Twitter Data Analysis using R

Jaichandran. R, Karthik .K, S.Rajaprakash, H.Dhrithimanth Prabhu, Yerasani Aditya Gowtham

***Abstract:** Big Data is an emerging concept in the field of Data mining. It has numerous applications in real life. Most data are coming from social media networking Websites comprising of structured and unstructured data including Text, video, images etc. The main characteristics can be understood by five v's. Twitter is one among the major evolving social media. Twitter Data analysis can be give you a wide perspective of public opinion regarding any product, public opinion etc which can be used to mine the knowledge from the data. For example, prediction analysis, product review, favourite among people tweets about GST (Goods and service tax).*

***Keywords:** Big data; twitter, R-studio, GST..*

I. INTRODUCTION

Twitter is a most popular social networking site for fast communication. More than 140 million active users publish over 400 million of 140 characters "Tweets" every day. Twitter's speed and ease of publication have made it an important communication medium for people from all walks of life. Twitter has played a prominent role in socio-political events [7]. Twitter has also been used to post damage reports and disaster preparedness information during large natural disasters, such as the Hurricane Sandy. It is the fastest medium to share the people opinion, mood and messages. It is used by numerous people including celebrities, sports star, politician, actor etc. Every people have to register with Twitter to post his opinions. Unregistered people can only view the tweets. Social media provides the valuable information regarding prediction, product review, popularity, feedback sentiment analysis etc [8]. Tweets are generated all over the world by the people from different religion, caste, habits or country. This paper is for the reader who is interested in understanding the basics of collecting, storing, and analyzing Twitter data using R. This paper discuss about how to collect Twitter data, looking at the free APIs provided by Twitter. We then go on to discuss how to store this data in a tangible way for use in real-time applications. We also focused on common measures and algorithms that are used to analyze social media data. We finish the analysis by discussing visual analytics. In this paper, a system is proposed to analyze the sentiment of people regarding introduction of GST in India.

Manuscript published on November 30, 2019.
* Correspondence Author

R.jaichandran S*, Dept. of Computer Science and Technology, Aarupadai Veedu Institute of Technology, Vinayaka Missions Research Foundation Chennai, India. rjaichandran@gmail.com

K. Karthik, Dept. of Computer Science and Technology, Aarupadai Veedu Institute of Technology, Vinayaka Missions Research Foundation Chennai, India., karthik@avit.ac.in

S.Rajaprakash, Dept. of Computer Science and Technology, Aarupadai Veedu Institute of Technology, Vinayaka Missions Research Foundation Chennai, India., srajaprakash_04@yahoo.com

Yerasani Aditya Gowtham Final year CSE, Aarupadai Veedu Institute of Technology, Vinayaka Missions Research Foundation Chennai, India

H.Dhrithimanth Prabhu H Final year CSE, Aarupadai Veedu Institute of Technology, Vinayaka Missions Research Foundation Chennai, India

II. LITERATURE REVIEW

Twitter Sentiment analysis is the recent trend in the field of Data mining. It uses the natural language processing system and Artificial Intelligence. Sentiment of people can be measured using scale broadly divided into positive, neutral and negative which can be represented as +1,0,-1 respectively. The Result can be figured using pie chart. In sentiment analysis, most popular thing is data mining.usage of text data is increases on social media. According to Abhishek Kesharwani et al., twitter is one of the most popular social media site used by the people to express their view. Such site generates many unnecessary and unrelated data. By using sentimental analysis we can easily filter the sentiment of the people. According to Ali Hasan et al., twitter is the platform where millions of tweets from many users are generated in huge amount. For analysing the sentiment of people R studio is very good open source platform. Ankit Gupta et al., proposed that Twitter sentiment analysis on demonetization could be accomplished using n-grams and sentiment classification score.in this paper they had classified the tweets into bi grams(n=2) and tri grams (n=3) and predict the probability of future word by only looking at the last word encountered. The sentiment classification scores had been provided by two types of polarity method: first Polarity score for sentence level, second polarity score for paragraph level. The polarity context had been divided further into neutral, negative and positive using R studio. Arvind Singh Raghuwanshi et al., proposed naïve based classifier algorithm for text classification and provides less computational time it mainly works on conditional probability theory.They collected the data for 6 days and calculate the range, term frequency, cumulative frequency and Zipf score on various fields such as funding, industry, income, Sensex. these data are analyzed and visualize using ggplot in R studio. They also calculated the sentiment score chart using five different excel files. Christian Nwankwo et al., proposed about predicting stock prize and sentiment co-relation using n-grams and word2vec model representation. They had shown strong correlation between stock prices of the company to the public through tweets on Twitter using R studio.They had proposed that about analysis of product twitter data through opinion mining they collected the data in JSON format and perform data cleaning and extract the required data using supervised machine learning techniques like naïve baize (NB) maximum entropy (ME) and support vector machines.

