Scientific Journal of Impact Factor (SJIF): 4.72

e-ISSN (O): 2348-4470
p-ISSN (P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 4, Issue 7, July -2017

Sentiment Analysis with Multilayer Perceptron using Emoticon Space Model

[1]Seema Chithore, [2]Mrs. D. A. Phalke

[1,2] *Department of Computer Engineering, D.Y.Patil College of Engineering Akurdi, Pune*

Abstract—*Sentiment analysis detects that text segment contains emotional or opinion content and determines emotional class of it so as to take into consideration. Now a days sentiment analysis is important to know people reviews about particular product or service as, from it people come to know the profit and loss in any occupations. Emoticons play important role in expressing the emotions. Emoticons promptly indicate emotion as compare to the text. Emoticon space model avoid overhead of manually assigning label to several known emoticons. ESM projects the microblog post into space and determine the co-ordinates, required to classify microblog post effectively. MLP have remarkable ability to derive meaning from complicated or imprecise data. MLP fetch out the patterns and identify data that are too complex and detect emotions out of it.*

Index Terms*— microblog sentiment analysis, emoticon space, polarity classification, subjectivity classification, emotion classification, Multilayer Perceptron(MLP).*

I. INTRODUCTION

Internet connects the people online. The first developed online communication media, such as e-mail or forums were based on text messages. After improvement and popularization of the internet connections gave the facility of phone calls, video conferences, the text-based message did not lose its popularity. Different strategies have developed to use emoticons, strings of symbols imitating body language. Today the use of emoticons in online conversation adds values to the facilitation of the online communication process in e-mails, instant messaging applications or blogs [1]. Now a days web provides a way to produce large volume of opinions by allowing people to express their opinion on various social networking sites like Twitter, Facebook [1]. Finding the sentiment or kind of reaction of the people within the tweet is important for planning the marketing strategies [2].

In various domains like politics, finance, industry, business [3] sentiment analysis plays a key role to know people review and take actions as per that so as to avoid negative circumstances about particular situation or considering profit in the market related to a product. Sentiment analysis plays a key role in identifying the nature of the review like those is positive or negative. From any of the positive, negative aspects in which area it is affecting most.

Sentiment analysis is done at different levels. Specifically at document, sentence, word and attribute level. In the document level polarity analysis, whole document is considered as a single entity. At the sentence level the sentence is considered as an entity. Word level polarity analysis contains analysis and classification of each word. Attribute level identifies each and every attribute of an entity and identifies the emotion of each attribute [4]. Other than this there are other forms such as using complex linguistic information, like the term part-of-speech or the term position within the given text. It is very simple to identify relation between terms then find sequence to identify meaning. Beside this many supervised classifiers are used for polarity analysis [5].
Text messages analysis some time is complex as those do not reflect clear meaning some time. Emoticons promptly reflect emotions as compare to the text. Emoticons signals are stronger and help in the overall classification phase. Those help to identify faster and more accurately. Number of emoticon affect the classification as stronger signals would get. Number of emotions are used which are broadly classified in few classes just like positive, negative. But going more deeply, each emoticon has its own unique meaning and classified in more deep emotion class like happy, sad, anger, clap etc [6].

There are various approaches of classification which are mainly divided into Machine learning approach and Lexicon based approach. In machine learning approaches various supervised and unsupervised algorithms and linguistic features are used. Lexicon based approach is based on collection of known terms. Hybrid approach combines this two main approaches as per need of application to give more accurate result [7]. In lexicon-based approaches, dictionaries can be created manually or automatically to expand the list of words by using seed words [8]. Including this there are other sentiment analysis methods Emoticons, LIWC, SentiStrength, SentiWordNet, SenticNet, SASA, Happiness Index, PANAS-t [9]. In some cases heuristic rules are used to find the relation among the words and influence on sentiment analysis is determined and sentiment is updates as per impact [10].

@IJAERD-2017, All rights Reserved

Social media plays very important role in communication at broader aspect. Acts a mediator between social networks, personal information channels and the mass media. Social media represent the flow of information generation, transferred and used. Lot of user generated data in the form of blogs, comment, opinions connects the producer and the consumer. This helps the producer to react to the consumer reaction as soon as possible [11]. Automatic part-of-speech tags and resources such as sentiment lexicons are the features have proved useful in multiple domain for sentiment analysis but are less useful in sentiment analysis of twitter data [12].

