World Applied Sciences Journal 35 (1): 54-59, 2017
ISSN 1818-4952
© IDOSI Publications, 2017
DOI: 10.5829/idosi.wasj.2017.54.59

Sentimental Analysis of Tweets Using Naive Bayes Algorithm

M. Vadivukarassi, N. Puviarasan and P. Aruna

Annamalai University, Tamil Nadu, India

Abstract: Sentiment analysis is a current research area in text mining. It is the stem of natural language processing or machine learning methods. It is the important sources of decision making and can be extracted, identified, evaluated from the online sentiments reviews. The main goal is to connect on Twitter and search for the tweets that contain a particular keyword and then evaluate the polarity of the tweets as positive and negative. In this paper, the keywords are collected from Twitter using Twitter API and the extracted raw data are preprocessed using Natural Language Toolkit techniques. The sentiments of the online tweets are evaluated based on feature selection of score words. In order to select the best features Chi Square test is used and Naïve Bayes classifier is used for training and testing the features and also evaluating the sentimental polarity. The proposed system is implemented using Python.

Key words: Feature selection · Natural Language Toolkit · Naïve Bayes classifier · Sentiment analysis

INTRODUCTION

Today, the micro blogging has become a very popular messaging tool between internet users. Millions of users can share their opinions in different aspects of life everyday in popular websites like Twitter, Tumblr and Face book. Twitter is a social website that offers the opportunities for the scrutiny of articulated humor [1]. Twitter supports brief explanation of ideas via short messages of tweets that are no longer than 140 characters. It allow for valuable and well-timed statement of information. These tweets are routinely posted as a stream on the user's report in Twitter and immediately sent to the user's network of followers. Twitter allows people to create profiles, communicate, and connect with other people on the service. The social link on twitter is asymmetric and can be conceptualized as a directed social network or follower network [2]. Towards specific product, organization, movies, events, news, issues, services and their attributes, the sentimental analysis is used to obtain the real influence of people. This can be useful in several ways and contains the computer science branches such as Natural Language Processing (NLP), text mining, information theory, machine learning and coding [3]. The main aim is to identify the sentiment of the tweets or reviews published in the web. It is possible to obtain a full report in the view of the requester, including a review about what people are feeling about an item without the need of find and read all topics and news related to it. This paper is made further as: Section 2 discusses literature survey examined till now. Section 3 describes block diagram of the proposed system. Section 4 presents result and analysis of the work using the graphical analysis. Section 5 ends with the conclusion.

Related Works: The engineering student's problem is mined using multi-label classifier to consider only negative aspects [4]. A system used to calculate the semantic distance from a word as good or bad had discussed. They also studied the work at a better level and used strings or words as classification topic. They grouped the words into two level, "good" and "bad" and then use certain functions to calculate the overall "goodness" or "badness" score for the documents. Some authors had suggested the approach to record different score scales to distinguish the views of users. They used the stars rating features and recorded online reviews of users on dissimilar services or products. These recorded stars rating are then utilized to recognize the service or product performance measures [5]. Researchers have also begun to examine various ways for routinely collecting the training data. The existing sentiment sites in Twitter have been exploited for collecting training data [6]. The hash tag is created for the training data, but they boundary their experiments to sentiment/non-sentiment categorization, rather than three way polarity

Corresponding Author: M. Vadivukarassi, Annamalai University, Tamil Nadu, India.

