Volume No:4, Issue No:9 (February-2019) ISSN No : 2454-423X (Online)
International Journal of Research in Advanced Computer Science Engineering
A Peer Reviewed Open Access International Journal
www.ijracse.com

Analysis on Tweets using Opinion Mining

Doppalapudi Haavani
Department of Computer Science and Engineering,
Gayatri Vidya Parishad College of Engineering (Autonomous),
Visakhapatnam, A.P - 530 048, India.

Mrs.G.Vani
Department of Computer Science and Engineering,
Gayatri Vidya Parishad College of Engineering (Autonomous),
Visakhapatnam, A.P - 530 048, India.

ABSTRACT

Twitter is a web application built to find out what is happening at any instance through its micro blogging feature, anywhere in the world. These messages contains positive or negative values based on sentiment analysis with respect to query. This is good for designing accurate and efficient sentiment classification system and finally using for polarity. Techniques used are Machine learning Approach (ML), Lexicon Based Approach (LB) and Hybrid Approach. Due to this large amount of usage we hope to achieve a reflection of public sentiment by analysing the sentiments expressed in the tweets. Analysing the public sentiment is important for many applications such as firms trying to find out the response of their products in the market, predicting political elections and predicting socioeconomic phenomena like stock exchange. The aim of this project is to develop a functional classifier for accurate and automatic sentiment classification of an unknown tweet stream.

Introduction

This project of analyzing sentiments of tweets comes under the domain of "Pattern Classification" and "Data Mining". Both of these terms are very closely related and intertwined, and they can be formally defined as the process of discovering "useful" patterns in large set of data, either automatically (unsupervised) or semi-automatically (supervised). The project would heavily rely on techniques of "Natural Language Processing" in extracting significant patterns and features from the large data set of tweets and on "Machine Learning" techniques for accurately classifying individual unlabelled data samples (tweets) according to whichever pattern model best describes them.

The features that can be used for modeling patterns and classification can be divided into two main groups: formal language based and informal blogging based. Language based features are those that deal with formal linguistics and include prior sentiment polarity of individual words and phrases, and parts of speech tagging of the sentence. Prior sentiment polarity means that some words and phrases have a natural innate tendency for expressing particular and specific sentiments in general. For example the word "excellent" has a strong positive connotation while the word "evil" possesses a strong negative connotation. So whenever a word with positive connotation is used in a sentence, chances are that the entire sentence would be expressing a positive sentiment. Parts of Speech tagging, on the other hand, is a syntactical approach to the problem. It means to automatically identify which part of speech each individual word of a sentence belongs to: noun, pronoun, adverb, adjective, verb, interjection, etc. Patterns can be extracted from analyzing the frequency distribution of these parts of speech (ether individually or collectively with some other part of speech) in a particular class of labeled tweets. Twitter based features are more informal and relate with how people express themselves on online social platforms and compress their sentiments in the limited space of 140 characters offered by twitter. They include twitter hashtags,

Cite this article as: Doppalapudi Haavani & Mrs.G.Vani, "Analysis on Tweets using Opinion Mining", International Journal of Research in Advanced Computer Science Engineering, Volume 4 Issue 9, 2019, Page 1-7.

Volume No:4, Issue No:9 (February-2019) ISSN No : 2454-423X (Online)
International Journal of Research in Advanced Computer Science Engineering
A Peer Reviewed Open Access International Journal
www.ijracse.com

retweets, word capitalization, word lengthenin,], question marks, presence of url in tweets, exclamation marks, internet emoticons and internet shorthand/slangs.

Classification techniques can also be divided into a two categories: Supervised vs. unsupervised and non-adaptive vs. adaptive/reinforcement techniques. Supervised approach is when we have pre-labeled data samples available and we use them to train our classifier. Training the classifier means to use the pre-labeled to extract features that best model the patterns and differences between each of the individual classes, and then classifying an unlabeled data sample according to whichever pattern best describes it. For example if we come up with a highly simplified model that neutral tweets contain 0.3 exclamation marks per tweet on average while sentiment-bearing tweets contain 0.8, and if the tweet we have to classify does contain 1 exclamation mark then (ignoring all other possible features) the tweet would be classified as subjective, since 1 exclamation mark is closer to the model of 0.8 exclamation marks. Unsupervised classification is when we do not have any labeled data for training. In addition to this adaptive classification techniques deal with feedback from the environment. In our case feedback from the environment can be in form of a human telling the classifier whether it has done a good or poor job in classifying a particular tweet and the classifier needs to learn from this feedback. There are two further types of adaptive techniques: Passive and active. Passive techniques are the ones which use the feedback only to learn about the environment (in this case this could mean improving our models for tweets belonging to each of the three classes) but not using this improved learning in our current classification algorithm, while the active approach continuously keeps changing its classification algorithm according to what it learns at real-time.

