# EMOTION EXTRACTION USING ENSEMBLE CLASSIFICATION MODEL IN DATA MINING

[1]Pooja Wadhwani, [2]Ajay Kumar

***Abstract - In past few years, a major growth has been seen in platforms of social media. People disseminate opinions, information, behavior, and announcements via social media. They share lot regarding them and their experiences. They frequently share regarding their feelings. This provides information wealth in real-time, regarding emotional state of communities or individuals. There is explosion of interaction on social media. People use Twitter to exchange experience, opinions, and feelings. This huge amount of data can provide important information for mental health research. In this paper, extraction of emotions from the text posted on social media using data mining techniques.***

***Keyword – emotion mining, social media, data mining, tweets, ensemble learning.***

## I. INTRODUCTION

The field of data mining and knowledge discovery has been attracting a significant amount of research attention. An enormous amount of data has been generated every day. Data are being collected and accumulated at a dramatic pace due to the rapid growing volumes of digital data. Data mining is the process of extracting useful information, patterns or inferences from large data repositories and it is used is various business domains. It involves finding valuable information and hidden inferences in large databases [1]. There are many applications of data mining like in medical field, text data mining, sentiment analysis etc.

This paper mainly focuses on emotion mining. Emotions constitute a key factor of human intelligence, which provides indicative characteristics of human behavior, colors the way of human communication and can play an important role in human computer interaction. Emotions play an important role in successful and effective human–human communication. In fact, in many situations, emotional intelligence is more important than IQ for successful interaction. The field of emotion analysis aims at determining emotions present in text such as happy, sad, anger, surprise, love, trust and anticipation. It is also referred to as Affective Computing. Affective computing has application in many areas such as Human-Computer interaction, Depression detection, Brand perception, Social Media sentiment analysis and for Decision making. Affective computing plays a crucial role in building affective interfaces in human-computer interaction with the ultimate goal to make computer understand the emotions and attitude of human so that a computer can interact and respond effectively during the interactions.

Many researchers have put effort in analysis emotions through various sensor channels on UI such as gestures, facial expressions, voice, pitch of the sound etc. This research is the part of digital image processing. However, less effort has been put in detecting emotions from text. Nevertheless, text is an important modality for analyzing emotions because most human knowledge is transmitted via text especially with the emerging field of internet and social media sites.

Emotions have cognitive bases and are shaped by several factors. They impact our basic leadership, influence our social connections and shape our every day conduct. With the quick development of feeling rich literary substance, for example, web-based social networking content, smaller scale blog entries, blog entries, and gathering exchanges, there is a developing need to create calculations and methods for recognizing feelings communicated in content. It has extraordinary ramifications for the investigations of suicide avoidance, decision making in various sectors such as government, business; representative efficiency, prosperity of individuals, client relationship administration, and so forth. Nonetheless, feeling distinguishing proof is very testing because of the numerous reasons. Not at all like Sentiment Analysis, it is a multi-class grouping issue that as a rule includes no less than six essential feelings. Content depicting an occasion or circumstance that causes the feeling can be without unequivocal feeling bearing words, along these lines the refinement between various feelings can be exceptionally unpretentious, which makes it hard to characterize feelings simply by catchphrases.

Table1: Twitter Emotion Analysis

| Emotion | Example |
|---|---|
| **Happiness** | #Happy #Enjoying #PartyyyTime |
| **Sadness** | #HeartBreak #FeelingSick |
| **Love** | #LoveMyFamily #Beautiful |
| **Surprise** | #incredibleBuilding #Wowww |
| **Anger** | #SoDisgusting #Bastard #HateIt |

The design isn't to distinguish particular feelings yet rather to tell if the content contains feelings or not [2]; at the end of the day, if the content is subjective mirroring the author's effect and passionate state or on the off chance that it is verifiable and target where the essayist does not express any sentiments.