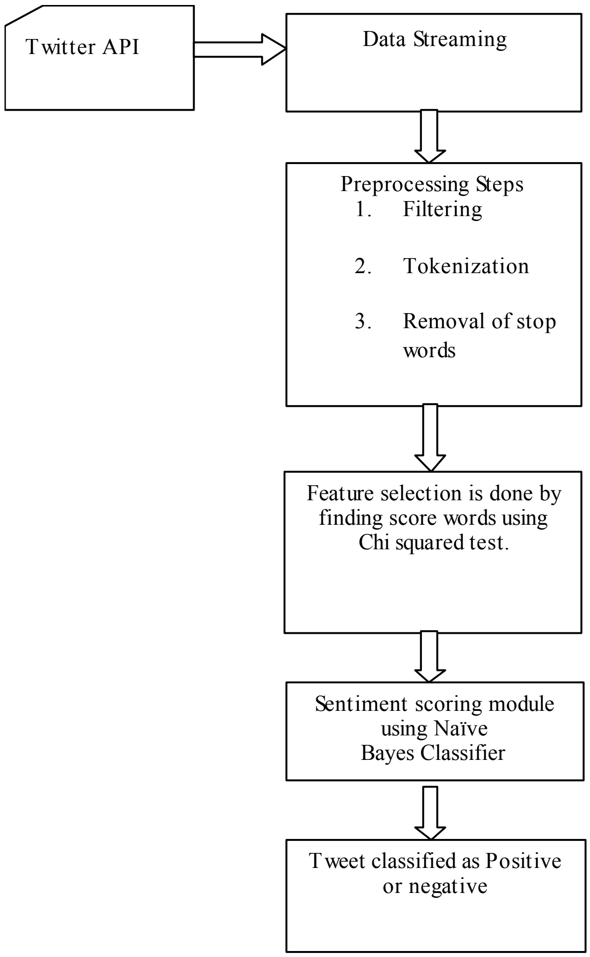

Twitter API
Data Streaming
Preprocessing Steps
1. Filtering
2. Tokenization
3. Removal of stop words
Feature selection is done by finding score words using Chi squared test.
Sentiment scoring module using Naïve Bayes Classifier
Tweet classified as Positive or negative

Fig. 1: Block diagram of the proposed system

classification has been produced [7]. An easy and whole system with the Hadoop support for sentiment mining on big datasets with Naive Bayes Classifier (NBC) has been presented. Instead of using Mahout Library, they implemented NBC to get fine grain power of the analysis procedure for a Hadoop execution. They have demonstrated that NBC is capable to scale up and to analyze the sentiment of millions film reviews with growing throughput [8]. Sentiments have analyzed on the document level. There is the most distinguished work between "Poor" and "Excellent" seed words to compute the semantic orientation by using mutual information approach. The sentiment orientation (SO) of the document was computed as the average of all such phrases. The accuracy results achieved to 66% [9]. Some researchers have classified review as neutral category also [10]. In the sentiment analysis domain, it involves filtering, tokenization of a particular corpus of text and removal of stop words as the preprocessing steps and then assigned a polarity score of the tweets [11]. The aggregated sums of these scores determine the sentiment behind the text and it is generally classified as positive or negative depending on the calculated score.

Proposed System: The sentences that represent observations or attitude that is expressed as positive or negative are called as sentiments. The flowchart describing an overall system design of twitter sentimental analysis is shown in Fig. 1. The users post their tweets in twitter. These tweets are extracted in the form of unstructured data. The unstructured dataset is converted into structured form then extracts features from structured review. The features of the words are selected and then classification technique is applied on extracted features to classify them into its sentiment polarity that is namely either positive or negative. Feature words representation based on Naïve Bayes classifier is the main algorithm proposed by information retrieval researchers to represent text corpus. It is an easy approach to convert unstructured text into structured data based on word by word and the grammar is neglected.

The standard algorithm is declared as follows:

Algorithm 1: Sentimental analysis in Twitter

```
Input: Tweets from twitter
Output: Sentimental Polarity of tweets
Method
Step 1: Get the input tweets
Step 2: Assign number of features n[i] where i ranges in
 10,100,1000,10000 and 15000.
Step 3: for n in n[i]
     Create wordscore()
     Find_best_words(wordscore,n)
     Evaluate (best_word_features)
Step 4: Create wordscore()
     a) Assign posword[j] and negword[k]
     b) Split the words by removing punctuations
     c) Build frequency distribution of all words
d) Find number of positive, negative and total number of words
e) Build dictionary of the wordscore based on the Chi-square test (i.e) word_score[t]
Step 5: By sorting the wordscore, bestwords are found
Step 6: Evaluate (best_word_features)
     a) Assign posfeatures[j] and negfeatures[k]
     b) Split the sentences into individual words.
c) Select ¾ of the features for training and ¼ of the features for testing
```

d) Train using Naïve Bayes Classifier
Step 7: Tweets are classified as positive and negative based on
the score words

The steps of this algorithm are discussed below:

Data Streaming: Data streaming is used to merge and access the real-time feeds and archived data for analytics. The collected raw tweets are applied as input to produce the sentimental polarity. Twitter presents two kinds of APIs to extract the tweets: Search API used for dumping old tweets while Streaming API used for dumping live tweets. Using Search API, training data set is built for sentiment classification; using Streaming API, the current results will be displayed. The twitter data is needed for classification and training the classifier. It is one of the most overstated parts of public networking site, it consists of various blogs, which are related to various areas worldwide.