Existing System

Pre processing:

- Removing all unwanted tweets
- Apply stop word filtering

Feature extraction:

- Emoticons: Extracted both positive and negative emoticons based on their frequency.
- Opinion lexicon: Dictionaries of positive and negative sentiment words.
- Punctuations
- Unigrams: Single words commonly used as features. These are used in tweets dataset.

Proposed System

Pre processing:

- Removing all unwanted tweets.
- Based on the API server to LSOW dataset data is converted into code and to analyze each and every tweet.

Feature extraction:

- By using bigram along with unigram we can enhance the performance of the tweets.
- After getting the tweets the data is automatically analyzed by using pie-chat and its shows the percentage of positive, negative and neutral.

IMPLEMENTATION

We will first present our results for the objective / subjective and positive / negative classifications. These results act as the first step of our classification approach. We only use the short-listed features for both of these results. This means that for the objective / subjective classification we have 5 features and for positive / negative classification we have 3 features. For both of these results we use the Naïve Bayes classification algorithm, because that is the algorithm we are employing in our actual classification approach at the first step. Furthermore all the figures reported are the result of 10-fold cross validation.

In addition to the above information, we make a condition while reporting the results of polarity classification (which differentiates between positive and negative classes) that only subjective labelled tweets are used to calculate these results. However, in case of final

Volume No:4, Issue No:9 (February-2019) ISSN No : 2454-423X (Online)
International Journal of Research in Advanced Computer Science Engineering
A Peer Reviewed Open Access International Journal
www.ijracse.com

classification approach, any such condition is removed and basically both objectivity and polarity classifications are applied to all tweets regardless of whether they are labelled objective or subjective.

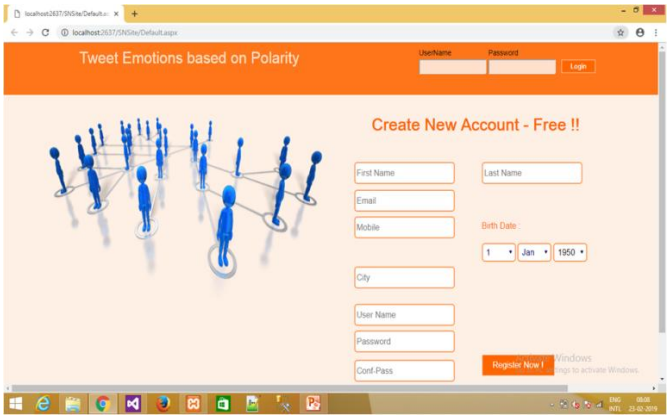
If we compare these results to those provided by Wilson et al. [16] (results are displayed in Table 2 and Table 3 of this report) we see that although the accuracy of neutral class falls from 82.1% to 73% if we use our classification instead of theirs. However, for all other classes we report significantly greater results. Although the results presented by Wilson et al. are not from Twitter data they are of phrase level sentiment analysis which is very close in concept to Twitter sentiment analysis.

Next we will compare our results with those presented by Go et al. [2]. The results presented by this paper are as follows:
If we compare these results to ours, we see that they are more or less similar. However, we arrive at comparable results with just 10 features and about 9,000 training data. In contrast to this, they used about 1.6 million noisy labels. Their labels were noisy in the sense that the tweets that contained positive emoticons were labelled as positive, while those with negative emoticons were labelled negative. The rest of the tweets (which did not contain any emoticon) were discarded from the data set. So in this way they hoped to achieve high results without human labelling but at the cost of using humongous large number amount of data set.