## II. RELATED WORK

**Mitra et al.[2002]**provided a survey of the available literature on data mining using soft computing. A categorization has been provided based on the different soft computing tools and their hybridizations used, the data mining function implemented, and the preference criterion selected by the model. The utility of the different soft computing methodologies is highlighted. Generally fuzzy sets are suitable for handling the issues related to understandability of patterns, incomplete/noisy data, mixed media information and human interaction, and can provide approximate solutions faster.

**Yassine et al. [2010]** proposed another system to describe passionate associations in informal communities, and afterward utilizing these attributes to recognize companions from colleagues. The objective is to remove the passionate substance of writings in online informal communities. The intrigue is in whether the content is an outflow of the essayist's feelings or not. For this reason, content mining strategies are performed on remarks recovered from an informal community. The system incorporates a model for information accumulation, database blueprints, information handling and information mining steps.

**Jiang et al. [2017]** proposed an innovative method Word Emotion Association Network (WEAN) to do emotion extraction and sentiment computing of news event. The proposed method consists of two parts: Word Emotion estimation by Word Emotion Association Network and word emotion refinement. In the word emotion computation phase, microblogs with emoticons are considered to calculate the corresponding emotion present in the microblog. For refinement of the emotions derived from the first phase, they used standard sentiment thesaurus. For testing, they used Malaysia Airlines MH370 news event as dataset and computed six basic types of emotions: love, joy, anger, sad, fear and surprise.

**Stojanovski et al. [2005]** exploit an convolutional neural network architecture for emotion analysis in Twitter messages related to sporting events on 2014 FIFA world Cup. In this paper, seven different kinds of emotions were evaluated using hashtag labeled tweets that were collected from Twitter Streaming API. The training of the network is performed on two samples containing 1000 and 10000 tweets on which this approach achieves 50.12% and 55.77% accuracy respectively. Moreover, they have presented the analysis of this approach on three different games that have great impact on Twitter users.

**Mishne et al. [2005]** addressed the task of classifying blog posts on the basis of mood of the writers. They obtained a huge corpus of blog posts from one of the largest online blogging communities Livejournal. The author took the advantage of the Livejournal that allows writers to update their current mood from the 132 given categories. Yahoo API was used to get a list of 1000 web pages containing a Livejournal blog post with each kind of mood. They used variety of feature sets such as Frequency-Counts, Length related, Semantic Orientation Features and the most useful one Mood PMI-IR (Pointwise Mutual Information). For the experimental analysis, they used SVMlight (Support Vector Machine package) for classifying the mood of a blog post. Two set of experiments were performed. The first set evaluates the specific individual mood in a blog post while in the second set of experiments moods were partitioned into two mood sets as positive and negative. This was done to enhance the performance of the classification. The uniqueness of this paper is represented in the results and discussion segment which shows that the substantial increase in the training data increases the accuracy of the classification algorithm.

**Yang et al. [2014]** exploit a novel approach for extracting emotions. They used graphical emojis, accentuation articulations alongside a conservative vocabulary to mark information. They gave a multi-name feeling arrangement calculation (MEC)to analyze emotions in short text of Weibo, which is a very famous online social networking site in china (just like Twitter). The approach they used is phycology independent for it worked well on different phycology theories for emotion classification. In the proposed approach they exploit K-nearest neighbor (KNN) for tweet level analysis and Naïve Bayes for Word level analysis of emotion. Moreover, their approach outperformed various state-of-art methods as discussed in experiment and results. The dataset contained tweets about Malaysia 370 missing flight and they concluded from their approach that the flare-up of Anger has a postponement in the wake of limit of Sadness.

**Catal et al. [2017]** exploit a sentiment classification model based on Vote ensemble classifier utilizes from three individual classifiers: Bagging, Naïve Bayes and Support Vector Machines. Moreover, in bagging they used SVM as base classifier. The main focus of this research is to improve the performance of machine learning classifiers for sentiment classification of Turkish reviews and documents. Their experimental results show that multiple classifier system based approaches are much better for sentiment classification of Turkish documents. They performed experiments on three different domains such as book review, movie reviews and shopping reviews. The authors concluded that this approach is not restricted to just one domain and can be extended to several other domains as well.

