ISSN: 0976-3031

Available Online at http://www.recentscientific.com

CODEN: IJRSFP (USA)

International Journal of Recent Scientific Research
Vol. 8, Issue, 7, pp. 18940-18945, August, 2017

International Journal of Recent Scientific Research

DOI: 10.24327/IJRSR

Research Article

TWITTER SENTIMENT ANALYSIS USING LEXICAL APPROACH AND ENHANCED NAIVE BAYES

Uma Maheswari P*., Pranesh Raagav S and Suriya Krishna M

Department of Computer Science and Engineering, Anna University, TamilNadu, India

DOI: http://dx.doi.org/10.24327/ijrsr.2017.0808.0597

ARTICLE INFO

Article History:

Received 15th May, 2017
Received in revised form 25th June, 2017
Accepted 28th July, 2017
Published online 28th August, 2017

Key Words:

Twitter sentiment analysis, lexical approach, social media analysis, microblogging, Enhanced naive Bayes.

ABSTRACT

Microblogging today has become a very popular communication tool in recent times. Millions of users share opinions on different aspects of life everyday in popular social networking sites such as Twitter, Tumblr and Facebook. Impressed by its growth, companies and media organisation are increasingly seeking ways to mine these social media for information about what people think about their companies and products. In this work, a hybrid method which performs the classification of tweet sentiment in Twitter has been executed. In this work, different methods for enhancing the accuracy of a Naive Bayes classifier for sentiment analysis has been explored and implemented. There is also a feature to compare two related trends in the twitter world which could be used to analyse the better product or better opinion in the region or worldwide.

INTRODUCTION

Social Media and Microblogging are generating a vast amount of sentiment rich data in the form of tweets, status updates, blog posts etc. “What other people think” has always been an important for most of us during the decision-making process. Sentiment analysis is extremely useful in social media monitoring as it allows us to gain an overview of the wider public opinion behind certain topics.

Twitter sentiment analysis is difficult when compared to general sentiment analysis due to the presence of slang words, emoticons and misspellings. For every tweet there are two important characteristics namely polarity and subjectivity which determine the appropriate context of the tweet. There are two approaches for determining them and they are lexical based and model based through machine learning. Even though the lexical based approach has no training process the machine learning model based approach is used for higher accuracy and redundancy elimination. Following are the major applications of sentiment analysis in real world scenarios and these make Sentiment Analysis important. They are Product and Service reviews, Reputation Monitoring, Result prediction and Decision Making.

There are existing systems which perform sentiment analysis based on the lexical approach in this proposed system we employ a combination of supervised machine learning techniques and lexical techniques (HYBRID approach)to develop a tool for visualizing and comparison of trends and mapping the opinions globally and also region wise. The lexical section of the system involves generating prior polarity score for the tweets in a standardised manner using dictionaries and this is used as a feature in the Machine Learning Classifier section. The Machine learning classifier module uses the Enhanced Naive Bayes approach to classify the polarity of the tweets and map them accordingly.

LITERATURE REVIEW

A 2-step automatic sentiment analysis method for classifying tweets has been designed (Barbosa.L. *et al*. Barbosa.L. and Feng.J., 2010) in which a noisy training set is used to reduce the labeling effort in developing classifiers. Firstly, the tweets were classified into subjective and objective tweets. After that, subjective tweets are classified as positive and negative tweets. A distant learning method which is used to acquire sentiment data (Go.A *et.al*., 2009) and since it mostly works with movie comments and tweets, additional features such as statements ending in positive emoticons like “:):-) ”are used as positive and negative emoticons like “:(:-(”are used as negative.