II. REVIEW OF LITERATURE

Author Jiang et al. [13] proposed Emoticon Space model (ESM) for microblog sentiment analysis, obtains a relatively high performance when the size of manually labeled training set is small and can effectively perform lexicon-based polarity classification which does not require manually labeled data. From the message the text and emoticons are extracted, the cosine similarity is calculated between then to project into space. After projecting co-ordinates from the space are taken as input to classification task.
Author Huang et al. [14] proposed the model used to explore the relationships among emoticon use, information richness, personal interaction, perceived usefulness and enjoyment. Author identified the potential effects of emoticons, specifically, on the relationships between emoticon use and factors related to the use of information and management.

Author Hassan et al. [15] introduced a new system that uses the different emotions recognized from the entire conversation in classifying the call. Proposed the model consists of three main phases as signal segmentation, emotions recognition and calls classification.

Author Liu et al.[16] presented how to adapt language models for Sentiment Analysis and proposed a very effective and efficient way to learn the emoticon model from Twitter API then proposed Emoticon smoothed language model(ESLAM) seamlessly integrate manually labeled data and noisy labeled data for training.

Author Yuasa et al. [17] experimentally showed emoticons convey emotions without the cognition of faces and that these results plays important role in investigation of abstract faces affect our behaviors. Author identified the brain activities that are related with emoticons, the most abstract faces by using functional MRI (fMRI) and described the results of the experiment with appropriate remarks.

Author Cho et al.[2] proposed the approach to sentiment classification at paragraph length using contextual information and sentiment-based domain dictionaries covering formal and informal vocabularies. To classify the sentiments of paragraph length texts in social network services, author calculates contextual information based on domain based keywords, the position of the sentence, and the flow of sentiments.

Author Hogenboom et al. [1] proposed a novel framework for automated sentiment analysis, which takes into account the information conveyed by emoticons. The goal of framework is to identify emoticons, decide their sentiment, and allocate the associated sentiment to the affected text in order to correctly classify the polarity of natural language text as either positive or negative. Author showed the sentiment associated with emoticons typically dominates the sentiment conveyed by textual cues in a text segment.
Author Recupero et al. [18] proposed Sentilo, is an unsupervised, domain independent system, with hybridizing natural language processing techniques and semantic web technologies it accomplish sentiment analysis. Sentilo provides its output as a RDF graph, and whenever possible it resolves holders and topics identity on Linked Data.

Author Saif et al. [19] Introduced a novel approach of adding semantics as additional features into the training set for sentiment analysis which results into better accuracy score for identifying both negative and positive sentiment. Author implemented a new set of semantic features for training a model for sentiment analysis of tweets, investigated three applications for adding such features into the training model, first by replacement, second by argumentation and by interpolation and show the superiority of the final approach.

Author Hu et al. [20] presented a mathematical optimization formulation that include the sentiment uniformity and emotional contagion theories into the supervised learning process; and utilize sparse learning to tackle noisy texts in Microblog. Author has pointed the problem of sentiment analysis in microblogging to take advantage of social relations for sentiment analysis. Author presented a novel supervised method to handle the noisy and short texts by integrating sentiment relations between the texts.

@IJAERD-2017, All rights Reserved

III. SOFTWARE REQUIREMENT SPECIFICATION

With the collected tweets the system performs the preprocessing. From the preprocessed tweets features gets extracted and are used for the classification task. The emotions of the tweets are identified and could be used to recommend the product to other users, owner of the product to know its positive and negative values from people. Tweets with English language are considered as an input to the system. The Twitter API provides up-to-date information; limited only by the rate of Twitter input. The functionality of the software is depending on any external services such as internet access that are required.

A. Hardware Specification

1) 4GB RAM for Windows 7 and above,
2) 20 GB of HARD DISK,
3) Processor Intel Pentium 3 and above or equivalent.

B. Software Specification

1) Windows OS,
2) Eclipse
3) JDK 1.8

IV. SYSTEM ARCHITECTURE / SYSTEM OVERVIEW

The proposed system is divided into two phases as Training and testing i.e. Recognition. Each phase has three steps as preprocessing of data, feature extraction and projection of tweets into ESM and Classification using MLP. Figure 1 shows the system architecture of the proposed system.