The proposed systems determine the opinion of the product and the system compares with old training old training sets if the product is updated. It classifies the tweets based on score which lies between -1 to +1 and visualize using graphs.

III. SYSTEM DESIGN

Figure 1 illustrates the methodology in analysing the twitter data.

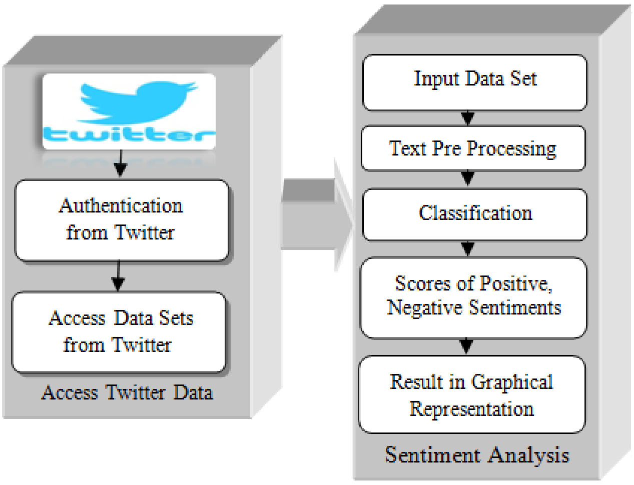
twitter
Authentication from Twitter
Access Data Sets from Twitter
Access Twitter Data
Input Data Set
Text Pre Processing
Classification
Scores of Positive, Negative Sentiments
Result in Graphical Representation
Sentiment Analysis

Fig. 1: Methodology

The various steps in this methodology can be described as below:

Twitter Authentication: It is the first process before mining any data from Twitter. Twitter provides a Twitter API to authenticate the user using twitter application. User have to create an application to connect with Twitter and streaming the data from it [9]. It provides the consumer key, consumer secret key access token and access token secret which can be used to establish the connection with Twitter and R-studio. These secret key are unique to the user and it is passed as a argument in setup_twitter_oauth() function. Three-way handshake is accomplished between user and twitter. Now it is ready to stream any data from the twitter.

```
    consumer_key    <- ‘xxxxxxxxxxxxx’
    consumer_secret <- 'xxxxxxxxxxxxx’
    access_token     <- ‘xxxxxxxxxxxxx’
    access_secret     <- ‘xxxxxxxxxxxxx’
    setup_twitter_oauth(consumer_key,consumer_secret,acces
s_token,access_secret)
```

Streaming Twitter Datasets: Twitter authentication requires a 6 digit pin during authentication provided by the twitter. This pin is generated for every transaction between twitter and R-studio application. Using this pin, twitter hashtags can be fetched and ready to mining the knowledge from the data. In this regard, we use searchTwitter() function to access the data and stored in the text format which can be shown as -

```
searchTwitter('GST',n=max,lang = 'en')
```

Text Processing: Now we have collected the data from the twitter using TwitterApi. It's time to filter the data by removing the unnecessary details of hashtag which is called as noise reduction from tweets such as punctuation, HTML tag, white space, numbers etc [10]. It can be done step by step as follows:

(i)Replacing Emoticons: An emotion is a pictorial representation of facial expression characters, usually punctuation marks, numbers, letter. Since, Twitter allows only limited numbers of words so user used emoticons to express the feeling and emotions using this icon. In this process, all the emoticons are replaced by the respective words in the tweets.

