International Journal of Scientific Research in Computer Science, Engineering and Information Technology

© 2017 IJSRCSEIT | Volume 2 | Issue 5 | ISSN : 2456-3307

Twitter Sentiment Analysis on GST tweets using R tool

D. Suganthi*1, Dr. A. Geetha2

*1M.Phil Research Scholar, Department of Computer Science, Chikkanna Government Arts College, Tirupur, Tamil Nadu, India

2Assistant Professor, Department of Computer Science, Chikkanna Government Arts College, Tirupur, Tamil Nadu, India

ABSTRACT

The Goods and Services Tax (GST) has revolutionized the Indian taxation system. This creates a big change in the financial standards of India. Twitter is the ninth largest social networking website in the world, only because of people can share information by way of the short message up to 140 characters called tweets. Twitter is the best source for the sentiment and opinion analysis. The tweets are classified as positive or negative or neutral based on the sentiments. This analysis can be done by classifying the dataset using various Machine Learning Algorithms. One option to perform sentiment analysis in R is to calculate a sentiment score for each tweet. This paper presents the sentiment analysis on the current tweets related to GST.

Keywords: Classification, Opinion mining, Sentiment analysis, Sentiment score, Support Vector Machine.

I. INTRODUCTION

Twitter is the best online platform for the sharing information and opinions, twitter is ninth largest social network in the world, and it has millions of active users. Users are mostly celebrities like politicians, film industry celebrities, sports stars and the common people registered as followers; users can tweet current information in the form of text, video, audio or any format, the rest of the users can react on that tweets[1]. In this way, information is shared among the world.

Feedback is required about the new product, about the government policy executions, and international talks, by the tweets and re-tweets it is the best platform for the feedback, opinion retrieval system in the social networks. Positives and negatives of the tweets are quickly identified using Sentiment analysis using twitter data [2]. In this paper, a system is proposed for the sentiment analysis on GST tweets data using R programming language.

R-studio is free and open source IDE for developing and deploying R applications which can be installed on top of Linux/Macintosh/Windows. R language is a scripting language for conducting statistical data computing and big data analytics; it has more than 10000 packages.

Sentiment analysis is performed on the input dataset that primarily performs data cleaning by removing the stop words, Punctuation, URLs, Special characters etc… followed by classifying the tweets as positive and negative by polarity of the words and also emotions of the words, and then generate the word cloud. Finally, that calculates sentiment score for each tweet and generates positive and negative sentiment.

II. SENTIMENT ANALYSIS

A. Sentence level

The task at this level goes to the sentences and determines whether each sentence expressed a positive, neutral or negative opinion. The first step is to identify whether the sentence is subjective or objective.

B. Document level

The task at this level is to classify whether a whole opinion document expresses a positive or negative sentiment.

C. Feature level

Both the sentence and document level analysis do not discover what exactly people liked and did not like. Instead of looking at language constructs, aspect level directly looks at the opinion itself.

III. PROPOSED METHODOLOGY

A. Steps to extract the tweets

(i) The first step is Creation of twitter application

(ii) In R tool, twitteR package act as interface to the Twitter web API.

(iii) ROAuth package is used for authentication.

(iv) Twitter authenticated credential object such as consumer key, consumer secret, access token, access secret are created.

(v) During authentication, redirection to a URL automatically when clicks on Authorize app, and enter the unique 7-digit number to get linked to the account [3].

B. Pre-processing

(i) **Cleaning text:** The process of cleaning text is carried out by removing unnecessary data from twitter data set such as HTML Tags, emoticons, White spaces, Numbers, URLs, Special symbols.

(ii) **Stop words Removal:** Stop words are the bag of words (such as is, at, which, on etc…) that are removed from the twitter data set, so that the resultant data set contains only required information for the analysis.

C. Lexical Analysis

Lexical analysis may be carried out using lexicon-based approach, which uses a set of positive and negative words. A database, created by Hui Lui contains 2006 positive and 4783 negative sentiment words, is loaded into R and the words in the tweets are compared with the words in the database and the sentiment is predicted[4].

D. Classification

Classification is done using supervised machine learning approaches like naïve Bayes, SVM, Maximum Entropy etc… In this work, the classification is carried out using naïve Bayes.

a) **Naïve Bayes Algorithm:** Naive Bayes classification model computes the posterior probability of a class is computed in Naive Bayes Classifier [5] which is based on the way words are distributed in the particular document. The positions of the word in the document are not considered for classification in this model as it uses bag of words feature extraction technique. Bayes Theorem is used to predict the probability where given feature set belongs to a particular label of the content.

E. Calculating sentiment score

Using Scoring Function score of every tweet has been calculated using Hui Lui lexicons.