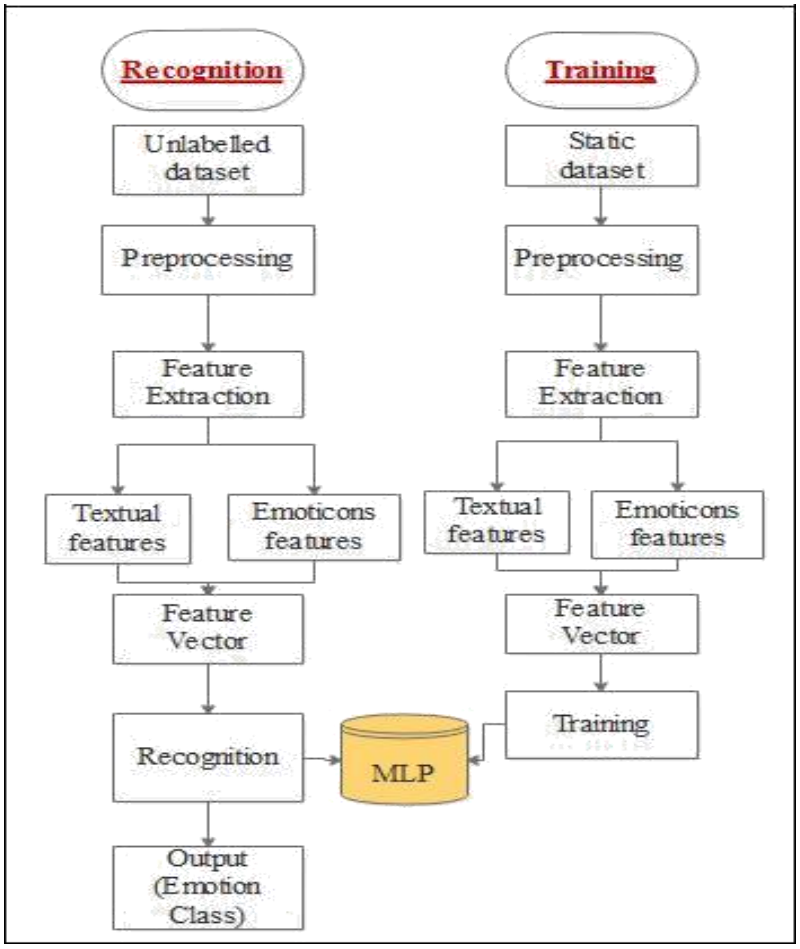

Figure 1. System Architecture of Proposed System

@IJAERD-2017, All rights Reserved

Firstly the input tweets are taken from the twitter API or saved file of tweets. Input data is preprocessed as follows:

* Filter English language.

* Stemming and Lemmatization.

* Spelling correction.

* Remove stop words, question words, URLs, special characters Expand abbreviations and Replace slangs.
* Part of speech tagging as

{Adjectives, Adverbs, Nouns, Verbs}
The emotions from the messages are tagged against 10 different classes defined in the proposed system, those are ”Happy”, ”Sad”, ”Angry”, ”Very Happy”, ”Clap”, ”Heart”, ”Sorrow”, ”Worried”, ”Surprise”, ”Mute”. MLP is trained using the coordinates of the posts as features which are obtained from projection of the blog into ESM. In ESM, number of frequently adopted emoticons is used to construct an emoticon space.
ESM consists of two phases:

1) Projection phase

2) Classification phase

Projection phase:

The Coordinates of the words are obtained from the semantic similarity between word and emoticon. Words coordinates are then considered for the calculation of the blog coordinates and then they are projected into the space [13].

Classification phase :

Coordinates of the post are used as features for the classification task. Distributed representation of words helps to learn semantic similarity between words and emoticons effectively. Based on the distributed representation post are projected into the emoticon space and then classification is done using supervised sentiment classifier [13].

Distributed Representation of Words :

Distributed representation corresponds to many relation between the two types of representation. The distributed representation of words is used mostly in neural probabilistic language models (NNLMs). In the distributed representation each word is represented by the vector. The words with similar meaning assumed to be having same vector. These vectors are then used for projection phase. word2vec is used to learn the distributed representation of words due to its fast speed[13].

