International Innovative Research Journal of Engineering and Technology

ISSN NO: 2456-1983

International Conference on Emerging Innovation in Engineering and Technology

ICEIET-2017

Sentiment Analysis On Micro-blogs

S. Prasath Sivasubramanian[1],N. Suganya[2]

[1] Assistant Professor of Computer Science
[2]PG Student of Computer Science
Kanchi Mamunivar Centre for Post Graduate Studies (Autonomous)
Lawspet, Puducherry
[1]mail2prasath@gmail.com,[2]sugan16695@gmail.com

Abstract— Online Micro-blogging used for finding opinions about certain entity in very short messages. Millions of users share their opinions about different content. Some of the micro-blog names are twitter, facebook, etc,. twitter is one of the most widely used micro-blogging site where people share their opinions in the form of tweets. Twitter is rich source for sentiment analysis. Sentiment analysis is a process of analyzing polarity of those opinions and categorize them into positive, negative and neutral. Here, SVM and lexicon based approaches are used to find the sentiment polarity of the given text.

Keywords— Dictionary based approach, Micro-blogs, Sentiment analysis, Support Vector Machine(SVM), Twitter**.**

I Introduction

Sentiment analysis is also called opinion mining. Sentiment analysis is a major part of Natural Language Processing(NLP) for analysing the web users review [1] . Micro-blogs is a network service, which provides the services to the web users and it allows users to post their opinions in the form of text. Now a days micro-blog websites has the huge amount of reviews. Micro-blogging websites have progress to become a source of varied kind of information. This is due to nature of micro[blogs on which people post real time messages about their opinions on a variety of topics, discuss current problem, complain, and express positive sentiment for products they use in daily life. Infact, companies manufacturing such products have started to poll these micro-blogs to get a sense of general sentiment for their product.[4] Many times these companies study user reactions and reply to users on micro-blogs. One challenge is to build technology to detect and summarize an overall sentiment. Micro-blogging is different from blogging as its content normally smaller in both total and actual file size.

Advantages of micro-blogging over traditional blogging:

- Developing content takes less time: The traditional blogs are takes time to complete our messages. Micro-blogs gives you the benefit of posting the information about the incident in a short time.

- Increases chances of frequent posts: micro-blogging involves the more frequent posts and shorter ones but traditional blogging involves exactly opposite less frequent post and longer.

- Share time sensitive or urgent information in an easier way: Huge numbers of the micro-blogging platforms have been made to be fast and easy to use. With the video, post, simple tweet, you can easily share to everyone on what's happening at this very moment.

- Communication with followers becomes easy and direct: In addition to communicate easily with greater short and frequent posts, micro-blogging platforms can be used to easily encourage and facilitate better interaction through liking, reblogging, tweeting , commenting and more.

- Convenient using with mobile and tabs: Micro-blogging gains too much of attention in present days and the main cause behind this is increasing trends of mobile browsing. It is difficult to consume, interact and write long and lengthy blog post in a tab or Smartphone that's why micro-blogging comes into play and provide small, easy and faster posts. Micro-blogging websites has a character limit from 140 to 200. Many number of micro-blogging websites are there. But twitter and face book are now considered as the most popular social networks and micro-blogging services. Twitter was created by Noah Glass, Biz Stone, Evan Williams and Jack Dorsey in March 2006 and launched in July 2006. Twitter has 310 Million monthly active user, 1 Billion Unique visits monthly to sites with embedded Tweets, 83% of active users are access through mobile application, consists of 3500 employees around the world, more than 35 offices across the world, 79% accounts are from outside U.S. , supports more than 40 languages and 40% employees of twitter are from technical background. All numbers approximate as of March 31, 2016. People post 600 tweet messages at every second in twitter. Sentiment analysis is on microblog text is quite difficult to analyze when compared to conventional text. Because micro-blogs has a lack of complete and correct sentence structures, misspelling words, poor grammatical structure, usage of abbreviations, and ill-formed words for example Haaaaaaaappy!!!!!, lol, OMG!, bcoz, tat. These are difficult to classify the sentiment polarity.