**Wang et al. [2012]** exploit a technique to automatically annotate a large amount of data. They extracted large amount of tweets (2.5 million) from twitter instead of using already annotated corpus which consists of just thousands of tweets. The main focus of this research is to study the effectiveness of different element mixes and the impact of the extent of the preparation information on the emotion analysis task. To automatically annotate data they extracted the tweets using 131 relevant keywords for seven emotion categories such as joy, sadness, anger, love, fear, thankfulness and surprise. They explored variety of features such as n- grams, emotion lexicons , parts-of –speech etc. Moreover, they performed experiment on two different machine learning algorithms LIBLINEAR and multinomial Naïve Bayes. The highest accuracy of 65.57% was achieved using an enormous dataset of 2 million tweets.

**Yeole et al. [2015]** presented an effective technique for emotion analysis of social media text. The uniqueness of this reseach is that they have not used only direct affective words for emotion extraction, but also the indirect sentences bearing emotions are taken into account and suitable NLP techniques are applied to calculate the relevant emotion. A novel technique for emotion extraction has been presented in which they have considered direct emotion bearing words and indirect emoticons and smiley faces as well. The feature extraction phase utilizes emotion dictionaries such as SentiWordNet 3.0. Effective preprocessing was done to remove noisy data and stop words. They used Fuzzy and rule based systems for the prediction of the emotion for a particular sentence.

**Perikos et al. [2016]** designed an ensemble classifier schema by combining statistical machine learning classification methods and knowledge based approach for the task of recognizing emotions in various domains such as news, headlines articles and social media posts. Moreover, the ensemble is based on the three classifiers: Naïve Bayes, Max. Entropy and knowledge based method. Furthermore, the majority voting scheme is used to combine the results of the classifiers. For training the classifiers, the corpus used is publicly available ISEAR dataset and Affective text datasets. However, for evaluation they created corpus from different sources and manual annotation was done by a human expert. They compared results with several traditional methods and the results show that ensemble classifiers are more effective than sole classifiers.

**Jain et al.[2017]** presented propelled system for location of feelings of clients in Multilanguage content information utilizing feeling speculations which manages phonetics and brain science. The feeling extraction framework is produced in light of various highlights bunches for the better comprehension of feeling dictionaries. Exact investigations of three constant occasions in areas like a Political decision, human services, and games are performed utilizing proposed system.

**Seol et al.[2008]** proposed emotion recognition system. Emotions can be expressed by various type of mediums like image, speech, facial expression, and so on. This paper focused on textual data. This hybrid system utilize two techniques, first is machine-learning method and keyword based.

**Roberts et al.[2012]** introduced a corpus collected from Twitter with annotated micro-blog posts (or "tweets") annotated at the tweet-level with seven emotions: ANGER, DISGUST, FEAR, JOY, LOVE, SADNESS, and SURPRISE and analyzed how emotions are distributed in the data we annotated and compare it to the distributions in other emotion-annotated corpora. This paper used the annotated corpus to train a classifier that automatically discovers the emotions in tweets and presented an analysis of the linguistic style used for expressing emotions.

**Houjeij et al. [2012]** designed a system that adopts a novel approach for emotional classification from human dialogue based on text and speech context. The main objective is to boost the accuracy of speech emotional classification by accounting for the features extracted from the spoken text.

**Dhawan et al.[2014]** presented a new perspective for studying emotions' expression in online social networks. This method is unsupervised; it basically utilizes k-means clustering and nearest neighbor algorithm. Experiments depicts high accuracy for model to determine subjectivity of texts and predicting emotions.

## III. PRESENT WORK

The methodology followed for the proposed work is as follows:

1. **Data Acquisition and Annotation**: The first and foremost step is to collect data for emotion analysis. Tweets are collected from twitter using Twitter API on Demonetization. Annotation of tweets is done manually on the basis of Ekman's six basic emotion and two more emotions (mixed emotion and no emotion) are added.