Pre-Processing of Extracted Data: In this stage, the tweets are mined using Twitter Streaming API. Initially, it cleans the unstructured textual data into structured textual data by removing the punctuations and additional symbols. It may outcome in inefficiencies and may change the accuracy of the overall process. So, the preprocessing techniques are required for obtaining better results. This raw data are to be preprocessed as shown in the following steps:

- *Filtering:* In this step, the special words, user names in twitter are removed.
- *Tokenization:* In this step, the texts are tokenized by splitting the text into spaces and punctuations marks and then form bottle of words.
- *Removal of Stop words:* Articles and other stop words are removed in this step.

Feature Selection: The word scores of the features are tested based on Chi-square method. It creates a list of all positive and negative words. Then build frequency distribution of all words and then frequency distribution of words within positive and negative labels. Finally, the number of positive and negative words as well as the total number of words and the dictionary of word score based on Chi-Square test is found. This test method gives good result for both positive and negative classes and it is also used to select feature from high dimensional data. So that word scores are found and the best number of words based on word scores are also extracted.

Chi-Square: Let the total number of tweets in the collection be 'n', the conditional probability of class 'i' for tweets which contain 'w' be $p_i(w)$, the global fraction of tweets containing the class i be 'Pi', and the global fraction of tweets which contain the word 'w' be 'F(w)'. γ_i is the way of measuring the correlation between conditions and classes. Therefore, the γ^2 of the word between word 'w' and class 'i' is defined as:

$$\gamma_i^2 = \frac{n, F(w)^2, (p_1(w) - P_1)^2}{F(w), (1 - F(w)), P_t, (1 - P_1)} \tag{1}$$

Sentiment Scoring: The polarity of words has the basic number of features based on the selection process. English language words assign a score by referring dictionary. This scoring module determines the score of sentiments during the analysis of data. Naïve Bayes classifiers are used to classify the sentiment. Naïve Bayes algorithm is used to classify the sentiment and this sentiment orientation performs well with more accuracy.

Naïve Bayes Classifier: It is used to predict the probability for a given words to belong to a particular class. It is used because of its easiness in both during training and classifying steps. Pre-processed data is given as input to train input set using Naïve Bayes classifier and that trained model is applied on test to generate either positive or negative sentiment. The Bayes theorem is as follows.

$$P(\frac{H}{X}) = \frac{P(\frac{H}{X})P(H)}{P(X)} \tag{2}$$

where X- Tuples, H-Hypothesis, P(H|X) represents Posterior probability of H conditioned on X i.e. the Probability that Hypothesis holds true given the value of X, P(H) represents Prior probability of H i.e the Probability that H holds true irrespective of the tuple values, P(X|H) represents posterior probability of X conditioned on H i.e. the Probability that X will have certain values for a given Hypothesis, P(X) represents Prior probability of X i.e the Probability that X will have certain values. The proposed system understands whether the tweet is positive or negative based on the dictionary methods of score. An experiment result of accuracy is evaluated using following information retrieval matrices. Accuracy is the performance evaluation parameter and it is calculated by number of correctly selected positive and negative words divide by total number of words present in the corpus. The formula is given as below.

$$Accuracy = \frac{\sum True\ Positive + \sum True\ Negative}{\sum Total\ number\ of\ words} \quad (3)$$

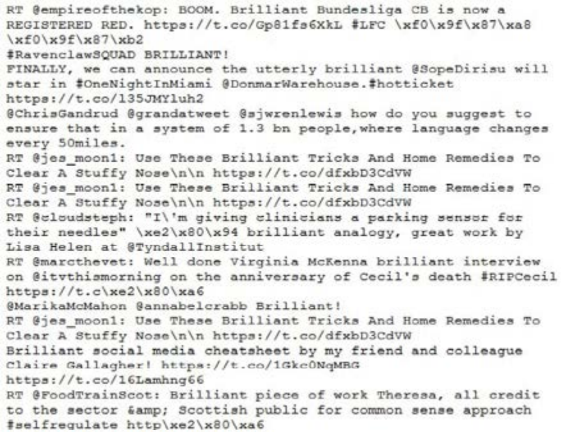
where True positive is number of tweets recognized as positive and true negative is number of tweets recognized as negative respectively.