Corresponding author:* **Uma Maheswari P
Department of Computer Science and Engineering, Anna University, TamilNadu, India

Akshay Amolik *et.al*.(2016) is proposing a highly accurate model of sentiment analysis of tweets with respect to latest reviews of upcoming Bollywood or Hollywood movies. With the help of feature vector and classifiers such as Support vector machine and Naïve Bayes, we are correctly classifying these tweets as positive, negative and neutral to give sentiment of each tweet Models were built using Naive Bayes, Max Entropy and Support Vector Machines (SVM) and compare their performances. The classification based on emoticons would not show the original sentiment of the tweet as it may have phrases which conflict with the emoticons., a new algorithm was proposed for predicting the sentiment scores based on Naive Bayes method (Vala Ali Rahani *et al.* Rohani and Shayaa.D, 2015). In SentiRobo approach, the fact that the sentiment score of each keyword in a sentence can be a good indicator for overall sentiment polarity of the whole sentence was followed. Domain-dependent solutions for analyzing social media sentiments provided more accurate predictions based on the specified list of positive and negative keywords. This operated in multiple languages (English and Malay).

Dictionary based approach is also called as lexical based approach or language model-based approach (Ding.X. *et al*., 2008) research the holistic lexicon benchmark and fully automatic lexicon expansion is used for opinion mining, where the data used for sentiment analysis is fixed and labeled and relationships are established in RDBMS and polarity is determined. Opinion lexicon were utilized (desirable and undesirable) to predict the polarity of phrases using these labeled words. Lexical Senti-wordnet database that are being produced has been used in a research (Dang.Y. *et al.* , 2010) for sentiment classification in online product review case, using senti-wordnet and determines review subjectivity. A lexicon enhanced method was developed to generate a set of sentiment words based on a sentiment lexicon as a new feature dimension. The downside of lexical based method is that it relatively depends on established lexical database and language model architecture. However, there is an upside where it takes no training process and has better generality. Therefore, machine learning is used as possible alternative approach in order to improve the accuracy and performance of the predictor. Twitter sentiment analysis was used to look for correlations between the sentiments during a major event like FIFA World Cup 2014 (Barnaghi.P *et al*., 2016). A text categorization method called Bayesian Logistic Regression (BLR) Classification was used for providing positive or negative sentiment on tweets. Detailed insights into opinions and trends around sporting events were provided. This paved way to improve the quality of matches by highlighting controversial ethical issues. Akshi Kumar and Teeja Mary Sebastian (2012) expounded a hybrid approach using both corpus based and dictionary based methods to determine the semantic orientation of the opinion words in tweets.

Senti Word Net is proposed (Baharudin.B *et al*., 2011) which is sentiment analysis lexical resource made up of synset from Word Net, a thesaurus like resource; they are allocated a sentiment score of positive, negative or neutral. The sentences are split into subjective and objective parts. Senti Word Net dictionary is used to check the semantic scores. Final weight is calculated after checking semantic orientation. The contextual information and the sense of each individual sentence are extracted according to the pattern structure of the sentence. It suffered due to the following reasons: Dependency on lexicons (Senti-Word) and lack of term sense disambiguation. A broad overview of some of the machine learning techniques used in sentiment classification is provided (Pang.B. *et al*., 2002). An overview of three well- known machine learning methods for text categorization, including Naive Bayes, Logistic Regression and Support Vector Machine was provided. Movie reviews were used for classifying sentiment as positive or negative. Opinions are classified as one of the two opposing sentiment polarities (positive or negative), however, they may also be labelled as neutral when there is a lack of conclusive opinion. This kind of labeling can be used to summarize the content of opinionated texts and documents. Warih Maharani performs Maharani, 2013 comparison of Opinion Mining using Lexical and Machine learning methods. In the Lexical Method, Labelled data and Lexical Database were used for opinion mining whereas in the Machine learning method K-nearest neighbours method was used. Examination through lexical based approach is done by implementing with and without stemming process. For k-NN method, the system accuracy value depends on how many K (amount of nearest neighbor) are utilised. Model based approach with machine learning produce better accuracy rate than lexical based approach. discussed the methods in sentiment analysis. A specific lexicon of emoticons was used to reduce manual tweet tagging for sentiment classification based on happy and sad emoticons, the training set was split into positive and negative samples (Pak.A. *et al*., 2010).