2. **Preprocessing and Filtration**: Each word in tweet is vital to make decision, therefore effective pre-processing of these tweets is an important task because these tweets are full of slang, misspellings and words from other languages. Therefore, in order to deal with this kind of noisy data , normalization of tweets is performed by intelligent text pre-processing techniques like tokenization, stop-word removal, stemming, lemmatization dimensionality reduction. Pre-processing of data is performed by the following techniques:

- Data Cleaning: Cleaning of data involves handling of missing values by ignoring that particular tuple. If any tuple or cell is empty then that will be filled with some specific value. Inconsistency of data may be handled manually. It also handles noisy data by implementing machine inspection, clustering, binning methods and regression. All the quotes("") from the sentences are removed, URL's are removed and other characters that are not considered to be in the category of texts are removed.
- Data integration: Data is always collected from various sources like data warehouse, internet etc. so the collected data in particular so not have any use. It has to be added altogether for further analysis. So this step will integrate the data collected from various sources.
- Data transformation: Transformation of data means to change the data from one form to another. For this purpose various methods like smoothing, normalization, aggregation and generalization are available for the transformation. Transformation steps are as follows:
  - Sentence Splitting: The first step involved is sentence splitting i.e the splitting of string into words. Identifying sentence boundaries in a document is not a smaller task.
  - Tokenization: Tokenization of words means to split a sentence into tokens or smallest unit of a sentence. Tokenization is an important task because many succeeding components need tokens clearly identified for analysis.

- Stop Word Filtering: There are a lot of words that do not have any meaning and can be removed from the input file. Words like "the", "and","for","or","if","that"; are referred to as stop words because they don't signify any meaning or sentiment. Therefore, removal of such words means stop word filtering and it also improves the performance of the system
- Stemming: A stemming algorithm is a process of linguistic normalization. In this process, the variations of a word are reduced to a common form. For instance, consider a simple example below:

  Communication  
  Communications  
  Communicative ------→ Communicate  
  Communicated  
  Communicating

**3.Classification:** Classification is done to automatically classify tweets into an emotion category. Ensemble machine learning algorithm is used to train the system.

In order to improve the performance of individual classifiers used in the paper is to the use of ensemble learning. While the advantage of using ensemble, methods is the improvement of the performance, the disadvantage is about the time it takes to finish the training phase. However, the main concern was to build a model which has a better performance compared to the individual classifiers. This model is novel because not only it uses ensemble method (Voting), but also it applies a meta classifier as one of its classifier component. Also, parameter optimization approach was used on its individual classifier. Each component in our model learns some parts of the classification problem and we combine these hypotheses to decide the probability level.

**4. Evaluation:** Evaluation of performance parameters is done using metrics like accuracy, precision, recall, F-measure. Moreover, FP rate, TP rate are calculated.

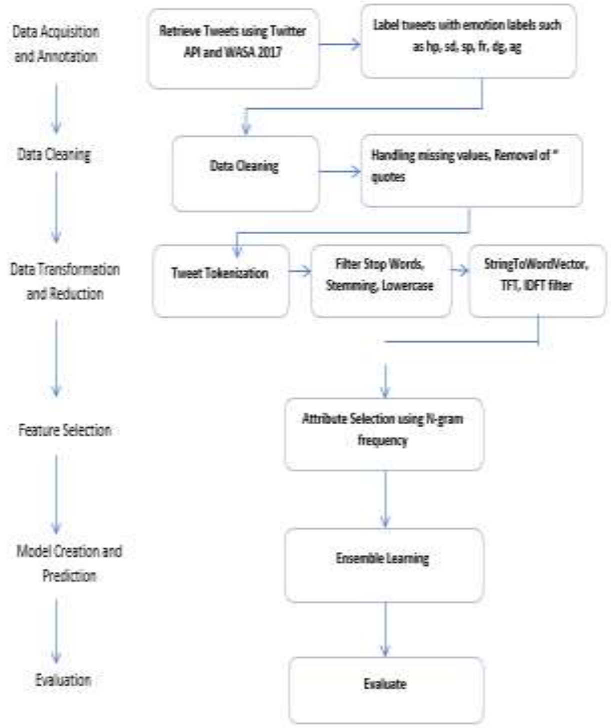
Data Acquisition and Annotation
Retrieve Tweets using Twitter API and WASA 2017
Label tweets with emotion labels such as hp, sd, sp, fr, dg, ag
Data Cleaning
Data Cleaning
Handling missing values, Removal of " quotes
Data Transformation and Reduction
Tweet Tokenization
Filter Stop Words, Stemming, Lowercase
StringToWordVector, TFT, IDFT filter
Feature Selection
Attribute Selection using N-gram frequency
Model Creation and Prediction
Ensemble Learning
Evaluation
Evaluate

