Volume: 5 Issue: XI Month of publication: November 2017

DOI: http://doi.org/10.22214/ijraset.2017.11107

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor:6.887
Volume 5 Issue X1, November 2017- Available at www.ijraset.com

Enhanced Classification Framework

Mr. Chandrashekhar Bora[1], Mr. Amit Mehra[2], Mr. Yash Kaul[3], Mrs. Harshali Patil[4], Mr. Rishi shah[5]
[1]Dept. of Computer Engineering, Thakur College of Engineering and Technology, Mumbai, India

Abstract: With the rise of social networking era, Micro blogging sites have millions of people sharing their thoughts daily because of its characteristic short and simple manner of expression. We propose and investigate a model to mine the sentiment from a popular real-time micro blogging service, Social networking sites, where users post real time reactions to and opinions about "everything". In this paper, we expound a hybrid approach using both quantity based and dictionary based methods to determine the semantic alignment of the opinion words in tweets. A case study is presented to illustrate the use and effectiveness of the proposed system.

Index Terms: Component, social networking sites, opinion mining, analysis, insert.

I. INTRODUCTION

Ongoing increase in wide-area network connectivity promise vastly augmented opportunities for collaboration and resource sharing. Now-a-days, various social networking sites like Social networking sites1, Facebook2, MySpace3, YouTube4 have gained so much popularity and we cannot

ignore them. They have become one of the most vital applications of Web 2.0 [1]. They allow people to build connection systems with other people in an easy and opportune way and allow them to share numerous kinds of information and to use a set of facilities like picture sharing, blogs, wikis etc.

It is apparent that the advent of these real-time information networking sites like Social networking sites have reproduced the creation of an unequaled public collection of opinions about every global entity that is of interest. Although Social networking sites may establishment for an excellent channel for opinion creation and presentation, it poses newer and different challenges and the process is incomplete without adept tools for analyzing those opinions to accelerate their consumption.

More recently, there have been several research projects that apply sentiment analysis to Social networking sites corpora in order to extract general public opinion regarding political issues. Due to the increase of intimidating and negative communication over social networking sites like Facebook and Social networking sites, recently the Government of India tried to allay concerns over restriction of these sites where Web users continued to speak out against any proposed restriction on posting of content. As reported in one of the Indian national newspaper "Union Minister for Communications and Information Minister, Kapil Sibal, proposed content screening & restricted of social networks like Social networking sites and Facebook". Originated by this the exploration carried out by us was to use sentiment analysis to device the public temperament and detect any rising hostile or negative feeling on social medias. Even though, we firmly believe that restricted is not right path to follow, this recent trend for research for sentiment mining in social networking sites can be utilized and extended for a range of practical applications that range from applications in business (marketing intelligence; product and service bench marking and improvement), applications as subcomponent technology (recommender systems; summarization; question answering) to applications in politics. This motivated us to propose a model which retrieves tweets on a certain topic through the Social networking sites API and computes the sentiment orientation/score of each tweet.

II. LITERATURE SURVEY

Applying sentiment analysis on Social networking sites is the upcoming trend with researchers recognizing the scientific trials and its potential applications. The challenges unique to this problem area are largely attributed to the dominantly informal tone of the micro blogging. Pak and Paroubek

justification the use micro blogging and more particularly Social networking sites as a quantity for sentiment analysis. They cited:

A. Micro blogging platforms are used by different people to express their opinion about different topics, thus it is a valuable source of people's opinions.

B. Social networking sites contains an enormous number of text posts and it grows every day. The collected quantity can be arbitrarily large.

C. Social networking sites's audience varies from regular users to celebrities, company representatives, politicians, and even country presidents. Therefore, it is possible to collect text posts of users from different social and interests groups.