METHODOLOGY

Lexical based Prior Polarity Scoring

To achieve the hybrid approach as shown in Figure 1, it is necessary that the Lexical Based approach is utilized Twitter Sentiment Analysis using Lexical Approach and Enhanced Naive Bayes to generate Prior Polarity Score for the tweet

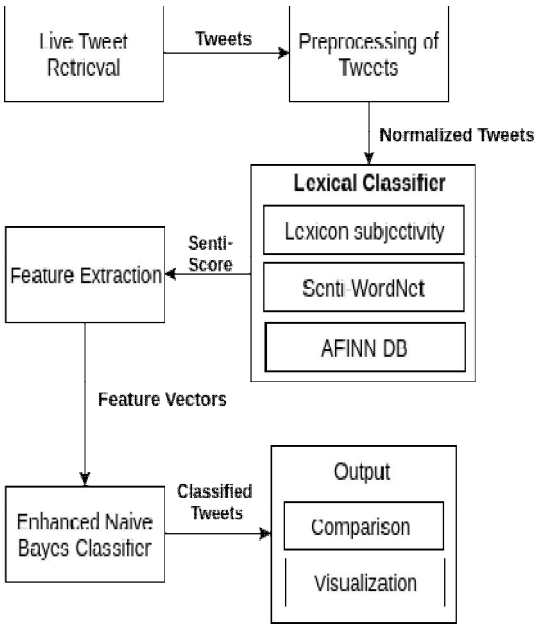


Figure 1 Block Diagram of Tweet Sentiment Analysis

which is used as a feature. These are standardised (Scores) results from three standard Dictionaries in the lexical model for

tweets are non-disambiguate in context. Dictionaries Used are AFINN DB, Senti-Word NET, and Lexicon Subjectivity DB. These well-established DBs are utilised in a hybrid approach for the orientation estimation which are used as features in the following module.

Algorithm Sentiment Orientation Calculation

Begin
For all Tweets $t_i \in T$;
Setup Dictionaries D_j where j =1, 2, 3;
Let the threshold be T_h ;
Calculate Senti−Score(SS) for each t_i with response to dictionary D_j ;
$\text{SS}_{\text{AVG}} = 1/m \sum_{j=1}^{m} SS(D_j)$; ***m=3;***

Standardise Senti−Score using Average−Weighted technique ;
$\text{SS}_{\text{w-AVG}}(W_j) = 1/m \sum_{j=1}^{m} W_j.SS(D_j)$;m=3;

where W is the weight of the dictionary
end.

Feature Extraction Module

Choosing the best feature set for sentiment analysis has the highest importance. Features are often preprocessed by various techniques in order to reduce the feature space. The collected dataset is used to extract features that will be used to train our sentiment classifier. Features like unigram, bigrams and trigrams were used, while for general information retrieval purposes, the frequency of a keyword's occurrence is a more suitable feature, since the overall sentiment may not necessarily be indicated through the repeated use of keywords. The basic reason for using feature selection (or reduction) methods for sentiment analysis is twofold: first, the reduced feature set decreases the computing demands for the classifier, and, second, removing irrelevant features can lead to better classification accuracy as it has a strong impact on the evaluation results. The process of obtaining bigrams from a Twitter post is as follows:

1. Filtering-we remove URL links (e.g. http://example. com), Twitter user names (e.g. @alex - with symbol @ indicating a user name), Twitter special words (such as "RT"6) and emoticons.
2. Tokenization-we segment text by splitting it by spaces and punctuation marks, and form a bag of words. However, we make sure that short forms such as "don't", "I'll", "she'd" will remain as one word.
3. Removing stopwords-we remove articles ("a", "an", "the") from the bag of words.
4. Constructing bigrams-a set of bigrams out of consecutive words. A negation (such as "no" and "not") is attached to a word which precedes it or follows it. For example, a sentence "I do not like fish" will form two bigrams: "I do+not", "do+not like", "not+like fish". Such a procedure allows to improve the accuracy of the classification since the negation plays a special role in an opinion and sentiment expression.