Figure 1: Flowchart of Proposed Technique

## IV. RESULTS

### Results Evaluation

Various parameters are used to evaluate the results of the proposed technique. These parameters are:

### Recall

Recall is measurement that is generally used to evaluate execution in content mining, and in content examination field like data recovery. This parameter is utilized for estimating completeness.

$$\text{Recall}=\frac{\text{True Positive}}{\text{True Positive+False Negative}}$$

### F-Measure

F-Measure is the harmonic mean of precision and recall. The esteem computed utilizing F-measure is a balance among precision and recall.

$$\text{F measure }=\frac{2*\text{recall}*\text{precision}}{\text{precision+recall}}$$

**Mean Absolute Error**

The MAE measures the ordinary size of the bumbles in a game plan of checks, without pondering their bearing. It evaluates precision for interminable components. The condition is given in the library references. Imparted in words, the MAE is the ordinary over the affirmation trial of the preeminent estimations of the differentiations among guess and the looking at discernment. The MAE is a straight score which infers that all the individual differences are weighted comparably in the typical.

$$MAE = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j|$$

**Root Mean Squared Error**

The RMSE is a quadratic scoring guideline which measures the normal size of the mistake. The condition for the RMSE is given in both of the references. Communicating the recipe in words, the distinction amongst figure and relating watched values are each squared and after that found the middle value of over the example. At last, the square foundation of the normal is taken. Since the mistakes are squared before they are found the middle value of, the RMSE gives a generally high weight to expansive blunders. This implies the RMSE is most valuable when expansive blunders are especially unwanted.

The MAE and the RMSE can be utilized together to analyze the variety in the mistakes in an arrangement of conjectures. The RMSE will dependably be bigger or equivalent to the MAE; the more noteworthy distinction between them, the more prominent the fluctuation in the individual blunders in the example. On the off chance that the RMSE=MAE, at that point every one of the mistakes are of a similar greatness.

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

Table 2: Class Parameters Comparison

| Parameter | SVM | Naïve Bayes | Proposed |
|---|---|---|---|
| Precision | 0.875 | 0.759 | 0.879 |
| Recall | 0.81 | 0.743 | 0.879 |
| F-measure | 0.824 | 0.747 | 0.878 |

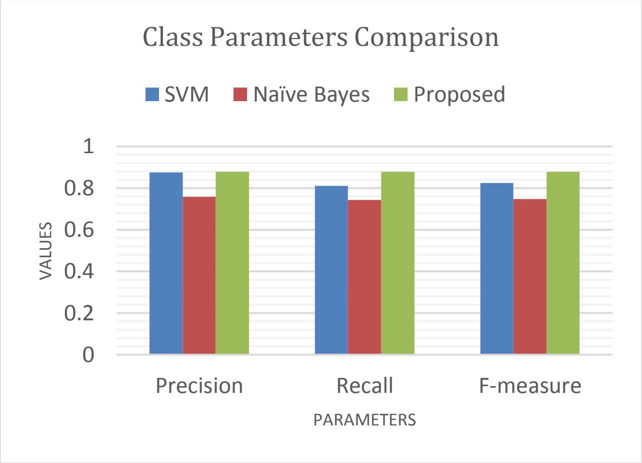

Figure 2: Showing the class parameters comparison of the proposed technique with the existing techniques

The figure above shows the results of proposed technique is better than the existing individual techniques as the values for precision, recall and Fmeasure is large in case of proposed ensemble technique.