RESULTS AND DISCUSSION

In the proposed system, Python is used for implementation. The twitter data programmatically can be accessed by creating an application in twitter that interacts with the Twitter API. In order to search for particular tweets, Oauth protocol is used for authentication. The first step is to login the Twitter (if you're not already logged in) and register a new application. The name and a description for the app are chosen. From this, a consumer key and a consumer secret are received. These are the application settings that should always be kept private.

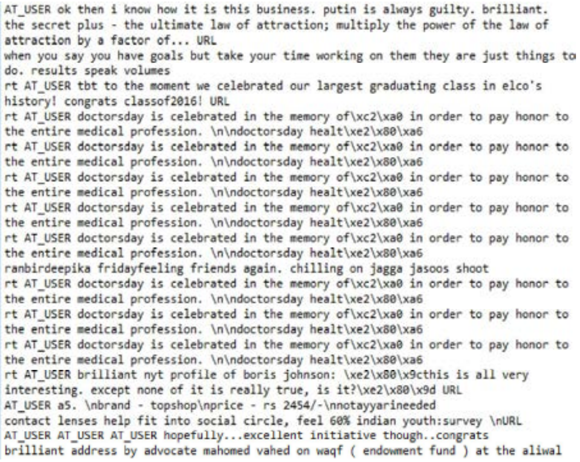
From the configuration page of the app, an access token and an access token secret are also required. Similarly to the consumer keys, these strings must also be kept private. They provide the application access to Twitter on behalf of the user account. The default permissions are read-only. Using these four keys, the only text data from the twitter are filtered based on the keyword in the particular location and languages.

In this section, the experimental results are presented based on extracting the particular positive and negative keywords in the location as India and also the language as English using Twitter Streaming API. The three positive keywords such as 'attraction',' brilliant' and 'celebrated' are filtered from twitter and also three negative keywords such as 'afraid', 'bloody', 'complaint' are filtered as shown in the Fig.2.



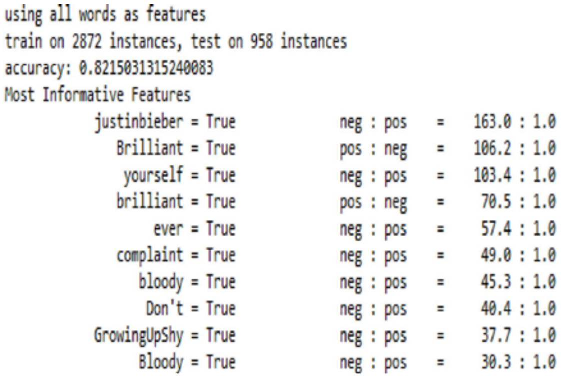
```
RT @empireofthekop: BOOM. Brilliant Bundesliga CB is now a
REGISTERED RED. https://t.co/Gp81fs6XkL #LFC \xf0\x9f\x87\xa8
\xf0\x9f\x87\xb2
#RavenclawSQUAD BRILLIANT!
FINALLY, we can announce the utterly brilliant @SopeDirisu will
star in #OneNightInMiami @DonmarWarehouse.#hotticket
https://t.co/135JMYluh2
@ChrisGandrud @grandatweet @sjwrenlewis how do you suggest to
ensure that in a system of 1.3 bn people,where language changes
every 50miles.
RT @jes_moon1: Use These Brilliant Tricks And Home Remedies To
Clear A Stuffy Nose\n\n https://t.co/dfxbD3CdVW
RT @jes_moon1: Use These Brilliant Tricks And Home Remedies To
Clear A Stuffy Nose\n\n https://t.co/dfxbD3CdVW
RT @cloudsteph: "I\'m giving clinicians a parking sensor for
their needles" \xe2\x80\x94 brilliant analogy, great work by
Lisa Helen at @TyndallInstitut
RT @marcthevet: Well done Virginia McKenna brilliant interview
on @itvthismorning on the anniversary of Cecil's death #RIPCecil
https://t.c\xe2\x80\xa6
@MarikaMcMahon @annabelcrabb Brilliant!
RT @jes_moon1: Use These Brilliant Tricks And Home Remedies To
Clear A Stuffy Nose\n\n https://t.co/dfxbD3CdVW
Brilliant social media cheatsheet by my friend and colleague
Claire Gallagher! https://t.co/1GkcONqMBG
https://t.co/16Lamhng66
RT @FoodTrainScot: Brilliant piece of work Theresa, all credit
to the sector &amp; Scottish public for common sense approach
#selfregulate http\xe2\x80\xa6
```