Classifier Model Module

The classification approach generally followed in this domain is a two-step approach. First Objectivity Classification is done which deals with classifying a tweet or a phrase as either objective or subjective. After this, Polarity Classification is performed (only on tweets classified as subjective by the objectivity classification). The polarity determination is done using the Enhanced Naive Bayes (as it produced better results than SVM Classifier) to determine whether the tweet is positive, negative or both. The model built is used to predict the sentiment of the new tweets. The algorithms used in training the classifier model are SVM Classifier and Enhanced Naive Bayes technique.

Naive Bayes

It is a simple probabilistic model based on the Bayes rule along with a strong independence assumption. It involves simplifying conditional independence assumption. The maximum likelihood probability of a word belonging to a particular class is given by,

$$P(x_i|c) = \frac{Count\ of\ x_i\ in\ tweet\ of\ class\ c}{total\ number\ of\ words\ in\ the\ tweet\ of\ class\ c} \quad (1)$$

Frequency count of the words are stored in hash tables. According to the Bayes Rule, the probability of a particular tweet (t) belonging to a class (Ci) is given by conditional dependence probability.

$$P(c_i|t) = \frac{P(t|c_i) * P(c_i)}{P(t)} \quad (2)$$

On Applying Conditional Independence assumption the equation becomes:

$$P(c_i|t) = \frac{(\pi P(x_i|c_j)) * P(c_j)}{P(t)} \quad (3)$$

Here, the xi are the individual words of the tweet.

Enhanced Naive Bayes

The main difference in the Enhanced Naive Bayes algorithm is the Laplacian Smoothing technique used in addition to the Normal Naive Bayes algorithm.

Laplacian Smoothing

With the Naïve Bayes Assumption, we can still end up with zero probabilities i.e. if the classifier encounters a word that has not been seen in the training set, the probability of both the classes would become zero. This is bad because all the other words in the tweet are ignored just because of a single rare word. To avoid this Laplacian Smoothing is used as follows:

$$P(x_i|c_j) = \frac{(Count(x_i)) + k}{(k+1) * (No\ of\ words\ in\ class\ \ C_j)} \quad (4)$$

where k is usually 1 and this tweet must be with unique words avoiding duplicates.

Table 1 Opinion Types table

Prediction	Class Positive	Class Negative
Predicted Negative	TN	FP
Predicted Positive	FN	TP

TN-True Negative, TP-True Positive, FN- False Negative, FP- False Positive.

Negation Handling

Negations are those words which affect the sentiment orientation of other words in a sentence. Examples of negation words include not, no, never, cannot, shouldn't, wouldn't, etc. Negation handling is an automatic way of determining the scope of negation and inverting the polarities of opinionated words that are actually affected by a negation. A state variable is used to monitor the negation status. It is set for a word which is followed by a "not" or "nt" and transforms into a "not word". The state variable is reset when a double negation or punctuation is encountered.

negated := False
for each word in document:
if negated = True :
Transform word to ' ' not_ ' ' + word.
if word is ' ' not ' ' or ' 'n ' t ' ' :
negated := not negated
if a punctuation mark is encountered
negated := False .

Mutual Information

Mutual information is a quantity that measures the mutual dependence of the two random variables. The mutual information of two discrete random variables X and Y can be defined as:

$$\left(I(X;Y) = \sum_{y \epsilon Y} \sum_{x \epsilon X} p(x,y) log \frac{p(x,y)}{p(x)p(y)} \right) \qquad (5)$$

p(x,y) is the joint probability distribution function of X and Y. p(x) and p(y) are the marginal probability distribution functions of X and Y respectively. Here X is an individual feature which can take 2 values Y is the class type positive or negative.