(ii) URL and Hashtags: The tweets generally contain hashtag to highlight the words.User used this because of limitation of words in the post. It must be processed to analyse the tweets in more appropriate manner.

(iii) Lowercase: Tweets contain both the cases of letters that is uppercase and lowercase. It gives the uneven meaning of the words such as TwITtEr DaTA. It is desirable case during mining of tweets so it must be converted into lowercase to reduce the anomaly.

(iv) Tokenization: It is needed in feature selection of the sentence. So all the tweets is converted into small tokens which is further processed by sentiment analyzer to rate the score of tweets.

(v) Stemming: It removes the prefix of the words containing etc. It is further reduces the complexity of analyzer to analyze in a proficient manner.

(vi) Stop Words: Stop words are the words which contains no meaning if it is present in the tweets such as like, is, are, the etc related to conjunction and preposition. so it is necessary to remove the stop words from the tweets to reduce the complexity of the sentence and analyze in smooth manner.

Sentiment Analysis and polarity classifier: The data we have collected is preprocessed and ready to analyse by the sentiment analyzer. The analyzing of data is categorized into two approaches that is Dictionary based approach and Machine Learning. In this paper we have dictionary-based approach to analyze the tweet sentiment using NLTK toolkit. Sentiment analysis and polarity classification is difficult task. In this case, we will use sentiWord net. It is a standard dictionary of positive, negative and neutral word used by various organisation for sentiment analysis. Polarity classifier will classify the words having positive, negative and neutral reviews. It can be demonstrated below in the next section of this paper.Visualization: The result of Sentient Analyzer can be visualize using various packages of R-studio such as word clouds, ggplot2, barChart, histogram etc. These tools can be used to visualize the result of Twitter Data which is shown in Result section..

IV. EXPERIMENTS AND RESULTS

The output of the analyzing twitter data of GST can be depicted by the figure 2 GST opinion analysis on Twitter (pie chart). It can be seen clearly in the chart that 10.9% people opinion is positive regarding the introduction of GST in India.

85.5% people does not support the GST implementation and 3.6% people opinion is neutral regarding GST.

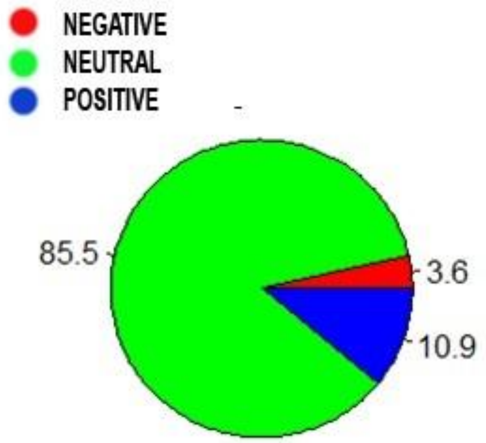

Fig. 2: GST Opening Analysis on Twitter

From the figure 3 Word cloud is generated using R-studio package which visualize the maximum number of frequency in the tweets. It can be used to find the most and least popular dataset in the tweets which can be shown below. It can be analyzed that the most popular dataset is gst, world, bank, second tax, rate and etc

.