Table 3: Error Comparison

| Parameter | SVM | Naïve Bayes | Proposed |
|---|---|---|---|
| Mean absolute error | 0.0634 | 0.1094 | 0.0588 |
| Root mean squared error | 0.2517 | 0.2444 | 0.1726 |
| relative Absolute error | 0.2284 | 0.393 | 0.2125 |
| root relative squared error | 0.6759 | 0.6557 | 0.464 |

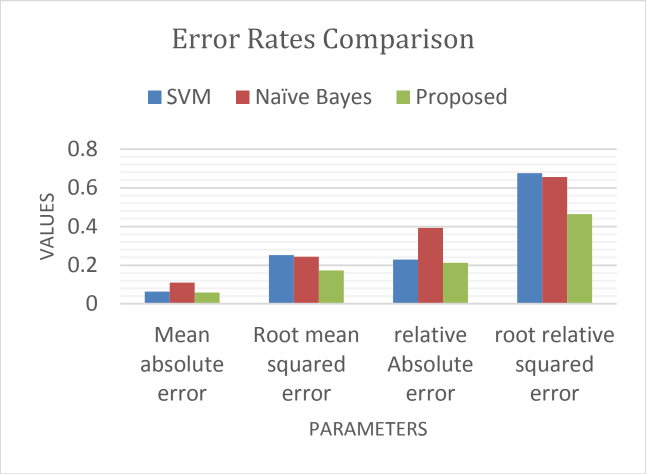

Figure 3: Showing the error rate comparison of the proposed technique with the existing techniques

Table 4: Classification Accuracy Comparison

| Parameter | SVM | Naïve Bayes | Proposed |
|---|---|---|---|
| Accuracy | 80.9904 | 74.2812 | 87.8594 |

The table above shows the comparison of accuracy of the proposed technique Ensemble learning is 87.8594 which is better than the existing naïve bayes and svm which is having 74.2812 and 80.9904.

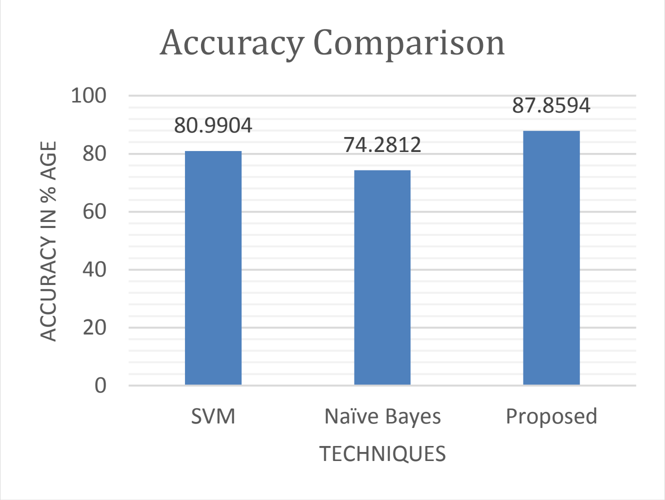

Figure 4: Showing comparison of Accuracy

The figure above shows the proposed technique ensemble learning is having highest accuracy which is 87.8594, the accuracy of the individual naïve bayes and svm is 74.2812 and 80.9904.

## V. CONCLUSION

The increasing number of social media websites by Internet users has raised the interest about the opportunity to understand the relation between people’s preferences and actual political behavior. We have observed that twitter is very commonly being used as a platform for deliberation by citizens of India. The social media is a powerful and reliable source of public opinion as far as a nation like India is concerned. The discussions on twitter are equivalent to traditional discussions and are capable enough to give a fair idea of emotions of general public.