Fig. 2: Extracting only tweets from Twitter



```
AT_USER ok then i know how it is this business. putin is always guilty. brilliant.
the secret plus - the ultimate law of attraction; multiply the power of the law of
attraction by a factor of... URL
when you say you have goals but take your time working on them they are just things to
do. results speak volumes
rt AT_USER tbt to the moment we celebrated our largest graduating class in elco's
history! congrats classof2016! URL
rt AT_USER doctorsday is celebrated in the memory of\xc2\xa0 in order to pay honor to
the entire medical profession. \n\ndoctorsday healt\xe2\x80\xa6
rt AT_USER doctorsday is celebrated in the memory of\xc2\xa0 in order to pay honor to
the entire medical profession. \n\ndoctorsday healt\xe2\x80\xa6
rt AT_USER doctorsday is celebrated in the memory of\xc2\xa0 in order to pay honor to
the entire medical profession. \n\ndoctorsday healt\xe2\x80\xa6
rt AT_USER doctorsday is celebrated in the memory of\xc2\xa0 in order to pay honor to
the entire medical profession. \n\ndoctorsday healt\xe2\x80\xa6
rt AT_USER doctorsday is celebrated in the memory of\xc2\xa0 in order to pay honor to
the entire medical profession. \n\ndoctorsday healt\xe2\x80\xa6
ranbirdeepika fridayfeeling friends again. chilling on jagga jasoos shoot
rt AT_USER doctorsday is celebrated in the memory of\xc2\xa0 in order to pay honor to
the entire medical profession. \n\ndoctorsday healt\xe2\x80\xa6
rt AT_USER doctorsday is celebrated in the memory of\xc2\xa0 in order to pay honor to
the entire medical profession. \n\ndoctorsday healt\xe2\x80\xa6
rt AT_USER doctorsday is celebrated in the memory of\xc2\xa0 in order to pay honor to
the entire medical profession. \n\ndoctorsday healt\xe2\x80\xa6
rt AT_USER brilliant nyt profile of boris johnson: \xe2\x80\x9cthis is all very
interesting. except none of it is really true, is it?\xe2\x80\x9d URL
AT_USER a5. \nbrand - topshop\nprice - rs 2454/-\nnotayyarineeded
contact lenses help fit into social circle, feel 60% indian youth:survey \nURL
AT_USER AT_USER AT_USER hopefully...excellent initiative though..congrats
brilliant address by advocate mahomed vahed on waqf ( endowment fund ) at the aliwal
```

Fig. 3: Preprocessed data of the tweets



```
using all words as features
train on 2872 instances, test on 958 instances
accuracy: 0.8215031315240083
Most Informative Features
          justinbieber = True           neg : pos    =    163.0 : 1.0
             Brilliant = True           pos : neg    =    106.2 : 1.0
              yourself = True           neg : pos    =    103.4 : 1.0
             brilliant = True           pos : neg    =     70.5 : 1.0
                  ever = True           neg : pos    =     57.4 : 1.0
             complaint = True           neg : pos    =     49.0 : 1.0
                bloody = True           neg : pos    =     45.3 : 1.0
                 Don't = True           neg : pos    =     40.4 : 1.0
          GrowingUpShy = True           neg : pos    =     37.7 : 1.0
                Bloody = True           neg : pos    =     30.3 : 1.0
```