RESULTS AND DISCUSSION

Sentiment140 allows you to discover the sentiment of a brand, product, or topic on Twitter. Sentiment140 uses classifiers based on machine learning algorithms and allow users to see the classification of individual tweets. The API lets users classify tweets and integrate sentiment analysis classifier functionality into their own websites or applications. The classifier evaluation is usually concerned with the classifier effectiveness rather than its efficiency. Evaluation of the correctness of the classifier predictability is more pivotal than its computational complexity. The confusion table is presented in the following Table 1

Data Set Format

The dataset is in a CSV format with emoticons removed.
Data file format has 6 fields

1. The polarity of the tweet (0 = negative, 4 = positive)
2. The id of the tweet (2087)
3. The date of the tweet (Sat May 16 23:58:44 UTC 2009)
4. The query (lyx). If there is no query, then this value is NOQUERY.
5. The user that tweeted (robotickilldozr)
6. The text of the tweet (Lyx is cool)

The available 50,000 tweets were divided into two sections having 40,000 and 10,000 for traning and testing purposes respectively.

Evaluation

Precision measures the proportion of positives that are correctly identified to the summation of false and true positives.

$$Precision = \frac{Number\ of\ TP}{Number\ of\ TP + Number\ of\ FP} \qquad (6)$$

Recall measures the proportion of positives that are correctly identified to the summation of false and true negatives.

$$Recall = \frac{Number\ of\ TP}{Number\ of\ TP + Number\ of\ FN} \qquad (7)$$

Specificity measures the proportion of negatives that are correctly identified.

$$Specificity = \frac{Number\ of\ TN}{Number\ of\ TN + Number\ of\ FP} \qquad (8)$$

Accuracy is the ratio of summation of true positives and true negatives to the total tweets retrieved.

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

F − Score can be used as a single measure of performance of the test for the positive class. The F-score is the harmonic mean of precision and recall.

$$F - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \qquad (10)$$

MCC is used as a measure of the quality of binary classifications.

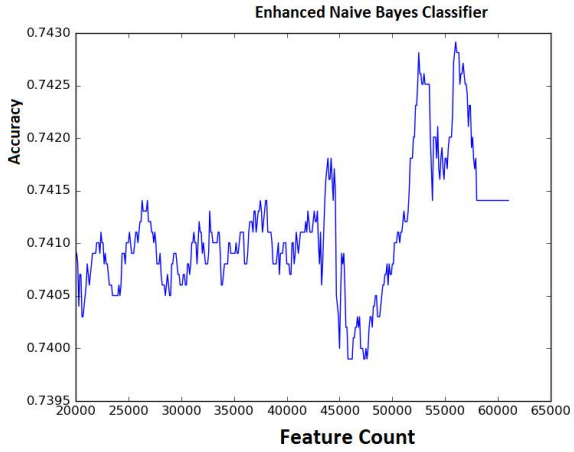
$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)\,(TP + FN)\,(TN + FP)\,(TN + FN)}} \qquad (11)$$

The Results of Enhanced Naive Bayes Classifier are tested and metric parameters were evaluated as shown in Table 2 and compared with those of the SVM classifier. Figure 2 shows the best possible feature count for the Enhanced Naive Bayes algorithm for improved performance. From the graph, the Figure 3 shows that The Results of Enhanced Naive Bayes classifier are tested and metric parameters were evaluated as shown in Table 2 and compared with those of the SVM classifier.

Table II Matric Evaluation Table-Enhanced Naïve Bayes

Metric evaluation parameter	Enhanced Naïve Bayes
Precision.	0.7511
Best K (Feature Count)	55900
Recall	0.7222
Matthew Correlation Coefficient	0.4831
Accuracy	0.7429
F-Score	0.7363

Graph 1 shows the best possible feature count for the Enhanced Naive Bayes algorithm for improved performance. From the graph, the Figure 3 shows the accuracy of Enhanced Naive Bayes outperforms the SVM classifier. In information retrieval, precision is a measure of result relevancy, while recall is a measure of how many truly relevant results are returned. Figure 4 shows that the Enhanced Naive Bayes was able to perform better than the SVM classifier in terms of precision also which means that false positive rate is lesser than SVM. High scores for both show that the classifier is returning accurate results (high precision), as well as returning a majority of all positive results (high recall). They are both sensitive to parameter optimization. In Figure 5 the Enhanced Naive Bayes is able to outperform the SVM classifier. Conventional Naïve Bayes Classifier and Support Vector Machine (SVM) more or less have the same efficiency in F-Score. Figure 6 also shows the Enhanced Naive Bayes algorithm is far ahead than the conventional SVM Classifier.