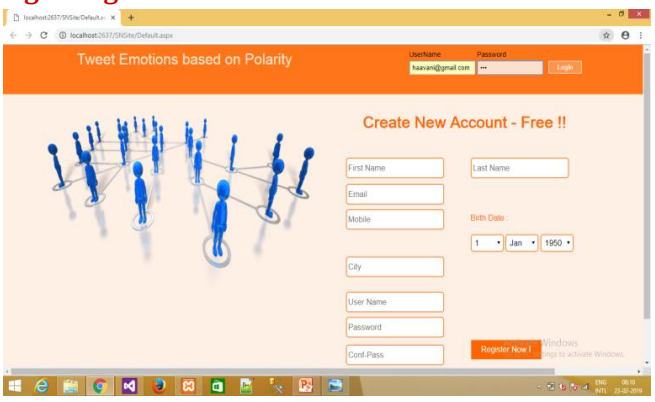
In comparison with these results, Koulompietal. [7] reports average F-measure of 68%. However when they include another portion of their data into their classification process (which they call the HASH data), their average F-measure drops to 65%. In contrast to this we achieve average F-measure of more than 70% which shows better performance than either of these results. Moreover we make use of only 8 features and 9,000 labelled tweets, while their process involves about 15 features in total and more than 220,000 tweets in their training set. Our unigram word models are also simpler than theirs, because they incorporate negation into their word models. However like in the case of (1-9) their tweets are not labelled by humans, but rather undergo noisy labelling in two ways: labels acquired from positive and negative emoticons and hashtags.

SCREEN SHOTS

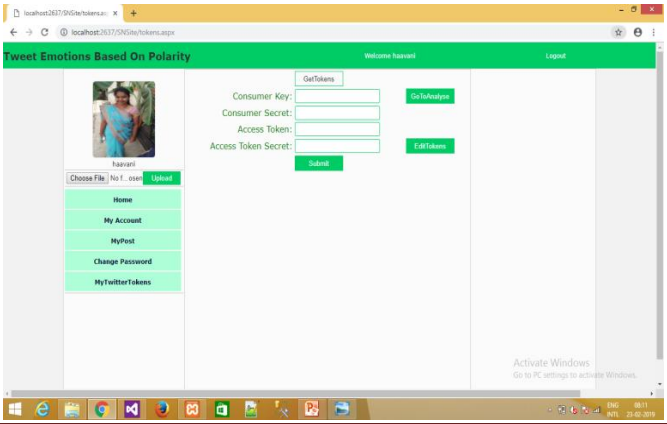
Registration Page



Login Page



Home page



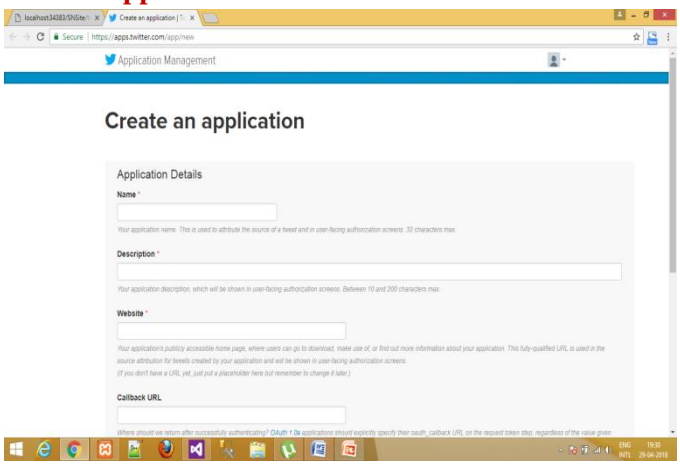
Volume No:4, Issue No:9 (February-2019) ISSN No : 2454-423X (Online)

International Journal of Research in Advanced Computer Science Engineering

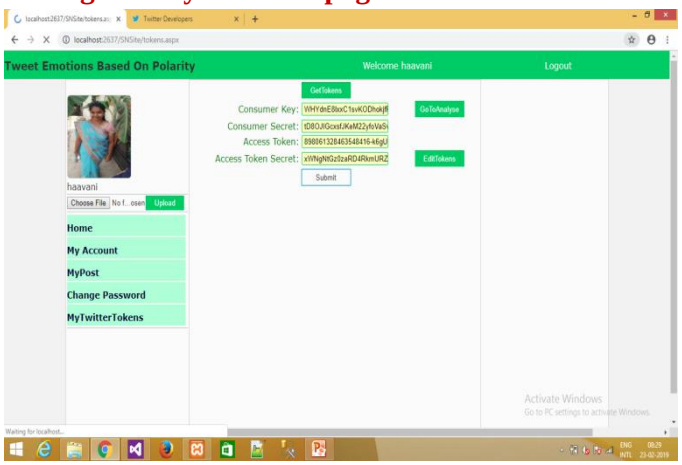
A Peer Reviewed Open Access International Journal