D. Social networking sites's audience is represented by users from many countries.

Parikh and Movassate implemented two Naive Bayes unigram models, a Naive Bayes bigram model and a Maximum Entropy model to classify tweets. They found that the Naive Bayes classifiers worked much better than the Maximum Entropy model could. Go et al. proposed a solution by using distant management, in which their training data consisted of tweets with emoticons. This approach was initially introduced by Read. The emoticons served as noisy labels. They build models using Naive Bayes, MaxEnt and Support Vector Machines (SVM). Their feature space consisted of unigrams, bigrams and POS. The reported that SVM outstripped other models and that unigram were more effective as features. Pak and Paroubek have done similar work but classify the tweets as impartial, positive and negative. In order to collect a quantity of neutral posts, they recovered text messages from Social networking sites accounts of popular newspapers and magazine, such as "New York Times", "Washington Posts" etc. Their classifier is based on the

Multinomial Naïve Bayes classifier that uses N-gram and POS-tags as features. Barbosa et al. too classified tweets as objective or individual and then the individual tweets were classified as positive or negative. The feature space used included features of tweets like re tweet, hash tags, link, punctuation and exclamation marks in conjunction with features like prior divergence of words and POS of words.

Mining for entity opinions in Social networking sites, Kalpana and Simran used a dataset of tweets spanning two months starting from June 2009. The dataset has roughly 60 million tweets. The entity was extracted using the Stanford NER, user tags and URLs were used to augment the entities found. A quantity of 200,000 product reviews that had been labeled as positive or unwelcome was used to train the model. Using this quantity the model computed the probability that a given unigram or bigram was being used in a positive context and the probability that it was being used in a negative context. Bifet and Frank used Social networking sites streaming data provided by Firehouse, which gave all messages from every user in real-time. They investigated with three fast incremental methods that were well-suited to deal with data streams: multinomial naive Bayes, stochastic gradient descent, and the Hoeffding tree. They concluded that SGD-based model, used with an appropriate learning rate was the best.

There are two basic procedures to detect sentiments from text. They are Symbolic techniques and Machine Learning techniques [3]. The next two sections deal with these techniques.

Much of the research in unsupervised sentiment organisation using symbolic techniques makes use of available lexical resources. Turney [4] used bag-of-words approach for sentiment analysis. In this approach, relationships between the individual words are not considered and a document is represented as a mere collection of words. To determine the overall sentiment, sentiment of every word is determined and this value is combined with some aggregation functions. He unyielding the division of a review based on the average semantic orientation of tuples extracted from the review where tuples are phrases having adjectives or adverbs. He determined the semantic orientation of tuples using the search engine Altavista.

Kamps et al. [5] used the lexical database WordNet [6] to determine the emotional content of a word along different dimensions. They developed a distance metric on WordNet and determined the semantic orientation of adjectives. WordNet database consists of words connected by synonym relations. Baroni et al. [7] developed a system using expression space model formalism that overcomes the difficulty in lexical substitution task. It represents the local context of a word along with its overall distribution. Balahur et al. [8] introduced EmotiNet, a conceptual representation of text that stores the structure and the semantics of real events for a specific domain. EmotiNet used the concept of Finite State Automata to identify the emotional responses activated by actions. One of the parikhcontributors of SemEval 2007 Task No. 14 [9] used bristly-grained and fine-grained approaches to identify sentiments in news headlines. In coarse-grained method, they achieved binary classification of emotions whereas in fine-grained approach, they classified emotions into different levels. Knowledge-based approach is found to be difficult due to the necessity of a huge lexical database. Social network generates huge amount of data every second, which is knowingly larger than the size of accessible lexical databases. Therefore, sentiment analysis often becomes hard and flawed.

III. PROBLEM DEFINITION

Applying sentiment analysis on Social networking sites is the upcoming trend with researchers recognizing the scientific trials and its potential applications. The challenges unique to this problem area are largely attributed to the dominantly informal tone of the micro blogging. Pak and Paroubek rational the use micro blogging and more particularly Social networking sites as a quantity for sentiment analysis. They cited:

E. Micro blogging platforms are used by different people to express their view about different topics, thus it is a appreciated source of people's opinions.

F. Social networking sites contain an enormous number of text posts and it grows every day. The collected quantity can be subjectively large.

G. Social networking sites's audience varies from regular users to celebrities, company representatives, politicians, and even country presidents. Therefore, it is possible to collect text posts of users from different social and interests groups.