Fig. 4: Extracting features using all words in the tweets

The structure of a tweet is discussed and digging into the processing steps for text analysis. A number of tweets is collected and stored them in database. The content of a tweet is embedded in the text to analyze by breaking the text down into words. Tokenization is used to split a stream of text into smaller units called tokens, usually words or phrases. The popular Natural language Toolkit (NLTK) library is used to tokenize a fictitious tweet. The raw data of the twitter are preprocessed using NLTK library techniques as shown in Fig. 3. The preprocessed tweets are appended for feature selection process in each list. The score words are evaluated based on Chi- Square test. The features which contain highest score that is the toppest feature of the tweets. It indicates that these features are the repeated words in the database. The one third of the features is selected for training and rest of the features is used for testing. Finally Naïve Bayes classifier is trained.

In the Fig. 4, the experiments are conducted using all words as features in the tweets. It is trained on 2872 instances and tested on 958 instances from the database.

Table 1: TOP 10 INFORMATIVE FEATURES OF THE TWEETS

Most Informative Features	Polarity	Word scores
Justinbieber	Negative	163.0
Brilliant	Positive	106.2
Yourself	Negative	103.4
Brilliant	Negative	70.5
Ever	Negative	57.4
Complaint	Negative	49.0
Bloody	Negative	45.3
Don't	Negative	40.4
GrowingUpShy	Negative	37.3
Bloody	Negative	30.3

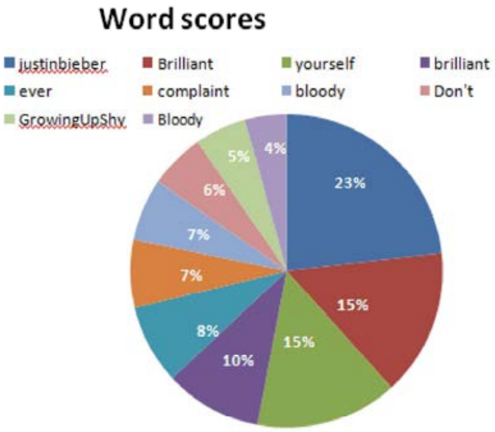


Fig. 5: Word scores tweets

It shows the top most useful features among the tweets along with sentimental polarity. Table 1 indicates the most informative features of the tweets.

The scores of words based on Chi-Squared tests are analyzed using the Pie chart as shown below in the Fig. 5. It shows the percentage of the top ten informative features from the tweets in the database.

Sentiment is scored based on the words contained in a tweet. The word 'justinbieber' contains 23% of score words in the database and it evaluates the polarity as negative. It is used to find the number of positive and negative words as well as the total number of words of the processed tweets. Twitter streaming API starts extracting tweets about the corresponding keywords and it redirects those extracted tweets perform preprocessing to remove all the unnecessary information and replace some information. Then, the Naïve Bayes classifier classifies sentiment words (positive or negative). The preprocessed tweets are evaluated based on sentimental analysis.