www.ijracse.com

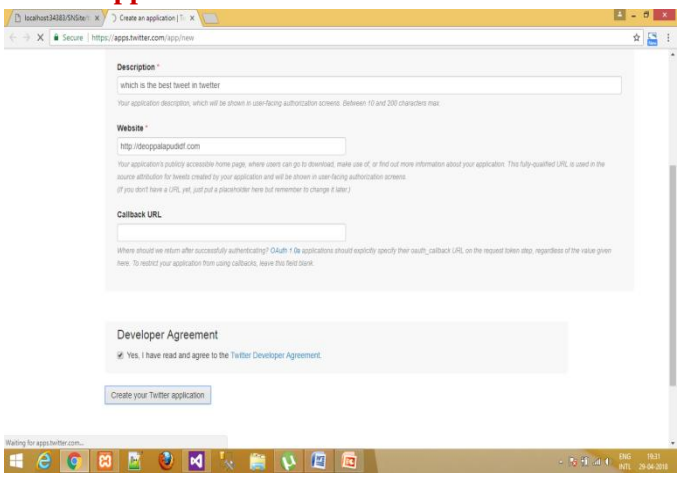
Twitter Application



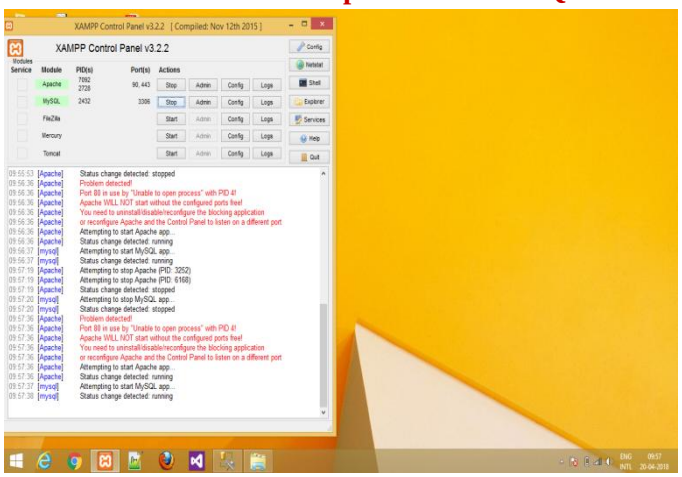
Enter given keys in home page



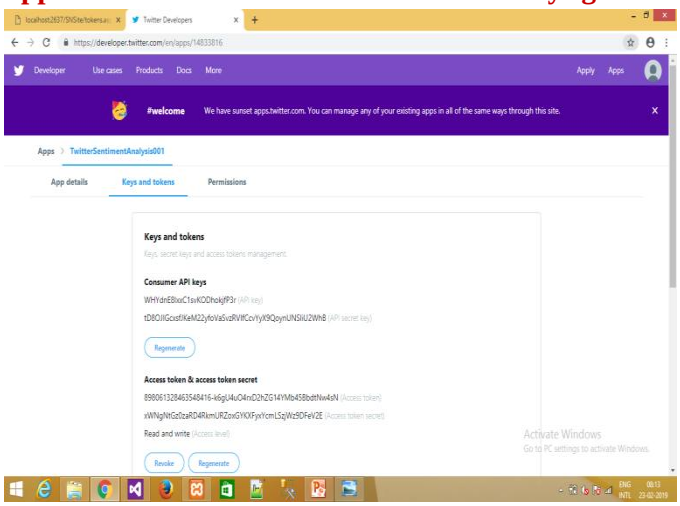
Create Application form



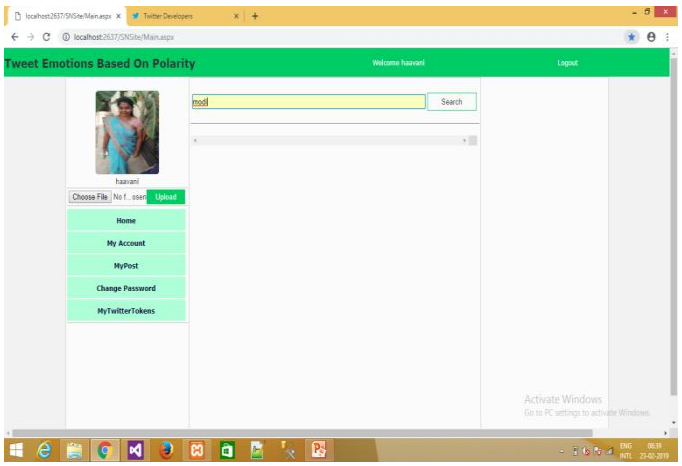
Start the XAMPP control apache and MYSQL