H. Social networking sites's audience is represented by users from many countries.

Parikh and Movassate realized two Naive Bayes unigram models, a Naive Bayes bigram prototypical and a Maximum Entropy model to classify tweets. They found that the Naive Bayes classifiers worked much better than the Maximum Entropy model could. Go et al. planned a solution by using unfriendly administration, in which their training data consisted of tweets with emoticons. This approach was initially introduced by Read. The emoticons served as noisy labels. They build models using Naive Bayes, MaxEnt and Support Vector Machines (SVM).

Their feature space consisted of unigrams, bigrams and POS. The testified that SVM outperformed other models and that unigram were more effective as features. Pak and Paroubek have done similar work but categorize the tweets as objective, optimistic and negative. In order to collect a amount of objective posts, they retrieved text messages from Social networking sites accounts of popular newspapers and magazine, such as "New York Times", "Washington Posts" etc. Their classifier is based on the multinomial Naïve Bayes classifier that uses N-gram and POS-tags as features. Barbosa et al. too classified tweets as objective or subjective and then the subjective tweets were classified as positive or negative.

The feature space used included features of tweets like re tweet, hash tags, link, punctuation and exclamation marks in conjunction with features like prior polarity of words and POS of words.

Mining for entity opinions in Social networking sites, Batra and Rao used a dataset of tweets spanning two months starting from June 2009. The dataset has roughly 60 million tweets. The entity was extracted using the Stanford NER, user tags and URLs were used to augment the entities found.

A quantity of 200,000 product reviews that had been labeled as positive or negative was used to train the model. Using this quantity the model computed the probability that a given unigram or bigram was being used in a positive context and the probability that it was being used in a negative context. Bifet and Frank used Social networking sites streaming data provided by Firehouse, which gave all messages from every user in real-time. They experimented with three fast incremental methods that were well-suited to deal with data streams: multinomial naive Bayes, stochastic gradient descent, and the Hoeffding tree. They concluded that SGD-based model, used with an appropriate learning rate was the best.

IV. PROPOSED WORK

A. Pre-processing of Tweets

We prepare the contract file that contains opinion needles, namely the adjective, adverb and verb along with emoticons (we have taken a model set of emoticons and manually assigned opinion strength to them). Also we identify some emotion intensifiers, namely, the percentage of the tweet in Caps, the length of repeated sequences & the number of exclamation marks, amongst others. Thus, we pre-process all the tweets as follows:

1) Remove all URLs (e.g. www.example.com), hash tags (e.g. #topic), targets (@username), special Social networking sites words ("e.g. RT").

2) Compute the percentage of the tweet in Caps.

3) Correct spellings; A sequence of repeated characters is tagged by a weight. We do this to differentiate between the regular usage and emphasized usage of a word.

4) Replace all the emoticons with their sentiment polarity (Table 1).

5) Remove all punctuations after counting the number of exclamation marks.

6) Using a POS tagger, the NL Processor linguistic Parser [15], we tag the adjectives, verbs and adverbs.

