TWITTER DATA SENTIMENT ANALYSIS AND WORD CLOUD

[1]Dr.R.Vijayabhanu, Assistant professor, Department of Computer Science, Avinashilingam Institute for Home science and Higher education for women, Coimbatore, India
[2]R.Shenbagam, PG Student, Department of Computer Science, Avinashilingam Institute for Home science and Higher education for women, Coimbatore, India

***Abstract*-** Social networking plays a significant role in our day-to-day life by interacting around the world about any business and social products or services. The opinions, suggestions and the reviews on various disciplines on social mediums share a remarkable place in marketing and promoting. Sentiment analysis or opinion mining is the study of public opinions, sentiments, attitudes, and emotions that are expressed by the people and a dynamic area in natural language processing and text mining in current years. Being a huge social website, Twitter consists of large users posting and disputing their information. One such trending domain like Politics has been analyzed in this paper. Sentiment analysis is a chain of various processes such as data mining and machine learning. In this paper, sentiment analysis has been carried out on the politics dataset by sentiment scoring, classification and performance evaluation.

***Keywords*: -** Sentiment Analysis, Opinion Mining, Natural Language Processing (NLP), Lexicons, Polarity Classification.

1. Introduction

Being an encouraging and active platform of interactions, Social Media becomes a library of volumes of various trending subjects revolving around the globe. One such application where people are free to post their opinions and suggestions about a commodity or have their comments on any discipline of public interest is Twitter. Around 200 million tweets per day are posted by the users as by communicating their thoughts on hot topics independently.

According to a couple of survey reports conducted among 2000 people had proved that about 81% of online users research at least once about a commodity and between 73% and 87% have reported that a major impact on the purchase of a product relies on the internet reviews [1]. Therefore, Sentiment analysis systems are being useful in almost every business and social domain because opinions are important to almost all communal activities which also a fundamental influencing factor of behaviours. Decision making is a complicated tasks, which rather be simplified with others' opinions. Sentiment Analysis is the computational handling of opinions, sentiments and subjectivity of text. But while managing Natural Language Processing (NLP), the problems such as Sentence boundary detection, Tokenization, Part-of-Speech assignment to individual words, Homographs, etc are include [2]. Thus, text mining a non-traditional information retrieval strategy to reduce efforts in obtaining information from large set of textual documents is required. The R, an open source programming language for statistical computing and graphics widely used and best known among statisticians and data miners for developing statistical software and data analysis can been used in such occasions [3].

In this paper, the sentiment analysis has been initiates with data mining process. The dataset has been pre-processed and subjected to sentiment scoring and sentiment classification. At last the performance evaluation has been achieved on polarity classification. To implement this scheme, Politics Dataset extracted from a Twitter account with 8000 tweets and 14 attributes has been used.

2. Literature Review

Study of Twitter Sentiment Analysis using Machine Learning Algorithms on Python [4] proposed by Bhumika Gupta *et al.* analysis the opinions of the Twitter users by combining various techniques and methodologies such as data mining and machine learning. An effective interrupted programming language like Python has been used for pre-processing the dataset. Scikit-learn, a Google Summer Code enhances the system with various machine learning classification algorithms and other tools of mining and analysis. The machine learning model broken up by the training data set provides a model of trained machine learning which has been further intermittent by the testing set from the original data set to provide a conclusive classified data. Thus, the paper enhance the sentiment analysis on the Twitter data with various sentiment classifiers and categories their result accurately.

Chetan Kaushik *et al.* has researched a paper on **A Scalable, Lexicon Based Technique for Sentiment Analysis** [5] by experimenting on sentiment analysis by using lexicon based technique rather than choosing a conventional text mining approaches. An expanded data set of around 6,74,412 tweets are studied using the Hadoop, an open source software. Being a flexible processing and analyzing technology, the Hadoop has been used in predictive analysis of both constructive and non-constructive data. The data set has been approached using sentiment lexicon or dictionary, thereby allocating the polarities of the words in their respective domains. To

optimize the time consumption, the feature detection of the subject towards the sentiments has been performed by using the Twitter hash tags. The negative and blind negations are focused during the polarity of the sentiments for the betterment of the output. Therefore sentiment calculations have been executed on each and every tweet thereby calculating the polarity score. Thus, the paper claims 73.5% of accuracy performed in 14.8 seconds on 6,74,412 tweets. Though the accuracy detected has not been comparative with the machine learning techniques but it has been better than other lexicon based approaches by consuming less time.