Application have been created and token keys given

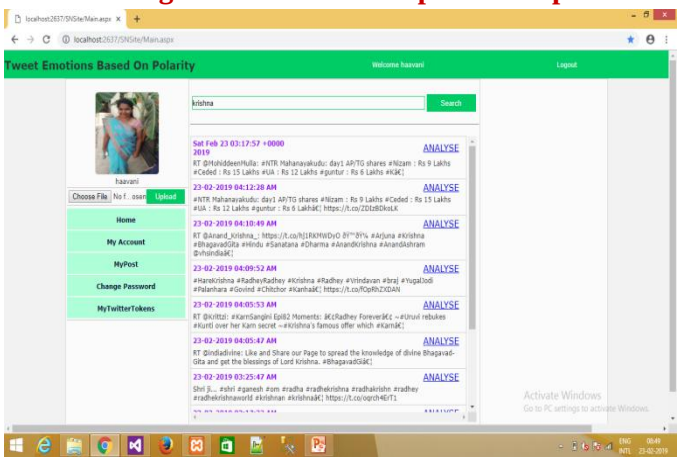


If anyone want to search the profile who has a twitter account will enter into the search bar

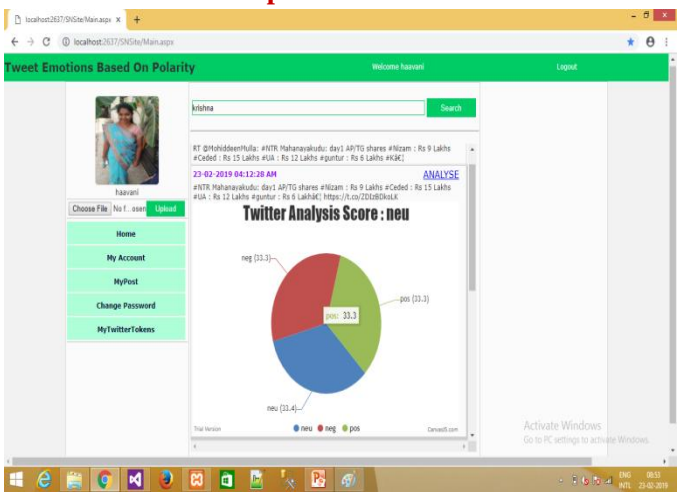


Volume No:4, Issue No:9 (February-2019) ISSN No : 2454-423X (Online)
International Journal of Research in Advanced Computer Science Engineering
A Peer Reviewed Open Access International Journal
www.ijracse.com

Now we will get the details of the particular person



Now click on the analysis button the details are shown in the form of pie-chart



CONCLUSION AND FUTURE RECOMMENDATIONS

The task of sentiment analysis, especially in the domain of micro-bloging, is still in the developing stage and far from complete. So we propose a couple of ideas which we feel are worth exploring in the future and may result in further improved performance.

Right now we have worked with only the very simplest unigram models; we can improve those models by adding extra information like closeness of the word with a negation word. We could specify a window prior to the word (a window could for example be of 2 or 3 words) under consideration and the effect of negation may be incorporated into the model if it lies within that window. The closer the negation word is to the unigram word whose prior polarity is to be calculated, the more it should affect the polarity. For example if the negation is right next to the word, it may simply reverse the polarity of that word and farther the negation is from the word the more minimized ifs effect should be.

Apart from this, we are currently only focusing on unigrams and the effect of bigrams and trigrams may be explored. As reported in the literature review section when bigrams are used along with unigrams this usually enhances performance.

However for bigrams and trigrams to be an effective feature we need a much more labeled data set than our meager 9,000 tweets.

Right now we are exploring Parts of Speech separate from the unigram models, we can try to incorporate POS information within our unigram models in future. So say instead of calculating a single probability for each word like P(word | obj) we could instead have multiple probabilities for each according to the Part of Speech the word belongs to. For example we may have P(word | obj, verb), P(word | obj, noun)and P(word | obj, adjective). Pang et al. [5] used a somewhat similar approach andclaims that appending POS information for every unigram results in no significant change in performance (with Naive Bayes performing slightly better and having a slight decrease in performance), while there is a significant decrease in accuracy if only adjective unigrams are used as features. However these results are for classification of reviews and may be verified for sentiment analysis on micro blogging websites like Twitter.