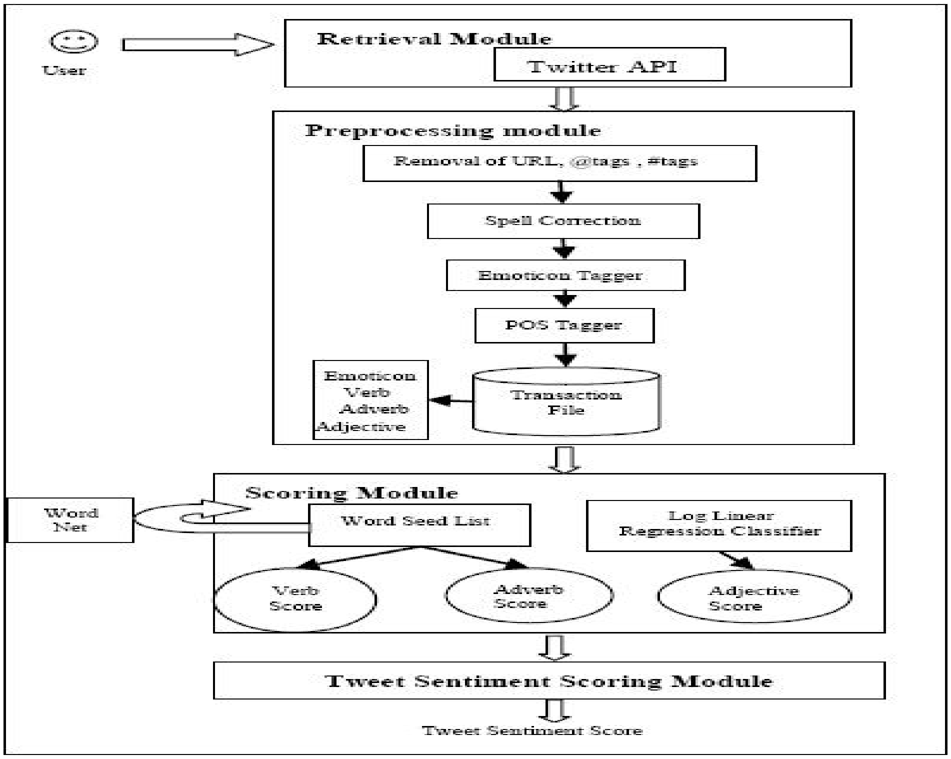
B. Scoring Module

The next phase is to find the semantic score of the opinion carriers i.e. the adjectives, verbs and adverbs. As cited previously, in our tactic we use quantity based method to find the semantic orientation of adjectives and the dictionary-based method to find the semantic alignment of verbs and adverbs.

C. Tweet Sentiment Scoring

As adverbs qualify adjectives and verbs, we group the corresponding adverb and adjective together and call it the adjective group; likewise we group the consistent verb and adverb together and call it the verb group. The adjective group strength is computed by the product of adjective score (adji) and adverb (advi) score, and the verb group asset as the product of verb score (vbi) and adverb score (advi). From time to time, there is no adverb in the opinion group, so the S (adv) is set as a default value 0.5 To compute the overall sentiment of the tweet, we average the asset of all opinion indicators like emoticons, exclamation marks, capitalization, word emphasis, adjective group and verb group as shown below: system architecture

$$S(T)=\frac{(1+(P_c+\log(Ns)+\log(N_X))/3)}{|OI(R)|}*\sum_{i=1}^{|OI(R)|} S(AG_i)+S(VG_i)+N_{ei}*S(E_i) \quad (3)$$



As adverbs qualify adjectives and verbs, we group the corresponding adverb and adjective together and call it the adjective group; similarly we group the corresponding verb and adverb together and call it the verb group. The adjective group strength is computed by the product of adjective score (adji) and adverb (advi) score, and the verb group strength as the product of verb score (vbi) and adverb score (advi). Sometimes, there is no adverb in the opinion group, so the S (adv) is set as a default value 0.5

To compute the overall sentiment of the tweet, we average the strength of all opinion indicators like emoticons, Exclamation marks, capitalization, word emphasis

$$S(T)=\frac{(1+(P_c+\log(Ns)+\log(N_X))/3)}{|OI(R)|}*\sum_{i=1}^{|OI(R)|} S(AG_i)+S(VG_i)+N_{ei}*S(E_i)$$

Where, |OI(R)| denotes the size of the set of opinion groups and emoticons extracted from the tweet, Pc denotes fraction of tweet in caps, Ns denotes the count of repeated letters, Nx denotes the count of exclamation marks, S (AGi) denotes score of the ith adjective group, S (VGi) denotes the score of the ith verb group, S (Ei) denotes the score of the ith emoticon Nei denotes the count of the ith emoticon. Pc, Ns and Nx represent emphasis on the sentiment to be conveyed so they can be collectively called sentiment intensifiers. If the score of the tweet is more than 1 or less than -1, the score is taken as 1 or -1 respectively. To clearly illustrate the effectiveness of the proposed method, a case study is presented with a sample tweet: <tweet>="@kirinv I hate revision, it's BOOOORING!!! I am totally unprepared for my exam tomorrow :(:(Things are not good...#exams"

A. The pre-processing of Tweet

A transaction file is created which contains the pre-processed opinion indicators.

