection System in this paper. This is a method for classifying Twitter messages into divergent emotional classes they convey. The system is able to detect and classify the emotion of the text tweets. The system performance has been improved up to 90% in the machine learning approach (MLA) when compared to the first approach that is Rule based approach (RBA). The proposed method can be used by the professionals/counseling agencies to monitor and track to recognize anxiety or systemic stress of population. This technology can measure public mood of people in the community, which may help social scientists to understand the better quality of life of population.

In this research paper, the investigations on existing systems reveal that a combination of RBA and MLA performs much better than an only rule based solution. This paper also shows that KNN along with a rule based approach performs the best when compared to Naïve Bayes, SVM and Decision trees.

REFERENCES

[1] Carlo Strapparava and Rada Mihalcea, "Annotating and Identifying Emotions in Text", *Intelligent Information Access*, Vol. 301, pp. 21-38, 2010.

[2] K.D. Mello Sidney and Art Graesser, "Language and Discourse Are Powerful Signals of Student Emotions during Tutoring", *IEEE Transactions on Learning Technologies*, Vol. 5, No. 4, pp. 304-317, 2012.

[3] Carlo Strapparava and Rada Mihalcea, "Learning to Identify Emotions in Text", *Proceedings of ACM Symposium on Applied Computing*, pp. 1556-1560, 2008.

[4] P.N. Johnson-Laird and K. Oatley, "The Language of Emotions: An Analysis of a Semantic Field", *Cognition and Emotion*, Vol. 3, No. 2, pp. 125-137, 1989.

[5] Binali Haji, Chen Wu and Vidyasagar Potdar, "Computational Approaches for Emotion Detection in Text", *Proceedings of 4th IEEE International Conference on Digital Ecosystems and Technologies*, pp. 13-16, 2010.

[6] Chun-Chieh Liu, Ting-Hao Yang, Chang-Tai Hsieh and Von-Wun Soo, "Towards Text-based Emotion Detection: A Survey and Possible Improvements", *Proceedings of International Conference on Information Management and Engineering*, pp. 611-617, 2009.

[7] Shiyang Wen and Xiaojun Wan, "Emotion Classification in Microblog Texts using Class Sequential Rules", *Proceedings of 28th International Conference on Artificial Intelligence*, pp. 23-29, 2014.

[8] K. Thambiratnam and S. Sridharan, "Dynamic match Phonelattice Searches for Very Fast and Accurate Unrestricted Vocabulary Keyword Spotting", *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, pp. 12-18, 2005.

[9] A. Moyal et al., "*Phonetic Search Methods for Large Speech Databases*", Springer, 2013.

[10] J. Garofolo, "TREC-9 Spoken Document Retrieval Track", Available at: https://trec.nist.gov/pubs/trec9/papers/cuhtk_trec9.pdf.

[11] Beth Logan et al., "An Experimental Study of an Audio Indexing System for the Web", *Proceedings of 6th International Conference on Spoken Language Processing*, pp. 16-20, 2000

[12] V. Frinken et al., "A Novel Word Spotting Method Based on Recurrent Neural Networks", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 34, No. 2, pp. 211-224, 2012.

[13] T.K. Landauer and S.T. Dumais, "A Solution to Plato's Problem: The Latent Semantic Analysis Theory of the Acquisition, Induction, and Representation of Knowledge", *Psychological Review*, Vol. 104, No. 2, pp. 211-240, 1997.

[14] E. Terra and C.L.A. Clarke, "Frequency Estimates for Statistical Word Similarity Measures", *Proceedings of North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pp. 244-251, 2003.

[15] M. Jarmasz and S. Szpakowicz, "Roget's Thesaurus and Semantic Similarity", *Proceedings of International Conference on Recent Advances in Natural Language Processing*, pp. 1-5, 2003.

[16] Dung T. Ho and Tru H. Cao. "A High-Order Hidden Markov Model for Emotion Detection from Textual Data", *Proceedings of International Conference on Knowledge Management and Acquisition for Intelligent Systems*, pp. 281-287, 2012.

[17] Klaus R. Scherer and Harald G. Wallbott, "Evidence for Universality and Cultural Variation of Differential Emotion Response Patterning", *Journal of Personality and Social Psychology*, Vol. 66, No. 2, pp. 310-317, 1994.

[18] Hui Yang et al., "A Hybrid Model for Automatic Emotion Recognition in Suicide Notes", *Biomedical Informatics Insights*, Vol. 5, pp. 17-24, 2012.

[19] F.J. Damerau, "A Technique for Computer Detection and Correction of Spelling Errors", *Communications of ACM*, Vol. 7, No. 3, pp. 171-176, 1964.

[20] Peter Norvig, "Simple Python Spell-Checker", Available at: https://github.com/pirate/spellchecker

[21] Daniel Jurafsky and James H. Martin, "*Speech and Language Processing: An Introduction to Natural Language Processing*", 2nd Edition, Springer, 2009.

[22] Maryam Hasan, Elke Rundensteiner and Emmanuel Agu, "EMOTEX: Detecting Emotions in Twitter Messages", *Proceedings of International Conference on Bigdata and Cyber Security*, pp. 27-31, 2014.

[23] Srinivasu Badugu and Matla Suhasini, "Emotion Detection on Twitter Data using Knowledge Base Approach", *International Journal of Computer Applications*, Vol. 162, No. 10, pp. 975-978, 2017.

[24] Munmun De Choudhury, Scott Counts and Michael Gamon, "Not All Moods Are Created Equal! Exploring Human Emotional States in Social Media", *Proceedings of 6th International AAAI Conference on Weblogs and Social Media*, pp. 402-408, 2012.

[25] Aamera Z.H. Khan, Mohammad Atique and V.M. Thakare, "Combining Lexicon-based and Learning-based Methods for Twitter Sentiment Analysis", *Proceedings of National Conference on Advanced Technologies in Computing and Networking*, pp. 1-2, 2015.

[26] Shadi Shaheen, Wassim El-Hajj, Hazem Hajj and Shady Elbassuoni, "Emotion Recognition from Text based on Automatically Generated Rules", *Proceedings of IEEE International Conference on Data Mining Workshop*, pp. 383-392, 2014.

[27] Vinay Kumar and Shishir Kumar. "*Predictive Analysis of Emotions for Improving Customer Services*", IGI Global Publisher, 2017.

[28] Lukas Povoda, Akshaj Arora, Sahitya Singh, Radim Burget and Malay Kishore Dutta, "Emotion Recognition from Helpdesk Messages", *Proceedings of International Workshop on Ultra-Modern Telecommunications and Control Systems*, pp. 310-313, 2015.

[29] David H. Olson, Douglas H. Sprenkle and Candyce S. Russell, "Circumplex Model of Marital and Family System: I. Cohesion and Adaptability Dimensions, Family Types and Clinical Applications", *Family Process*, Vol. 18, No. 1, pp. 3-28, 1979.

[30] NLTK 3.4 Documentation, Available at: https://www.nltk.org/_modules/nltk/tokenize.html.