Sentiment Analysis on Twitter data [6] by Varsha Sahayak *et al.* describes a paper on sentiment analysis by using data miming and machine learning methodologies. The sentiments are categories by polarity of the words such as positive, negative or neutral. The lexicon-based approach of the proposal has been based on both the corpus-based and dictionary-based techniques. The initial step involves data mining of the tweets like extracting and pre-processing the data set. The extracted tweet data are classified by using machine leaning approaches such as Naive Bayes classifier or Support Vector Machines (SVMs). Thereby has been used as the training data set, the process of classification takes places followed by the sentiment scoring. Thus, the opinion mining has been executed automatically by classifying the sentiments of tweets obtained from Twitter data set.

The paper **Sentiment Analysis on Twitter** [7] directed by Akshi Kumar *et al.* demonstrates a novel approach of opinion mining by combining both the corpus-based and dictionary-based algorithms. The tweets from the Twitter dataset are pre-processed by using data mining techniques to extract the sentiments which are subjected to the scoring module. The semantic score of the sentiment carriers i.e. the parts of speech such as adverbs, verbs and adjectives are analyzed by the novel approach. The verbs and adverbs are scrutinized by the dictionary-based method while the adjectives are analyzed by the corpus-based method. A linear equation of letters, adjective groups, exclamations, emotions, etc has been used to calculate the tweet sentiment score respectively and to obtain the opinions on the Twitter data set.

Walaa Medhat *et al.* have studied a paper on **Sentiment Analysis Algorithms and Applications: A Survey** [8]. The paper illustrates a complete synopsis on Sentiment Analysis (SA) or Opinion Mining (OM). The classification in sentiment analysis can takes place either in following three ways such as document-level, sentence-level and aspect-level. The paper briefly discuss about both the machine learning and lexicon-based approaches. The feature selection during sentiment classification has been studied well. The illustration based on the number of articles that has been targeted for the sentiment analysis with percentages invokes the better understating of the subject. Various problems that might frequently occur during the opinion miming have been focused. Therefore, the survey paper has been played its importance in empowering the knowledge of Sentiment Analysis (SA).

3. Proposed System

The lexicon-based sentiment analysis consists of Dictionary-based and corpus-based approaches. The lack of finding the opinion words with respect to domain and context specific orientation is the major drawback of dictionary-based approach which luckily can be overwhelmed by the corpus-based approach. A collection of known and pre-defined sentiment terms are recognized as Sentiment Lexicons plays a vital role in lexicon sentiment analysis.

The proposed system consists of four phases:-

a. **Data mining / Text mining**: Pre-processing the raw data to extract the knowledge.

b. **Sentiment Score**: Loading of the predefined lexicons and calculating the positive / negative score by comparing the tweets terms with positive/negative term corpus and summing the occurrence count.

c. **Sentiment Classification**: The essential phase of organizing the text according to the opinion's; sentiment polarities (positive, negative, neutral) [9]. The pre-processed tweets are taken as input to sentiment classification which can be done by using two lexicons such as emotion lexicon and subjectivity lexicon. The emotion lexicon is used to classify the emotions, such as joy, fear, sadness, disgust, surprise, anger and unknown while the subjectivity lexicon is used to classify the polarity such as positive, negative and neutral.

d. **Performance evaluation:** Compares both the values of sentiment score and lexicon based algorithm and visualizes in the form of graph.

The system has been implemented in the R Studio which has been consider to be the best at text mining and pre-processing and it helps in the improvement of the system accuracy [10]. The systematic flow has been illustrated in the Fig1.