Micro blogging and more particularly Twitter is used for the following reasons:

1. Micro-blogging platforms are used by different people to express their opinion about different topics, which is valuable source of people's opinions.
2. Twitter contains an enormous number of text posts and it grows every data. The collected dataset can be very large.
3. Twitter's audience varies from regular users to celebrities, company representatives, politicians. It is possible to collect text posts of users from different social and interests groups.
4. Twitter's audience is represented by users from many countries.
5. The tweets have its own conventions. The following are some of the examples of tweets conventions.
   - tweet—tweets are limited to 140 characters. This is different from other lexicons. [6]
   - Hashtags ("#") "#BigData" and "#Sentiment" are hash tags. These are used to organize tweets on particular topics.[6]
   - URLs—the URLs are used to track the external sources.
   - Retweet ("RT")—this is the easiest and most common way to share someone else's content.
   - Emoticons (Smileys)—Smileys or emoticons are regularly used in tweets to express the user state of mind
   - Colloquial expressions—most of the users express their situation of mind in an abnormal way.

II Problem Definition

This research paper focuses on using twitter, the most popular micro-blogging platform which is used for sentiment analysis process. The tweets are important for analysis because tweets arrive at a high frequency. Machine learning algorithm and lexicon based approaches used to process the tweets under very strict constraints. Algorithms are support vector machine(SVM) and dictionary based approach. The purpose of this paper is to determine the people's reaction accurately by comparing the results of these two algorithms.

III Sentiment Analysis And Process

The figure 1. depicted below show the flow of operations of Sentiment analysis.

It includes the process data collection, preprocessing, feature extraction, sentiment classification and projection of output. Various steps that are used in preprocessing are: Tokenization, Data filtering, stop word removal and stemming.

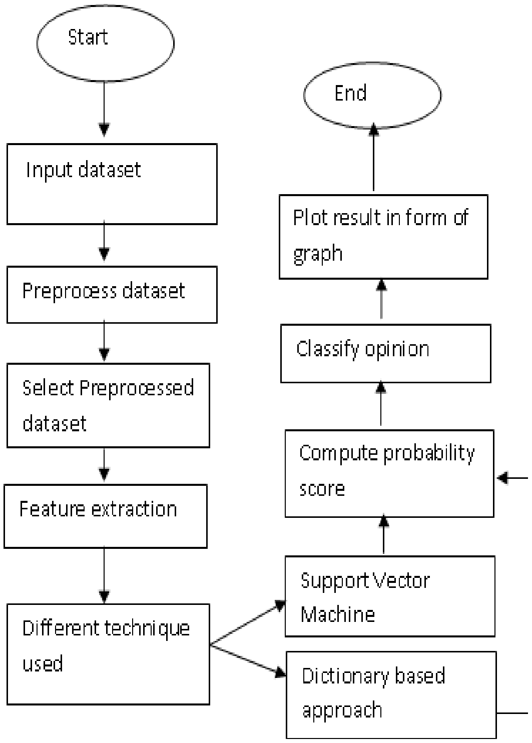

Figure 1 Sentiment Analysis Flow Chart

A. Data Collection

Sentiment analysis consists of collecting data from user generated content from blogs, forums, social networks These data are disorganized, expressed in different ways by using various vocabularies, user slangs, context of writing about the topic etc.. Manual analysis is almost impossible. Therefore, text analytics and natural language processing are used to extract and classify.

B. Pre-Processing

Pre-processing starts the text preparation into a more structured representation. This includes the following steps.

Tokenization: Tokenization is used to identify all words in a given text.

Data Filtering: People use a lot of casual language on twitter. For example, 'happy' is used in the form of 'haaaaaaappy'. Though this implies the same word 'happy', the classifiers consider these as two different words. To improve this and make words more similar to generic words, such sets of repeated letters are replaced by two occurrences. Thus haaaaappy would be replaced by haappy.