Word Projection :

While distributed representation the post are preprocessed and each emoticon is treated separately. The Words with similar meaning are considered to have similar vector. Thus the by measuring the similarity between the representation vector of word wi and emoticon e_j helps to calculate the semantic similarity. The similarity of the representation vector is measured by cosine distance which is formalized as in equation 1.

$$\text{similarity}(w_i;\ e_j) = wi.\frac{e_j}{|w_i||e_j|} \quad (1)$$

where
w_i and e_j are the representation vectors of w_i and e_j.
Specifically, if $w_i = e_j$, the similarity between the representation vectors is 1. Equation 1 is used as the measurement of semantic similarity between w_i and e_j and use the semantic similarity as the coordinate of the word w_i in dimension j [13].

@IJAERD-2017, All rights Reserved

Microblog Post Projection :

Simple method for projecting the emotions into the emoticon space is by using the semantic similarity between the posts and emoticons. Semantic similarity cannot be determined directly while the coordinates of the post are obtained by the using the mathematical operation on the coordinates of the words. There two simple strategies for post projection [13].

1) Basic ESM (B-ESM)

2) Extended ESM (E-ESM)

Basic ESM (B-ESM) :

One way to project the post into the emoticon space is to obtain the post coordinates by summing up the coordinates of the words and assume P be the coordinate of the post.

$$P = \sum_{w_i \in P} C(i,j) \qquad (2)$$

Due to the simplicity of the strategy the name is given as Basic ESM [13].

Extended ESM(E-ESM):

Subjective post contains several sentiment words; one could be representing the sentiment of overall post. Sentiment word could be showing semantic similarity with some emoticons and dissimilarity with others. Thus the coordinates of the words may vary respectively. The sentiment of this post is indicated by maximum and minimum values of word coordinates in certain dimensions.[13]. E-ESM integrates all this information. Using the B-ESM, the minimum and maximum values are added [13].

Supervised Sentiment Classification

After projection of the post into the emoticon space the classification is done. In the classification the coordinates are considered as features. The B-ESM is used as an advantage of the emoticons signals is taken [13].

In this paper with E-ESM strategy the Multi Layer perceptron (MLP) is used. MLP is a feed forward neural network having one or more layers called as hidden layer between input and output layer as shown in fig.2.
MLP is effectively used in pattern classification, recognition, prediction and approximation. Linearly not separable problems are solved by MLP [22].

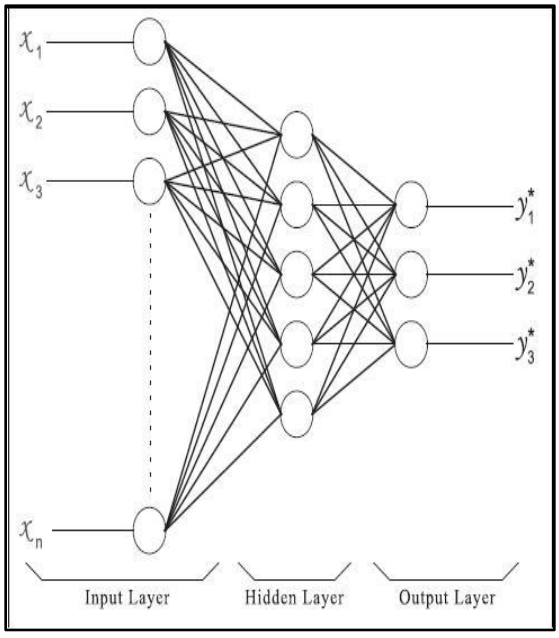

Figure 2. The structure of the MLPNN model [21]

@IJAERD-2017, All rights Reserved

In the classification phase, the multilayer perceptron neural network is used. In the classification the tasks are divided as follows and all formulas are taken from reference [22].

* Input calculation.

* Forward Propagate.

* Back Propagate Error.
* Train Network.