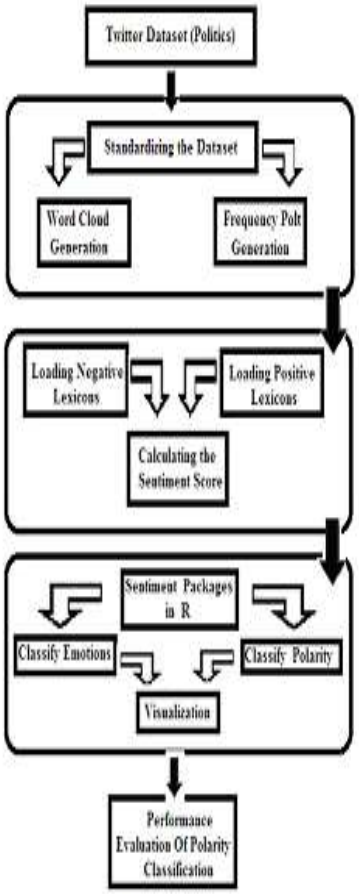

Fig 1. System Flow Diagram

4. Module Description

4. 1. Load and standardizing the dataset

The raw data has been pre-processed to extract the usable dataset are loaded into the R studio as shown in the Fig 2.

Fig 2. Top Trending Tweets Today

Text mining methodologies allows in highlighting the most frequently used keywords in a paragraph referred as the word cloud or the text cloud or the tag cloud, which is a visual representation of text data. In this work, the word cloud has been generated for Politics Dataset in which the most frequently used spot words are displayed in larger size and bolder format as shown in the Fig 3.

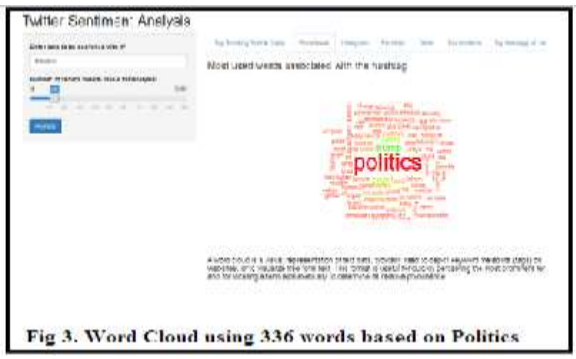

Fig 3. Word Cloud using 336 words based on Politics

A graphical representation of patterns in a set of data by plotting how often particular values of a measure occur is a frequency plot. The horizontal axis of a frequency plot graph shows grouping of a continuous measure (e.g. age, time, no. of occurrences) while vertical axis shows the number of times that a value in that group was seen. Here, the frequency plot has been generated for larger words in the word cloud generation.

4.2. Sentiment Score

The positive and negative lexicons are downloaded using the available links and loaded in R studio**.** The corpus contains around 6789 words. The positive / negative score are calculated by comparing the tweets terms with positive/negative term corpus and summing the occurrence count. The histograms of positive sentiments and negative sentiments are demonstrated in the Fig 4 and Fig 5 respectively.

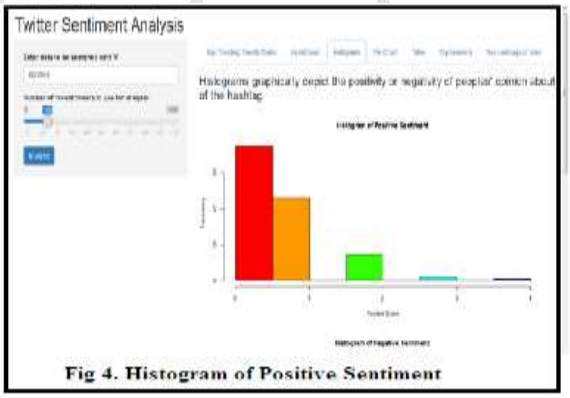

Fig 4. Histogram of Positive Sentiment

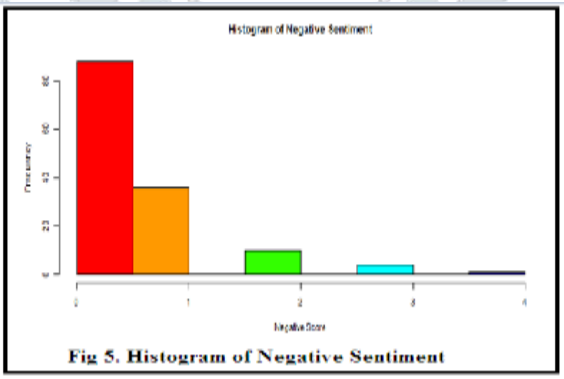

Fig 5. Histogram of Negative Sentiment

The sentiment score is the more precise numerical representation of the sentiment polarity. It calculates the score for each tweet. The more number of words that matches the positive list from the tweet gives positive score and the number of words that matches the negative list from the tweet gives negative score which has been explained in the Fig 6.