Stop Word Removal: It used to eliminate that words that occurs frequently such as article, prepositions, conjunction and adverbs. These stop words depends on language of the text in questions. For example, words like the, and, before, while, and so on do not contribute to the sentiment. Remove all URLs (e.g. www.xyz.com), hash tags (e.g. #topic), targets (@username). Remove all punctuations, symbols, and numbers. Expand Acronyms (we can use a acronym dictionary).

Stemming: In information retrieval, stemming is the process of reducing a word to its root form. For example, walking, walker, walked all these words are derived from the root word walk. Hence, the stemmed form of all the above words is walk` The raw data is pre-processed to improve quality.

C. Feature Extraction

Features in reviews are extracted so that it helps to know which feature has positive comment and which one has negative.

D. Sentiment Classification

In this step, subjective sentences are classified in positive, negative, good, bad; like, dislike, but classification can be made by using multiple points.

E. Presentation of Output

The main objective of sentiment analysis is to convert unstructured text into meaningful information. When the analysis is finished, the text results are displayed on different types graphs. Also time can be analyzed and can be graphically displayed constructing a sentiment time line with the chosen value (frequency, percentages, and averages) over time.

IV Sentiment Analysis Techniques

Sentiment analysis has two main techniques: machine learning based and lexicon based techniques. The machine learning approach(ML) applies the famous ML algorithms and uses linguistic features. The lexicon-based approach depends on a sentiment lexicon, a collection of known and precompiled sentiment terms. It is divided into dictionary based approach and corpus-based approach. This corpus-based approach used statistical or semantic methods to find sentiment polarity. [8]

The text classification methods using ML approach can be divided into supervised and unsupervised learning methods.

The lexicon –based approach depends on finding the opinion lexicon which is used to analyse the text. This method has two approaches. They are dictionary based approach and corpus based approach. Corpus based approach has two methods, they are statistical and semantic methods.

A. Machine Learning Approach

Machine learning approaches work by training an algorithm with a training dataset before applying it to the actual data set. Machine learning techniques first trains the algorithm with some particular inputs with known outputs so that later it can work with new unknown data. [5]

Support Vector Machine: It is a non-probabilistic classifier. It requires a large amount of training set. It is done by classifying points using a (d-1)-dimensional hyper plane. SVM finds a hyper plane with largest possible margin Support Vector Machines make use of the concept of decision planes that define decision boundaries. A decision plane is one that divide a set of objects having different class membership. The objects are mapped or rearranged using a mathematical function known as kernel and this is known as mapping or transformation. After transformation, the mapped objects are linearly separable and as a result the complex structures having curves to split the objects can be avoided. It works by plotting the training data in multidimensional space; it then tries to split the classes with a hyperplane. If the classes are not immediately linearly separable in the multidimensional space the algorithm will add a new dimension in an attempt to further separate the classes. It will continue this process until it is able to separate the training data into its two separate classes using a hyper plane.[2] A basic representation of how it splits the data is shown in figure given below.

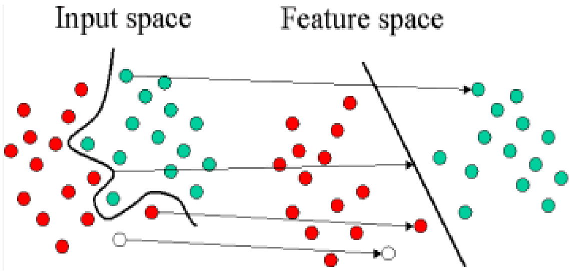

Figure 2 SVM Basic Operation

B. Lexicon based approach

Lexicon based approach [3] is an unsupervised technique, classification is done by comparing the features of a given text against sentiment lexicons whose sentiment values are determined prior to their use. Sentiment lexicon has lists of words and expressions used to express people's subjective feelings and opinions. There are many main approaches used to collect the opinion word list. Manual approach is very time consuming and it is not used alone. It is usually combined with the other approaches as a final check to avoid the mistakes.