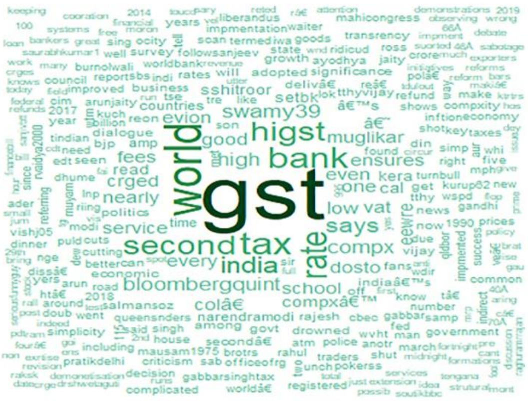

Fig. 3: Word Cloud for GST

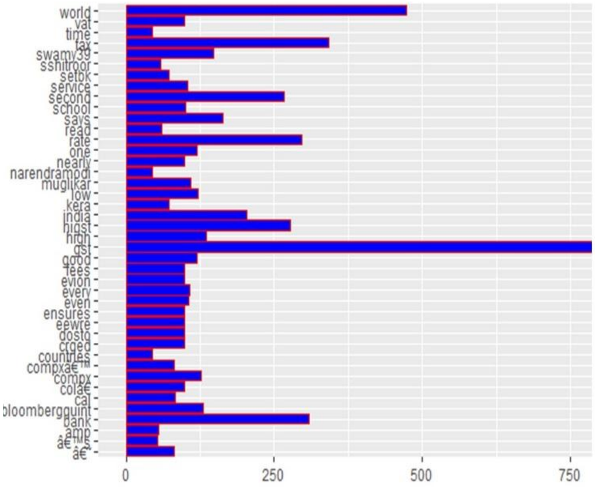

Fig. 4: Frequent terms in GST Analysis

From figure 3 we can conclude that the number of frequent terms are analyzed using R-studio. It can be used for finding the most frequent terms are used by the users through tweets. From the plot we can say that the terms gst, world, tax, rate and etc.. are the mostly used by the users generated from the different places around the globe. Fig. 4: Frequent terms in GST Analysis From figure 3 we can conclude that the number of frequent terms are analyzed using R-studio. It can be used for finding the most frequent terms are used by the users through tweets. From the plot we can say that the terms gst, world, tax, rate and etc.. are the mostly used by the users generated from the different places around the globe.

V. CONCLUSION

This paper analysed Twitter datasets on GST. It gives a wide perspective on the public opinion on the GST. Based on the result, we can conclude that large number of people has welcomed the new tax program launched by Indian government. In this process, text cleaning, sentiment analyser, polarity classifier is used to complete the sentiment analysis. It is a sample public opinion on GST. It can be-be implemented in a wide range of topics under Big Data.

REFERENCES

1. Abhishek Kesharwani, and Rakesh Bharti, “Movie Rating Prediction Based on Twitter Sentiment Analysis”, Journal of Advanced Computing and Communication Techniques, Vol. 5, Issue 1, pp. 6-10, 2017.
2. Ali Hasan, Sana Moin, Ahmad Karim, and Shahaboddin Shamshirband, “Machine Learning-Based Sentiment Analysis for Twitter Accounts”, Mathematical and Computational Applications, vol. 23, Issue 11, pp. 1-15, 2018.
3. Ankit Gupta, Jyotika Pruthi, and Neha Sahu, “Sentiment Analysis of Tweets using Machine Learning Approach”, International Journal of Computer Science and Mobile Computing, vol. 6, Issue. 4, pg. 444-458, 2017.
4. Arun K, Srinagesh A, Ramesh M, “Twitter Sentiment Analysis on Demonetization Tweets in India using R Language”, International Journal of Computer Engineering in Research Trends, Vol. 4, Issue 6, pp. 252-258, 2017.
5. Arvind Singh Raghuwanshi, Satish Kumar Pawar, “Polarity Classification of Twitter Data Using Sentiment Analysis”, International Journal on Recent and Innovation Trends in Computing and Communication, Vol. 5, Issue 6, pp.434-439, 2017.
6. Christian Nwankwo, Hayden Wimmer, Jie Du, “Predicting Movie Rating using Sentiment Analysis of Tweets”, Midwest DSI Annual Conference Grand Rapids, Michigan, pp. 68-79, 2017.
7. N.M.Dhanya and Harish, U.C., “Sentiment Analysis of Twitter Data on Demonetization using Machine Learning Techniques”, Lecture Notes in Computational Vision and Biomechanics, Vol. 28, pp. 227-237,2018.
8. Pappu Rajan A, and Victor S.P “Web Sentiment Analysis for Scoring Positive or Negative Words using Tweeter Data”, International Journal of Computer Applications, Vol. 96, no. 6, pp. 33-37, 2014.
9. Rupawari Jadhav, and M.S. Wakode, “Survey: Sentiment Analysis of Twitter Data for Stock MarketPrediction”, International Journal of Advanced Research in Computer Communication Engineering, Vol. 6, Issue 3, pp. 558-562, 2017.
10. Smailović J., Grčar M., Lavrač N., Žnidaršič M. (2013) Predictive Sentiment Analysis of Tweets: A Stock Market Application. In: Holzinger A., Pasi G. (eds) Human-Computer Interaction and Knowledge Discovery in Complex, Unstructured, Big Data. Lecture Notes in Computer Science, vol 7947. Springer, Berlin, Heidelberg.