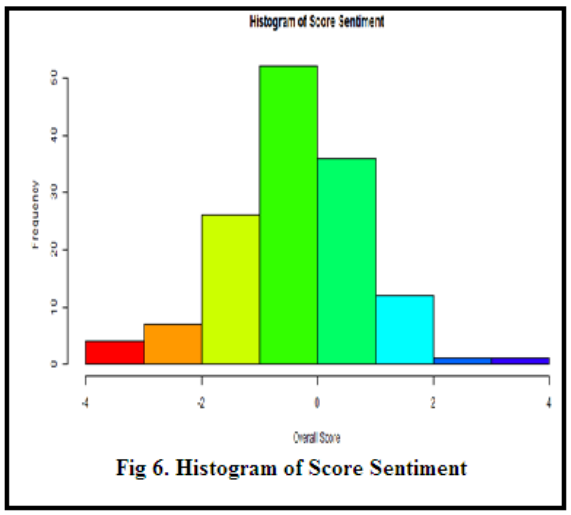

Fig 6. Histogram of Score Sentiment

4.3. Sentiment Classifications

R package sentiment by Timothy Jurka has a function that helps us to analyze some text and classify it in different types of emotions: anger, disgust, fear, joy, sadness, surprise and unknown. The emotion classification can be performed using the algorithm called lexicon based sentiment algorithm trained on Valitutti's emotion lexicon. Another function from sentiment package named classify polarity allows us to classify some text as positive or negative. In this case, the polarity classification can be done by using a lexicon based algorithm trained on Janyce Wiebe's subjectivity lexicon. The following Fig 7 indicates the pie chat of negative and positive sentiments.

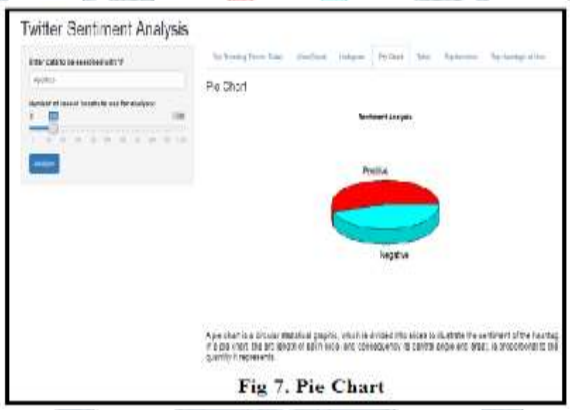

Fig 7. Pie Chart

4.4. Performance Evaluation

While the performance evaluation on polarity classifications, both the values of sentiment score and lexicon based algorithm are compared and visualized in the form of excel sheets. By the performance evaluation and result shown, the lexicon based sentiment algorithm accuracy has been better than the sentiment score accuracy.

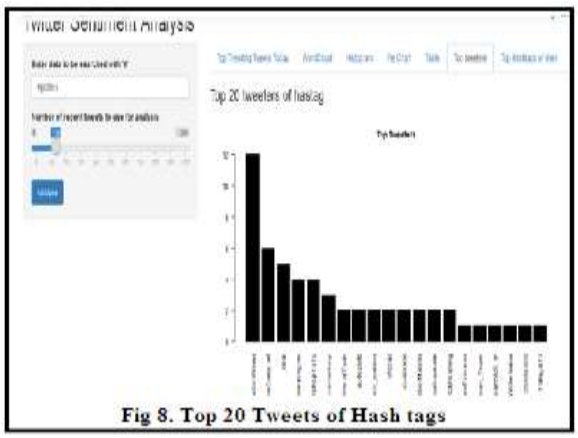

Fig 8. Top 20 Tweets of Hash tags

The above Fig 8 is a bar chart representation of top 20 tweets of hash tags and the following Fig 9 contains the Top 20 Tweeters.