Graph 1 Graphical Analysis of Enhanced Naive Bayes-Best

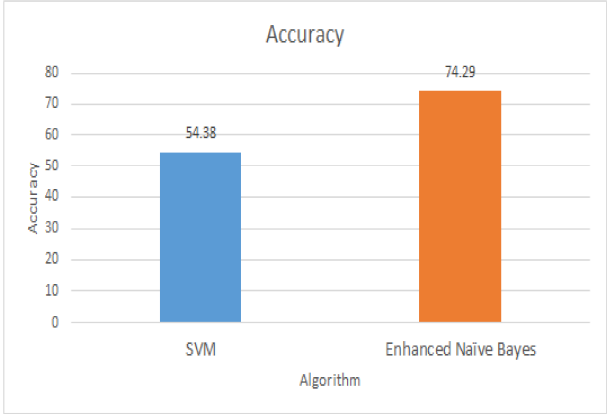


Figure 2 Accuracy Comparison

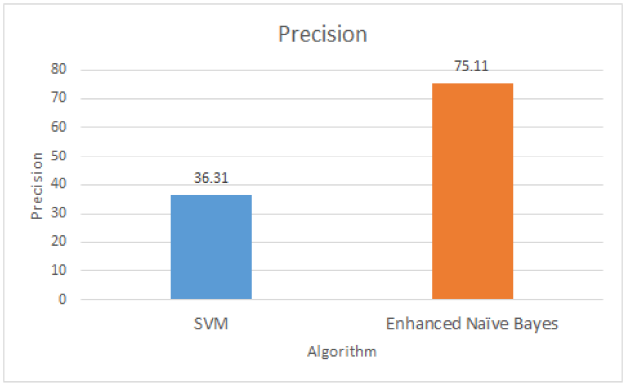


Figure 3 Precision Comparison

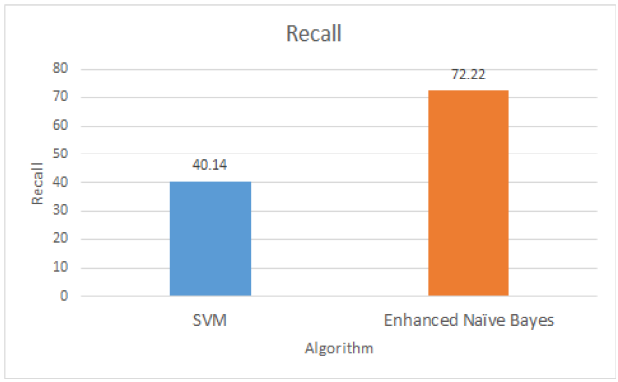


Figure 5 Recall Comparison

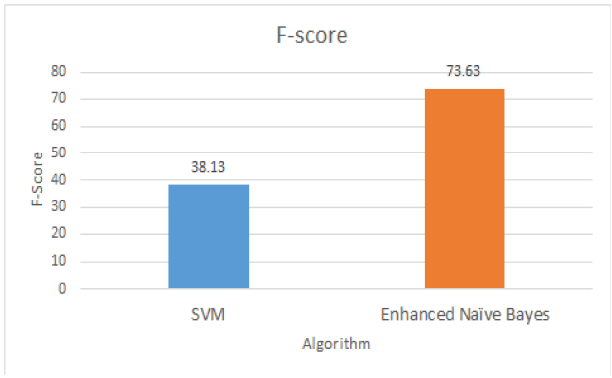


Figure 6 F-Score Comparison

CONCLUSION

Studies in Sentiment Analysis approaches have existed for more than a decade and now are exploited by enterprises as an important tool for strategic marketing, planning and maneuvering. The goal of this project was to classify the tweets and determine the polarity of the opinions from the large twitter community using a hybrid approach. Based on the result of system tests and analysis it can be concluded that scoring result using Hybrid approach (lexical database and Enhanced Naive Bayes) is able to classifying opinion into positive and negative. The results show that a simple Naive Bayes classifier can be enhanced to match the classification accuracy of more complicated models for sentiment analysis by choosing the right type and count of features and removing noise by appropriate feature selection. Naïve Bayes classifiers due to their conditional independence assumptions are extremely fast to train and can scale over large datasets. They are also robust to noise and less prone to over fitting. Ease of implementation is also a major advantage of Naive Bayes. They were thought

to be less accurate than their more sophisticated counter parts like support vector machines and logistic regression but the results show that a significantly high accuracy can be achieved if features are pruned.

References

Barbosa.L. and Feng.J., (2010), "Robust Sentiment Detection on Twitter from Biased and Noisy Data". In:Proceedings of the 23rd International Conference on Computational Linguistics: Posters, Association for Computational Linguistics, 36 - 44, Beijing, August.

R. Piryani *et al*., (2017), "Analytical mapping of opinion mining and sentiment analysis research during 2000-2015", *Journal of Information Processing and Management*, Elsevier Publications, 52, 122-150.