Dictionary Based Approach: Dictionary based approach to compile sentiment words in an obvious approach because most dictionaries list synonyms and antonyms for each word. An example of that dictionary is WordNet, which is used to develop a thesaurus called SentiWordNet. This approach has a simple technique to use a few seed sentiment words to bootstrap based on the synonym and antonym structure of a dictionary. Initially a small set of sentiment words(Seed) are collected manually. The algorithm then grows this seed words by searching in the WordNet or another online dictionary for their synonyms and antonyms. Then newly found words are added to the seed list. The next iteration begins. The iterative process ends when no more new words can be found. After the process competes, finally a manual inspection step was used to check and correct the mistakes .[7]

V Implementation

Here, R programming language used for implementation. R language offers maximum support for sentiment analysis. The reason for using R language is, when the dataset is big, it is fast and efficient in terms of performance. This language has a wide scope of performing the analysis using SVM, by its rich set of libraries. The packages in the R tool are updated regularly and have greater number of probabilistic and statistical functions. The reason we are using R language is when the dataset is big, it is fast and efficient in terms of performing classification and feature extraction. The packages in the R tool are updated regularly.

In this paper the recent Demonetization process that takes place in India is centred and the twitter dataset for the demonetization was collected for doing sentiment analysis. Demonetization dataset consists of 80,000 tweets. After collecting the dataset, pre-processing steps is done on demonetization dataset. We remove the retweeted data sets and the dataset is reduced to 4000 in number. Then we have to create two separate dataset called positive and negative. , Later the dataset is compared with the text data. If the word is available on positive entries of the text means it considered as a positive text. If it is available in negative entries of the text means it is considered as a negative tweet. Otherwise it is considered as a neutral polarity. This implementation is done by using R with necessary libraries installed for analysing the micro-blogs. The result is computed and is projected in the figure shown below.

In the figure 3 (Graph) shown below shows almost 90% of the tweets falls on the neutral scale. The positive tweets spans almost 52% and 40% of the tweets falls on the negative scale. Though with this we can derive a conclusion about the impact created by the demonetization, the graph also shows details about false positive and false negative tweets. Both false positive and false negative impact are almost equal on the scale. The false positive and false negative impact can be studied better using specialized algorithms like lexicon or dictionary based approaches.

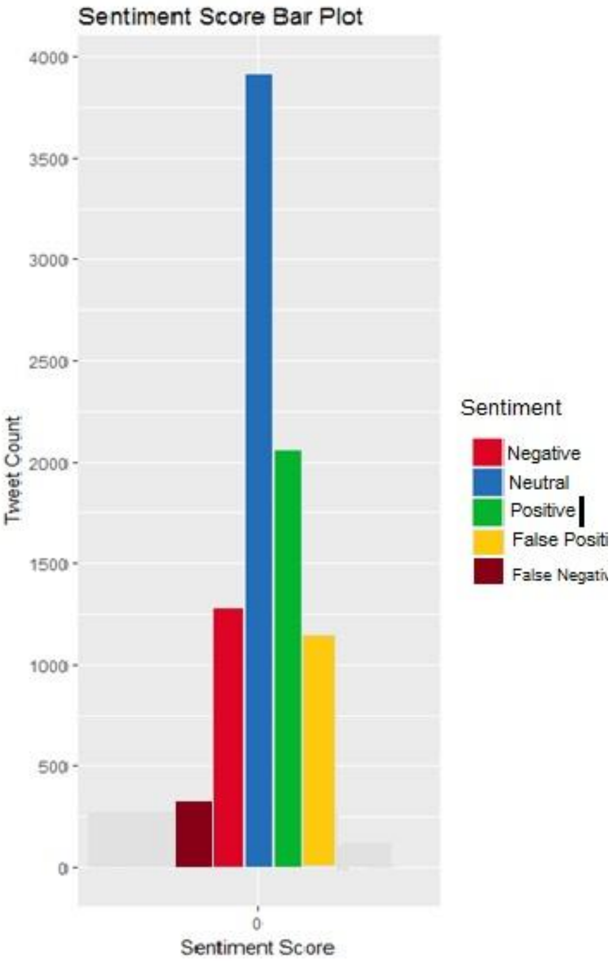

Figure 3 Graph Showing The Impact Of Tweets On Positve, Neutral , Negative , False Positive And False Negative Scale