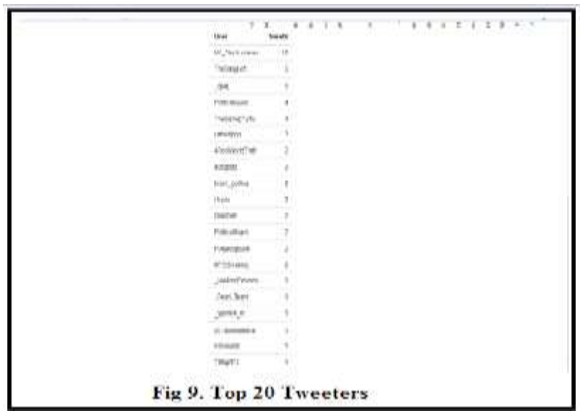

Fig 9. Top 20 Tweeters

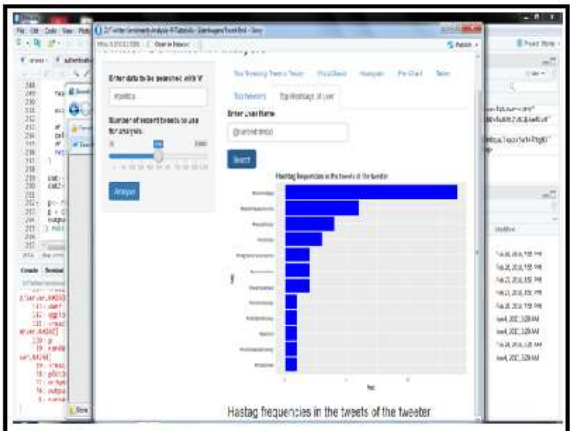

Fig 10. Hash tag Frequency in the Tweets of the Tweeter

Hash tag frequency in the tweets of the tweeter has been visualized in the Fig 10.

5. Conclusion

People are using social media on daily basis to communicate their opinions on a wide variety of subjects, products and services, which has made it a rich resource for text mining and sentiment analysis. The sentiment classification done by using two lexicons such as emotion lexicon and subjectivity lexicon are visualized in the form of bar charts for better understanding. The performance evaluation of polarity classifications by comparing both the values of sentiment score and lexicon based sentiment algorithm are visualized in the form of graphs has enhances the system accuracy which has been consider better than the sentiment score accuracy.

References

[1] Opinion Mining and Sentiment Analysis, B. Pang, L. Lee, Foundation and Trends in Information retrieval, Vol 2, No 1-2, 2008, Pp: 1-135.

[2] Natural Language Processing: An Introducyion, Prakash M Nadkarni, Lucila Ohno-Machado, Wendy W Chapman, J Am Med Inform Assoc, 18, July 2011, Pp: 544-551.

[3] Txt Mining in R, Kasper Welbers, Wouter Van Atteveldt, Kenneth Benoit, Communication Methods and Measures, Vol 11, No 4, Nov 2017, Pp: 245-265.

[4] Study of Twitter Sentiment Analysis using Machine Learning Algorithms on Python, Bhumika Gupta, Monika Negi, Kanika Vishwakarma, Goldi Rawat, Priyanka Badhani, IJCA, Vol 165, No 9, May 2017, Pp: 29-34.

[5] A Scalable, Lexicon based Technique for Sentiment Analysis, Chetan Kaushik, Atul Mishra, IJFCST, Vol 4, No 5, Sept 2014, Pp: 35-43.

[6] Sentiment Analysis on Twitter data, Varsha Sahayak, Vijaya Shete, Apashabi Pathan, IJIRAE, Issue 11, Vol 2, Jan 2015, Pp: 178-183.

[7] Sentiment Analysis on Twitter, Akshi Kumar, Teeja Mary Sebastian, IJCSI, Vol 9, Issue 4, No 3, July 2012, Pp: 372-378.

[8] Sentiment Analysis Algorithms and Applications: A Survey, Walaa Medhat, Ahmed Hassan, Hoda Korashy, ASEJ, May 2014, Pp: 1093-1113.

[9] Sentiment Classification and Polarity Shifting, Shoushan Li, Sophia Yat Mei Lee, Ying Chen, Chu- ren Huang, Guodong Zhou, ICCL, Beijing, Aug 2010, Pp: 635-643.

[10] Text Mining Scientific Articles using the R Language, Carlos A. S. J. Gulo, Thiago R. P. M. Rubio, DSIE 1st Edt, 2015, Pp: 60-69.