Input calculation:

$$y = (w_1x_1 + w_2x_2 + \ldots + w_nx_n) \quad (3)$$

Where,

w_i is the weight of the interconnected link between neurons of the two layer either input and hidden or hidden and output.
x_i is the input neurons

Forward Propagate:

Sigmoid activation function is used in the proposed system as the data is not linearly separable.

$$f(v) = \frac{1}{1+e^{-y}} \quad (4)$$

Back Propagate Error:

This is broken down into two sections. fc={fi,fii}

fi= Transfer Derivative.
fii= Error Backpropagation.

Transfer Derivative:

$$\text{derivative} = \text{output} * (1 - \text{output}) \quad (5)$$

Error Backpropagation:
Error Backpropagation at the output layer is performed as:

$$error = (expected - output) * derivative \quad (6)$$

Error Backpropagation at the hidden layer is performed as:

$$error = \left(w_i - error_j\right) * derivative_i \quad (7)$$

Train Network :
Network weights are updated as follows:

$$weight = weight + learning_{rate} * error * input \quad (8)$$

In the proposed system, In the training phase the MLP structure is created for each data class considering the input of the neurons in the layer and the weight of the connected links. Considering multiple dataset most apparent structure is selected. In the testing phase, unlabeled data is tested against all the structure to get classified against the particular class.

@IJAERD-2017, All rights Reserved

V. SYSTEM ANALYSIS

A. Algorithm

Vector representation of words is learned from word2vec. The representation vectors of words form a matrix $M_w \in R^{dV}$, where V is the size of the vocabulary and d is the dimension of the representation vectors. Each column of M_w denotes the representation vector of the corresponding word. Suppose E emoticons (denoted as (e1, e2, ..., e_E)) are selected to construct the emoticon space. Representation vectors of these emoticons are searched in matrix M_w and receive a matrix $M_e \epsilon R^{dE}$. M_e is a sub-matrix of M_w. Each column in Me denotes the representation vector of the corresponding emoticon [13]

Algorithm: Calculation of the Coordinates of Words and Emoticons

Require:

Distributed representation matrix of words, M_w.

Distributed representation matrix of emoticons, M_e.

for each i in [1 : E] do do

for each j in [1 : V] do do

C(i, j) = similarity(M_w(:, j),M_e(:, i))

end for

end for

B. Dataset

In this paper Twitter dataset is used. Blogs are given as input to the system. Twitter is an online news and social networking service in which user can post and read messages called as tweets. The size of the tweet is restricted to 140 characters. In this registered users can post and read the messages while the unregistered users can only read the posts. Using different authentication keys particular account posts can be accessed. 200 tweets are allowed on the twitter timeline with 16 pages at a instance. 20 messages are extracted at a instance in the proposed approach as an input to the system[23].

VI. IMPLEMENTATION

A proposed system is implemented in Java. The system is mainly divided into three modules as preprocessing, feature extraction and classification using supervised multilayer perceptron neural network.

Tweets are given as input to the system. The secured connection is established and up to date tweets are extracted for the processing. In tweets using internet the emoticons are reflected as it is from the Twitter emoji scanner - GitHub, the respective image is extracted with the url as https://abs.twimg.com/emoji/v1/72x72/image.png.

Table I shows the example emoticons used in the system with their clear sentiment meaning. The proposed system consist of two phases, first is the training and testing. In both the phases data is projected into space to calculate the co-ordinates and finding the cosine similarity between the text and emoticons used in the data. In the training phase data is trained with multilayer perceptron neural network. The structures for different classes are created with training data by calculating their vectors. Using the structure the testing data which is unlabeled is test against identified classes with their structure.

@IJAERD-2017, All rights Reserved

TABLE I : EMOTICONS HAVING CLEAR EMOTIONAL MEANINGS[13].

Sentiment	Example Emoticons
Positive	
Negative	
Happiness	
Like	
Sadness	
Disgust	

Output of the data and sigmoid value is calculated. Activation function is sigmoid is applied as data is not linearly separable. With the structure it comes to know the data is nonlinearly separable and effectively class determination is done by multilayer perceptron neural network.

VII. PERFORMANCE MEASURES USED

Accuracy is used for the results evaluations. The equations used for these performance measures are presented as follows:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (9)$$

VIII. RESULT TABLES

The expected accuracy of the proposed system is as follows.