VI CONCLUSION AND FUTURE WORK

Applying sentiment analysis to mine the huge amount of unstructured data is the primary focus of major organization and researchers are putting effort to find the best system for sentiment analysis. Demonetization is one of the current debate topics among the micro-blogs. Government of India have interested to find the answer to the question, by analyzing the micro-blogs to understand the impact and feeling of the people of India on their demonetization move?. Demonetization dataset used to find out the view of different people on the demonetization by analyzing the micro-blogs from twitter.This work can be further developed by taking the impact of false positive and false negative tweets by using the Lexicon and Dictionary based approach. These two algorithms are used to predict the polarization of tweets. Finally comparison can be made by using the output which is produced by these two algorithms.

ACKNOWLEGEMENTS

The authors sincerely acknowledge all who has contributed directly or indirectly for the resources and literature used in this paper.

REFERENCES

[1] Alessia D'Andrea, Fernando Ferri, Patrizia Grifoni, Tiziana Guzzo "Approaches, Tools and Applications for Sentiment Analysis Implementation". International Journal of Computer Applications(0975-8887), Volume 125-No.3, September 2015.

[2] Ana s Collomb, Crina Costea , Damien Joyeux, Omar Hasan ,Lionel Brunie "A Study and Comparison of Sentiment Analysis Methods for Reputation Evaluation" 2016.

[3] Devika M D, Sunitha c, Amal Ganesh "Sentiment Analysis:S Comparative Study On Different Approaches". Fourth ``International Conference on Recent Trends in Computer Science & Engineering. 2016.

[4] Fröhlich, B. and Plate, J. "The cubic mouse: a new device for three-dimensional input", In

Proceedings of the SIGCHI Conference on Human Factors in Computing System,2000

[5] K.S.Ilakiya1, Mrs. M.Lovelin Ponn Felciah2 Challenges and techniques for sentiment analysis: a survey - IJCSMC, Vol. 4, Issue. 3, March 2015, pg.301 – 307

[6] Sannella, M. J, "Constraint Satisfaction and Debugging for Interactive User Interfaces" Doctoral Thesis. UMI Order Number: UMI Order No. GAX95-09398., University of Washington,1994

[7] S. Vohra, Prof. J. B Teraiya "A Comparative Study Of Sentiment Analysis Techniques". Journal Of Information, Knowledge And Reseach Computer Engineering

[8] Walaa Medhat, Ahmed Hassan, Hoda Korashy " Sentiment Analysis Algorithms and Applications: A Survey**".** Ain Shams Engineering
Journal. December 2014

S.PRASATH SIVASUBRAMANIAN,

Assistant Professor of Computer Science

Kanchi Mamunivar Centre for PG.Studies, Puducherry. A College with Potential for Excellence, ReAccredited by NAAC with "A" grade.

Having a Teaching experience of nearly 21+ years and a research experience of 7 years, presently working the PG Department of Computer Science. Obtained M.Tech from Pondichery University and is pursuing Ph,D currently.

Area of Research interest includes, Big Data Analytics, SOA and Adhoc Networks.

N. SUGANYA-PG.Student of Computer Science

Kanchi Mamunivar Centre for PG.Studies, Puducherry.

A College with Potential for Excellence, ReAccredited by NAAC with A grade.

Presently studying Final year Master Degree in Computer Science and is currently working on Data Analytics for Final Semester Project work. Her area of interest includes Analysing Microbogs, Topic Modelling.