1) *Extracting Opinion Intensifiers*: The opinion intensifiers are computed for the tweet as follows.

a) Fraction of tweet in caps: There are a total of 18 words in the sentence out of which one is in all caps. Therefore, Pc=1/18=0.055

b) Length of repeated sequence, Ns=3

c) Number of Exclamation marks, Nx=3

2) *Extracting Opinion Words:* After the tweet is preprocessed, it is tagged using a POS tagger and the adjective and verb groups are extracted.

The list of Adjective Groups extracted**:**

AG1=totally unprepared

AG2=not good

AG3=boring

The list of Verb Groups extracted:

VG1=hate

The list of Emoticons extracted:

E1 = :(

Ne1 = 2

B. Scoring Module Now that we have our adjective group and verb group, we have to find their semantic orientation. Calculation is based on ke

1) Score of Adjective Group

S (AG1) = S (totally unprepared) =0.8*-0.5 == -0.4

S (AG2) = S (not good) =-0.8*1= -0.8

S (AG3) = S (boring) = 0.5*-0.25 = -0.125

5.2.2 Score of Verb Group

S (VG1) = S (hate) = 0.5*-0.75 = 0.375

C. Tweet Sentiment Scoring

By the formula distinct in equation 3 we can compute the sentiment strength of the tweet as follows:

( ) * ( )

5

1.33 S(T)

5

i 1 *i i ei i* □ □ □ □*S(AG)*□ *S VG* □ *N S E* □

((0.4) (0.8) (0.125) (0.5) 2(0.5))

5

(1.33)

0.751

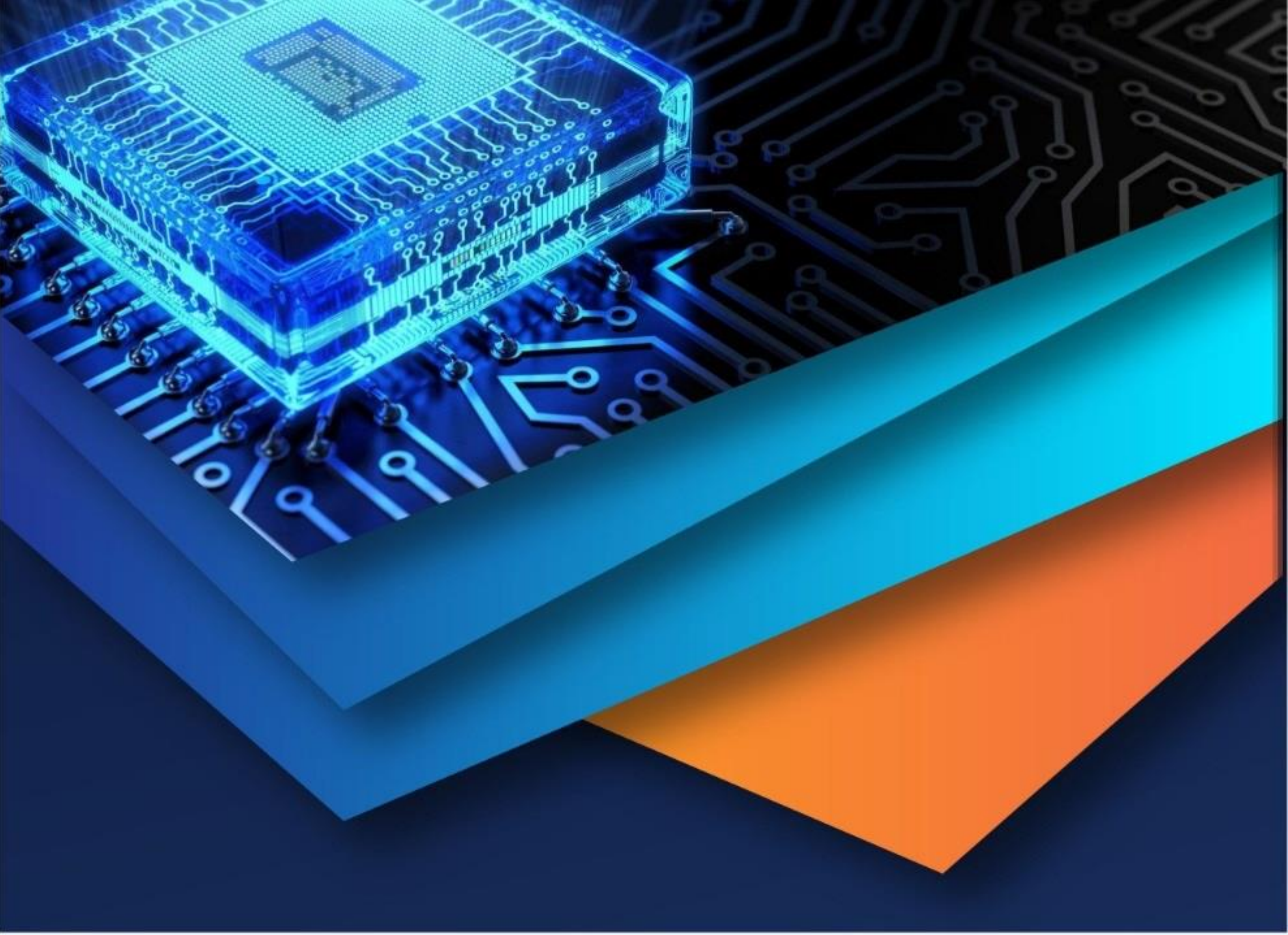
As we have got a negative value, we can safely classify the tweet as negative. We applied our approach to a sample set of 10 tweets. The semantic analysis results obtained are depicted in table 3 below.