Emotions have cognitive bases and are shaped by several factors. They impact our basic leadership, influence our social connections and shape our every day conduct. With the quick development of feeling rich printed content, for example, web-based social networking content, smaller scale blog entries, blog entries, and gathering discourses, there is a developing need to create calculations and strategies for distinguishing feelings communicated in content. It has incredible ramifications for the investigations of suicide aversion, decision making in various sectors such as government, business; representative efficiency, prosperity of individuals, client relationship administration, and so forth. Content depicting an occasion or circumstance that causes the feeling can be without express feeling bearing words, subsequently the qualification between various feelings can be extremely unpretentious, which makes it hard to arrange feelings absolutely by catchphrases’. This work mainly focuses on ensemble learning based emotion extraction. The expected results would be the increase in accuracy of learning algorithm.

In future, neural network with fuzzy can be used for the same purpose and also to reduce the training time of ensemble learning, features can be optimized using optimization techniques. One of the tasks is to consider emotion intensity for classification. Explore the relation between emotion classes and emotion intensity. Content-based analysis of emotion data is yet another possible line of research. Data sets containing emoticons, stickers and other images with texts representing emotions can also be taken into consideration in future.

## REFERENCES

[1] Mitra, Sushmita, Sankar K. Pal, and PabitraMitra. 2002. Data mining in soft computing framework: a survey. IEEE transactions on neural networks. 13(1): 3-14.

[2] Mohamed Yassine and Hazem Hajj. 2010. A Framework for Emotion Mining from Text in Online Social Networks. IEEE International Conference on Data Mining Workshops.

[3] Jiang, Dandan. 2017. Sentiment Computing for the News Event Based on the Social Media Big Data. IEEE Access 5: 2373-2382.

[4] Stojanovski, Dario. 2015. Emotion identification in FIFA world cup tweets using convolutional neural network. Innovations in Information Technology (IIT), 11th International Conference on. IEEE.

[5] Mishne, Gilad. 2005. Experiments with mood classification in blog posts. Proceedings of ACM SIGIR 2005 workshop on stylistic analysis of text for information access. 19.

[6] Yang, Jun, Lan Jiang, Chongjun Wang, and JunyuanXie. 2014. Multi-label Emotion Classification for Tweets in Weibo: Method and Application. In Tools with Artificial Intelligence (ICTAI), 2014 IEEE 26th International Conference. 424-428.

[7] Catal, Cagatay, and Mehmet Nangir. 2017. A sentiment classification model based on multiple classifiers. Applied Soft Computing 50: 135-141.

[8] Wang, Wenbo, Lu Chen, KrishnaprasadThirunarayan, and Amit P. Sheth. 2012. Harnessing twitter" big data" for automatic emotion identification. In Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom), 587-592.

[9] Yeole, Ashwini V., P. V. Chavan, and M. C. Nikose. 2015. Opinion mining for emotions determination. In Innovations in Information, Embedded and Communication Systems (ICIIECS), 2015 International Conference. 1-5.

[10] Perikos, Isidoros, and IoannisHatzilygeroudis. 2016. Recognizing emotions in text using ensemble of classifiers. Engineering Applications of Artificial Intelligence 51: 191-201.

[11] Vinay Kumar Jain, Shishir Kumar, Steven Lawrence Fernandes. 2017. Extraction of emotions from multilingual text using intelligent textprocessing and computational linguistics. Journal of Computational Science. 316–326.

[12] Yong-Soo Seol, Dong-Joo Kim, Han-Woo Kim. 2008. Emotion Recognition from Text Using Knowledge – based ANN. International Technical Conference on Circuits/Systems. 1569-1572.

[13] Kirk Roberts, Michael A. Roach, Joseph Johnson, Josh Guthrie, Sanda M. Harabagiu. 2012. EmpaTweet: Annotating and Detecting Emotions on Twitter. 3806-3813.

[14] Ali Houjeij, Layla Hamieh, Nader Mehdi, Hazem Hajj. 2012. A Novel Approach for Emotion Classification based on Fusion of Text and Speech. IEEE, International Conference on Telecommunications.

[15] Sanjeev Dhawan, Kulvinder Singh, Vandana Khanchi. August 2014. A Framework for Polarity Classification and Emotion Mining from Text. International Journal Of Engineering And Computer Science. 3(8):7431-7436.