Akshi Kumar and Teeja Mary Sebastian (2012), "Sentiment Analysis on Twitter", *IJCSI International Journal of Computer Science* Issues, 9(4), No 3, 372-378.

Barnaghi.P., Breslin.J.G, and Ghaffari.P.(2016), "Opinion Mining and Sentiment Polarity on Twitter and Correlation Between Events and Sentiment". In: Big Data (Big Data), IEEE International Conference on, 2686-2691.

Akshay Amolik *et al*. (2016), "Twitter Sentiment Analysis of Movie Reviews using Machine Learning Techniques", *International Journal of Engineering and Technology*, 7(6), 2038-2044.

Dang.Y., Zhang.Y., and Chen.H.(2010), "A Lexicon-Enhanced Method for Sentiment Classification: An Experiment on Online Product Reviews". IEEE Intelligent Systems, University of Arizona, 25(4), 46-53.

Ding.X., Liu.B., and Yu.P. (2008), "A Holistic Lexicon-Based Approach to Opinion Mining". In: Proceedings of the first ACM International Conference on Web search and Data Mining (WSDM'08), Chicago, 231-240, California, USA.

Khan.A. and Baharudin.B. (2011), "Sentiment Classification by Sentence Level Semantic Orientation using SentiWordNet from Online Reviews and Blogs". In: Universiti Teknologi PETRONAS Perak, Malaysia, August, Vol-2 , 539-549.

Maharani, Warih (2013), "Microblogging Sentiment Analysis with Lexical Based and Machine Learning Approaches". In: International Conference of Information and Communication Technology (ICoICT), IEEE, 439-443.

Pak.A. and Paroubek.P. (2010) , "Twitter as a Corpus for Sentiment Analysis and Opinion Mining". In: Proceedings of the 7th International Conference on LREC, 1320-1326.

Pang.B., Lee.L, and Vaithyanathan.S (2002), "Thumbs-up? Sentiment classification using machine learning techniques". In: Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing, Cornell University Ithaca, NY USA, 79-86

Rohani, Vala Ali and Shahid Shayaa.D.(2015), "Utilizing Machine Learning in Sentiment Analysis: SentiRobo Approach". In: Computer and Information Sciences (ICCOINS) 3rd International Conference, 397-402.

How to cite this article:

Uma Maheswari P *et al*.2017, Twitter Sentiment Analysis Using Lexical approach and Enhanced Naive Bayes. *Int J Recent Sci Res.* 8(8), pp. 18940-18945. DOI: http://dx.doi.org/10.24327/ijrsr.2017.0808.0597

****