AUTHORS PROFILE

Dr.R.Jaichandran is currently working as Head of the department of CSE in Aarupadai Veedu Institute of technology an ambit institution of Vinayaka Missions Research Foundation (Deemed to be University), Tamil Nadu, India. He has 13 years of experience in academics, industry, research, anddevelopment activities. Published 33 research papers in referred Journals and Conferences. His area of Interest includes Wireless Sensor Networks, Internet of Things (IoT), Ethical Hacking, Big data Analytics, and Embedded systems. He has delivered 33 Special lecturers in various reputed organizations in topics like Ethical Hacking, Mobile Phone Hacking, Big data Analysis, Internet of Things (IOT), Cloud Computing, Networking etc. Attended Seminars/Workshops/Faculty development programs conducted by various reputed Organizations. Received grants from reputed organizations like Tamil Nadu State Council for Science and Technology, and Computer Society of India.He has peer Reviewed Manuscripts in reputed international Journals and Conferences. He is a member in following professional societies: International Association of Computer Science and Information Technology (IACSIT), Association of Computer Electronics and Electrical Engineering (ACEEE), International Association of Engineers (IAENG), Computer Society of India (CSI),Indian Society of Technical Education (ISTE)..

CSI and ISTE.

Dr.S.Rajaprakash M.sc, M.Phil M.E Ph.D. currently working as Associate professor of CSE in Aarupadai Veedu Institute of Technology an ambit institution of Vinayaka Missions Research Foundation (Deemed to be University), Tamil Nadu, India. He has 18 years of experience in academics, research, and development activities. Published 17 research papers in referred Journals and Conferences. His area of Interest Artificial Intelligence, Computational Intelligence, Discrete Mathematics and Automata theory. .Received grants from Tamil Nadu State Council for Science and Technology .He has peer Reviewed Manuscripts in reputed international Journals and Conferences. He is a member in following professional societies: CSI and ISTE and Ramanujam Mathematical Society.

Mr. K.Karthik ME (Ph.D) currently working as Assistant professor Aarupadai Veedu Institute of Technology an ambit institution of Vinayaka Missions Research Foundation (Deemed to be University), Tamil Nadu, India published more than 7 national and international journal and conference and organizing committee for 4 international conference,2 national conference and 15 years of teaching experience with 4 years of research experience. He is a member in following professional societies: CSI and ISTE.

H.Dhrithimanth Prabhu H Final year CSE, Aarupadai Veedu Institute of Technology, Vinayaka Missions Research Foundation Chennai, India

Yerasani Aditya Gowtham Final year CSE, Aarupadai Veedu Institute of Technology, Vinayaka Missions Research Foundation Chennai, India

.