One more feature we that is worth exploring is whether the information about relative position of word in a tweet has any effect on the performance of the classifier. Although Pang et al. explored a similar feature and

Volume No:4, Issue No:9 (February-2019) ISSN No : 2454-423X (Online)
International Journal of Research in Advanced Computer Science Engineering
A Peer Reviewed Open Access International Journal
www.ijracse.com

reported negative results, their results were based on reviews which are very different from tweets and they worked on an extremely simple model.

In this research we are focussing on general sentiment analysis. There is potential of work in the field of sentiment analysis with partially known context. For example we noticed that users generally use our website for specific types of keywords which can divided into a couple of distinct classes, namely: politics/politicians, celebrities, products/brands, sports/sportsmen, media/movies/music. So we can attempt to perform separate sentiment analysis on tweets that only belong to one of these classes (i.e. the training data would not be general but specific to one of these categories) and compare the results we get if we apply general sentiment analysis on it instead.

Last but not the least, we can attempt to model human confidence in our system. For example if we have 5 human labellers labelling each tweet, we can plot the tweet in the 2-dimensional objectivity / subjectivity and positivity / negativity plane while differentiating between tweets in which all 5 labels agree, only 4 agree, only 3 agree or no majority vote is reached. We could develop our custom cost function for coming up with optimized class boundaries such that highest weightage is given to those tweets in which all 5 labels agree and as the number of agreements start decreasing, so do the weights assigned. In this way the effects of human confidence can be visualized in sentiment analysis.

REFERENCES

[1] Albert Biffet and Eibe Frank.Sentiment Knowledge Discovery in Twitter Streaming Data.Discovery Science, Lecture Notes in Computer Science, 2010, Volume 6332/2010, 1-15, DOI: 10.1007/978-3-642-16184-1_1

[2] Alec Go, RichaBhayani and Lei Huang. Twitter Sentiment Classification using Distant Supervision. Project Technical Report, Stanford University, 2009.

[3] Alexander Pak and Patrick Paroubek.Twitter as a Corpus for Sentiment Analysis and Opinion Mining.In Proceedings of international conference on Language Resources and Evaluation (LREC), 2010.

[4] AndranikTumasjan, Timm O. Sprenger, Philipp G. Sandner and Isabell M. Welpe.Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment.In Proceedings of AAAI Conference on Weblogs and Social Media (ICWSM), 2010.

[5] Bo Pang, Lillian Lee and ShivakumarVaithyanathan. Thumbs up? Sentiment Classification using Machine Learning Techniques.In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2002.

[6] Chenhao Tan, Lilian Lee, Jie Tang, Long Jiang, Ming Zhou and Ping Li. User Level Sentiment Analysis Incorporating Social Networks. In Proceedings of ACM Special Interest Group on Knowledge Discovery and Data Mining (SIGKDD), 2011.

[7] EfthymiosKouloumpis, Theresa Wilson and Johanna Moore. Twitter Sentiment Analysis: The Good the Bad and the OMG! In Proceedings of AAAI Conference on Weblogs and Social Media (ICWSM), 2011.

[8] Hatzivassiloglou, V., &McKeown, K.R..Predicting the semantic orientation of adjectives.In Proceedings of the 35th Annual Meeting of the ACL and the 8th Conference of the European Chapter of the ACL, 2009.

[9] Johann Bollen, Alberto Pepe and Huina Mao.Modelling Public Mood and Emotion: Twitter Sentiment and socio-economic phenomena.In Proceedings of AAAI Conference on Weblogs and Social Media (ICWSM), 2011.

[10] Luciano Barbosa and JunlanFeng.Robust Sentiment Detection on Twitter from Biased and Noisy Data.In

Volume No:4, Issue No:9 (February-2019) ISSN No : 2454-423X (Online)
International Journal of Research in Advanced Computer Science Engineering
A Peer Reviewed Open Access International Journal
www.ijracse.com

Proceedings of the international conference on Computational Linguistics (COLING), 2010.

[11] Peter D. Turney. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. In Proceedings of the Annual Meeting of the Association of Computational Linguistics (ACL), 2002.

[12] Rudy Prabowo and Mike Thelwall. Sentiment Analysis: A Combined Approach. Journal of Infometrics, Volume 3, Issue 2, April 2009, Pages 143-157, 2009