TABLE II : Accuracy of the proposed system

Number of Emoticons	Accuracy
2	0.821
4	0.832
6	0.843
8	0.846
10	0.849
12	0.847
14	0.856
16	0.861
18	0.868
20	0.876
22	0.85

@IJAERD-2017, All rights Reserved

In the result the projection phase would be calculating the vector representation and words coordinates. Using this word coordinates as features the multilayer perceptron neural net-work gives more accurate polarity analysis of the blog. The comparison of the result of proposed system with Emoticon Space Model is as shown in figure 3.

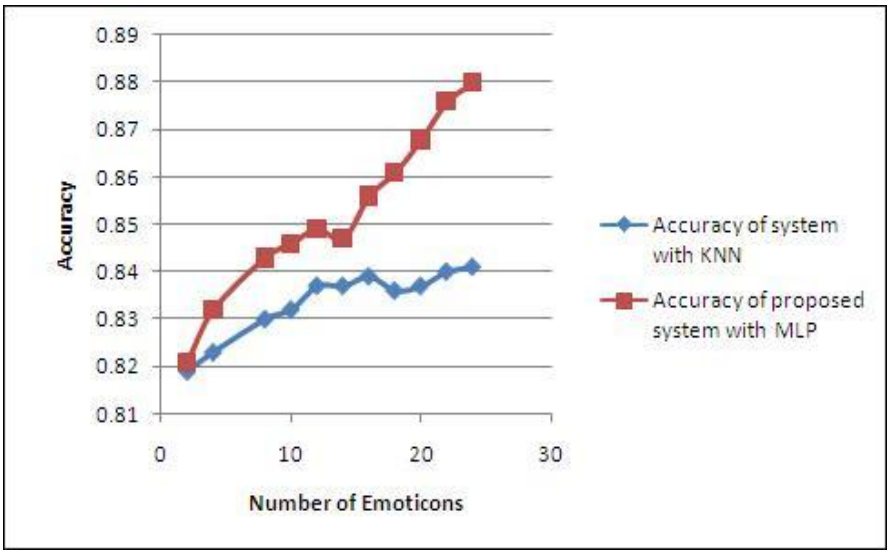

Figure. 3. Comparison of the proposed system accuracy with existing system.

IX. CONCLUSION

In this paper it is shown that messages with emoticons could be classified more accurately by using multilayer perceptron even if training data is minimum by projecting input data into emoticon space model. Emoticons express different moods, emotions, feelings in microblog environment. So they are considered as important part of microblog sentiment analysis. Some emoticons have clear emotional meaning while some do not express clearly. By projecting words and microblog posts into an emoticon space, the ESM model helps identify subjectivity, polarity, and emotion in microblog environments. MLP helps to deeply classify message with proper class even if particularly related data is not train by relating the unknown data with the classified and trained data. In future the projection phase and classification phase could be combined. Also, using different dataset the training and testing of system can be tested.

ACKNOWLEDGMENT

I am very grateful to my guide Mrs. D. A. Phalke for her insightful and detailed comments and suggestions, which have helped me to improve the paper significantly. I take this opportunity to convey my sincere thanks to our beloved Head of the Department Dr. Mrs. Neeta Deshpande for her continual support and encouragement throughout the course of this paper writing. I express my profound thanks to our P.G. Coordinator Mrs. V.L. Kolhe for their advice and valuable guidance that helped me in making this paper interesting and successful one.

REFERENCES

[1] A. Hogenboom, D. Bal, F. Frasincar, M. Bal, F. de Jong, and U. Kaymak, “Exploiting emoticons in sentiment analysis,” in Proceedings of the 28th Annual ACM Symposium on Applied Computing, pp. 703–710, ACM, 2013.
[2] S.-H. Cho and H.-B. Kang, “Statistical text analysis and sentiment classification in social media,” in 2012 IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp. 1112–1117, IEEE, 2012.
[3] B. Pang and L. Lee, “Opinion mining and sentiment analysis,” Foun-dations and trends in information retrieval, vol. 2, no. 1-2, pp. 1–135, 2008.
[4] A. C. E. Lima, L. N. de Castro, and J. M. Corchado, “A polarity analysis framework for twitter messages,” Applied Mathematics and Computation, vol. 270, pp. 756–767, 2015.
[5] J. Z. Ferreira, J. Rodrigues, M. Cristo, and D. F. de Oliveira, “Multi-entity polarity analysis in financial documents,” in Proceedings of the 20th Brazilian Symposium on Multimedia and the Web, pp. 115–122, ACM, 2014.
[6] M. Ptaszynski, J. Maciejewski, P. Dybala, R. Rzepka, and K. Araki, “Cao: A fully automatic emoticon analysis system based on theory of kinesics,” IEEE Transactions on Affective Computing, vol. 1, no. 1, pp. 46–59, 2010.