Table 2: Number of Features Selected and Accuracy

No of features	Accuracy
10	0.500
100	0.500
1000	0. 635
10000	0.763
15000	0.791

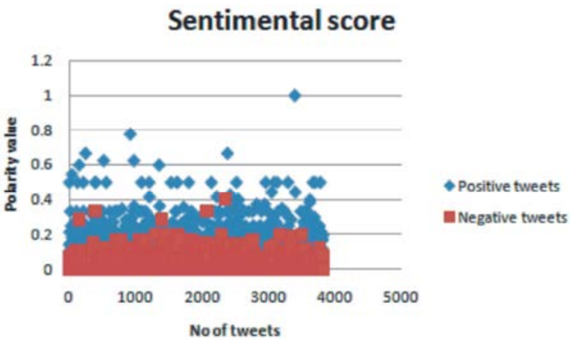


Fig. 6: Graphical analysis of sentimental score for keywords

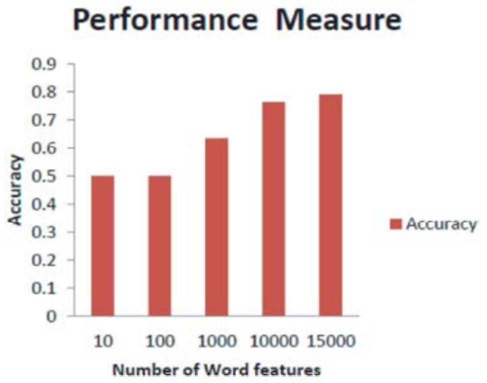


Fig. 7: Graphical analysis of various numbers of features

Table 2 shows the accuracy of the word features. The algorithm is applied to all tweets mentioning both positive and negative tweets. The Naïve Bayes classifier approach gained an average accuracy of 0.821 for all the number of features in the tweets. If 10 features are selected, then the accuracy is 0.500. Similarly 100, 1000, 10000, 15000 features of the tweets are selected, and then the accuracy of each features are 0.500, 0.635, 0.763 and 0.791 respectively.

The graphical analysis of sentimental score for keywords as positive and negative is shown in Fig. 6. Then, the graphical analysis of different number of features of the words is selected with accuracy based on Chi-Square test as shown in the Fig. 7.

CONCLUSION

Twitter is an excellent initial point for social media analysis. People directly share their opinions through Twitter to the general public. One of the very common analyses which can perform on a large number of tweets is sentiment analysis. In the proposed work, tweets are collected using Twitter streaming API from twitter. The collected tweets are preprocessed using Natural Language Toolkit techniques. The features of the tweets are selected based on Chi-Square test and Naïve Bayes classifier is used to classify the tweets as positive and negative. This proposed work is implemented using Python. The experiments are conducted based on different features such as 10,100,1000,10000 respectively. It is found that if the number of features increases, the accuracy of the selected features also increases. This proposed system would be easy for user to obtain the summarized report about the opinion from Twitter. It is also used to support them in decision making process in their daily life activities.

REFERENCES

1. Jaba Sheela, L., 2016. A Review of Sentiment Analysis in Twitter Data Using Hadoop, International Journal of Database Theory and Application, 9(1): 77-86,2016.
2. Farzindar Atefeh And Wael Khreich, 2013. A Survey Of Techniques For Event Detection In Twitter, Computational Intelligence, Volume 0, Number 0, 2013.
3. Sai Krishna, D., G Akshay Kulkarni and A. Mohan, 2015. Sentiment Analysis-Time Variant Analytics, International Journal of Advanced Research in Computer Science and Software Engineering, 5(3).
4. Chen, X., M. Vorvoreanu and K. Madhavan, 2014. Mining Social Media Data for Understanding Students' Learning Experiences, IEEE Transactions on Learning Technologies 7(3).
5. Pang, B. and L. Lee, 2005. Seeing Stars: Exploiting Class Relationshipsfor Sentiment Categorization with Respect to Rating Scales, Proc.43rd Ann. Meeting on Assoc. for Computational Linguistics (ACL), pp: 115-124.
6. Barbosa, L. and J. Feng, 2010. Robust sentiment detection on twitter from biased and noisy data. In Proc. of Coling, 2010.
7. Davidov, D., O. Tsur and A. Rappoport, 2010. Enhanced sentiment learning using twitter hashtags and smileys, In Proceedings of Coling, 2010.
8. Liu, Bingwei, Erik Blasch, Yu Chen, Dan Shen and Genshe Chen, 2013. Scalable Sentiment Classification for Big Data Analysis Using Naive Bayes Classifier In Big Data, 2013 IEEE International Conference on, pp. 99-104. IEEE, 2013.
9. Richa, S., N. Shweta and J. Rekha, 2014. Opinion Mining Of Movie Reviews At Document Level, International Journal on Information Theory (IJIT), 3(3).
10. Vimalkumar B. Vaghela and Bhumika M. Jadav, 2016. Analysis of Various Sentiment Classification Techniques, International Journal of Computer Applications (0975-8887) Volume 140 – No.3, April 2016.
11. Devang Jhaveri, Aunsh Chaudhari and Lakshmi Kurup, 2015. Twitter Sentiment Analysis on E-commerce Websites in India, International Journal of Computer Applications, (0975-8887): 127(18).