I. CONCLUSIONS

The work presented in this paper specifies a novel approach for sentiment analysis on Social networking sites data. To reveal the sentiment, we extracted the opinion words (a amalgamation of the adjectives along with the verbs and adverbs) in the tweets. The quantity-founded method was used to find the semantic positioning of adjectives and the dictionary-based method to find the semantic orientation of verbs and adverbs. The global tweet sentiment was then calculated using a linear equation which assimilated emotion intensifiers too. This work is probing in nature and the prototype evaluated is a preliminary prototype. The initial results show that it is a rousing technique.

REFERENCES

[1] L. Colazzo, A. Molinari and N. Villa. “Collaboration vs. Participation: the Role of Virtual Communities in a Web 2.0 world”, International Conference on Education Technology and Computer, 2009, pp.321-325.
[2] nlp.stanford.edu/courses/cs224n/2011/reports/patlai.pdf
[3] National Daily, Economic Times: articles.economictimes.indiatimes.com › Collections › Facebook
[4] K. Dave, S. Lawrence, and D.M. Pennock. “Mining the peanut gallery: Opinion extraction and semantic classification of product reviews”. In Proceedings of the 12th International Conference on World Wide Web (WWW), 2003, pp. 519–528.
[5] A. Pak and P. Paroubek. “Social networking sites as a Quantity for Sentiment Analysis and Opinion Mining”. In Proceedings of the Seventh Conference on International Language Resources and Evaluation, 2010, pp.1320–1326.
[6] R. Parikh and M. Movassate, “Sentiment Analysis of User- Generated Social networking sites Updates using Various Classification Techniques”, CS224N Final Report, 2009
[7] A. Go, R. Bhayani, L.Huang. “Social networking sites Sentiment Classification Using Distant Supervision”. Stanford University, Technical Paper ,2009
[8] J. Read. “Using emoticons to reduce dependency in machine learning techniques for sentiment classification”. In Proceedings of ACL-05, 43nd Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2005
[9] L. Barbosa, J. Feng. “Robust Sentiment Detection on Social networking sites from Biased and Noisy Data”. COLING 2010: Poster Volume, pp. 36-44.
[10] S. Batra and D. Rao, ”Entity Based Sentiment Analysis on Social networking sites”, Stanford University, 2010.



10.22214/IJRASET

INDEX COPERNICUS
45.98

IMPACT FACTOR:
7.129

TOGETHER WE REACH THE GOAL
IMPACT FACTOR:
7.429