@IJAERD-2017, All rights Reserved

[7] W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: A survey," Ain Shams Engineering Journal, vol. 5, no. 4, pp. 1093–1113, 2014.
[8] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, "Lexicon-based methods for sentiment analysis," Computational linguistics, vol. 37, no. 2, pp. 267–307, 2011.
[9] P. Gonc¸alves, M. Araujo,´ F. Benevenuto, and M. Cha, "Comparing and combining sentiment analysis methods," in Proceedings of the first ACM conference on Online social networks, pp. 27–38, ACM, 2013.
[10] N. Lalithamani, L. S. Thati, and R. Adhikesavan, "Sentence-level senti-ment polarity calculation for customer reviews by considering complex sentential structures," IJRET: International Journal of Research in Engi-neering and Technology, vol. 3, 2014.
[11] J. Leskovec, "Social media analytics: tracking, modeling and predicting the flow of information through networks," in Proceedings of the 20th international conference companion on World wide web, pp. 277–278, ACM, 2011.
[12] E. Kouloumpis, T. Wilson, and J. D. Moore, "Twitter sentiment analysis: The good the bad and the omg!," Icwsm, vol. 11, pp. 538–541, 2011.
[13] F. Jiang, Y.-Q. Liu, H.-B. Luan, J.-S. Sun, X. Zhu, M. Zhang, and S.-P. Ma, "Microblog sentiment analysis with emoticon space model," Journal of Computer Science and Technology, vol. 30, no. 5, pp. 1120–1129, 2015.
[14] A. H. Huang, D. C. Yen, and X. Zhang, "Exploring the potential effects of emoticons," Information & Management, vol. 45, no. 7, pp. 466–473, 2008.
[15] E. A. Hassan, N. El Gayar, and M. G. Moustafa, "Emotions analysis of speech for call classification," in 2010 10th International Conference on Intelligent Systems Design and Applications, pp. 242–247, IEEE, 2010.
[16] K.-L. Liu, W.-J. Li, and M. Guo, "Emoticon smoothed language models for twitter sentiment analysis.," in AAAI, 2012.
[17] M. Yuasa, K. Saito, and N. Mukawa, "Emoticons convey emotions without cognition of faces: an fmri study," in CHI'06 Extended Abstracts on Human Factors in Computing Systems, pp. 1565–1570, ACM, 2006.
[18] D. R. Recupero, V. Presutti, S. Consoli, A. Gangemi, and A. G. Nuzzolese, "Sentilo: frame-based sentiment analysis," Cognitive Com-putation, vol. 7, no. 2, pp. 211–225, 2015.
[19] H. Saif, Y. He, and H. Alani, "Semantic sentiment analysis of twitter," in International Semantic Web Conference, pp. 508–524, Springer, 2012.
[20] X. Hu, L. Tang, J. Tang, and H. Liu, "Exploiting social relations for sentiment analysis in microblogging," in Proceedings of the sixth ACM international conference on Web search and data mining, pp. 537–546, ACM, 2013.
[21] U. Orhan, M. Hekim, and M. Ozer, "Eeg signals classification using the k-means clustering and a multilayer perceptron neural network model," Expert Systems with Applications, vol. 38, no. 10, pp. 13475–13481, 2011.
[22] "Multi layer perceptron." Available:https://en.wikipedia.org/wiki/ Multilayer perceptron, [accessed January 4, 2017].
[23] "Google." Available:https://www.google.co.in/?gws rd=ssl#q=twitter, [accessed January 4, 2017].
[24] G. Paltoglou and M. Thelwall, "Twitter, myspace, digg: Unsupervised sentiment analysis in social media," ACM Transactions on Intelligent Systems and Technology (TIST), vol. 3, no. 4, p. 66, 2012.

@IJAERD-2017, All rights Reserved