Sentiment Score = Σ positive words – Σ Negative words

a) **Polarity types**

(i) Positive polarity - Number of positive words are greater than number of negative words.

(ii) Negative polarity - Number of negative words are greater than number of positive words.

(iii) Neutral polarity - Number of positive and negative words are same or is no existence of any opinion words.

F. Visualization

Sentiment analysis can be visualized by graphical representation using R-studio, there are a rich set of graphical packages are available in R. In this paper, word clouds and bar charts are used to represent the outcomes of the sentiment analysis.

IV. EXPERIMENTS AND RESULTS

A. Collecting GST Related Tweets

Before mining any data from Twitter using APIs, we have to authenticate with Twitter using an application created on Twitter. Once the application is created, we get access to consumer key, consumer secret, access token, access secret using which the API has to authenticate itself with the Twitter Authentication server.

```
consumer_key<-„xxxxxxxxxxxx'
consumer_secret<-'xxxxxxxxxxx'
access_token<-'xxxxxxxxxxxxxx'
access_secret<-„xxxxxxxxxxxx'

setup_twitter_oauth (consumer_key, consumer_secret,
access_token, access_secret)
```

Access twitter data sets

Once API is authenticated with Twitter Authentication service, a token is generated and is made available to API for every transaction with the Twitter server. Using this token, tweets are mined using hashtags. We use searchTwitter() function to access the data. In this work we extract 5000 tweets on GST.

```
searchTwitter ('GST', n=5000, lang = 'en')
```

B. Classification of tweets

Classification by polarity

The polarity operation is applied on pre-processed data set, the data which contains cleaned data with bigram features, the polarity function can generate the sentiment scores for each tweet, if it is negative or positive tweets, and we need what are the positives and negative from the public [6].

These are represented by bar chat in figure 1 and word clouds are generated. Figure 3 is shows that the word cloud.

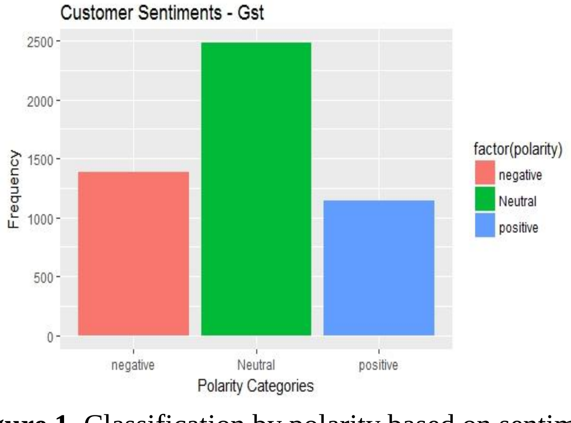

Figure 1. Classification by polarity based on sentiment score

Naïve bayes algorithm are applied in dataset and the results are displayed in bar chat figure 2 depicts that.

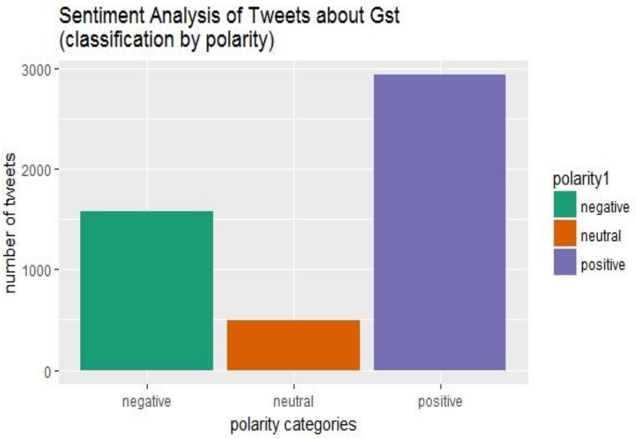

Figure 2. Classification by polarity using naïve bayes

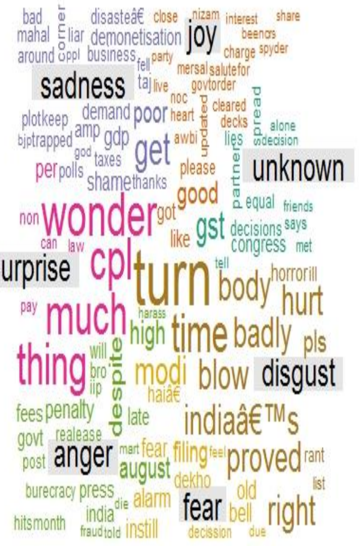

Figure 3. Word cloud of GST tweets

Classification by emotions

Figure 4 shows that classification by emotion using naïve bayes. The emotions are classified as anger, joy, surprise, sadness, fear, and the other category is unknown [7].

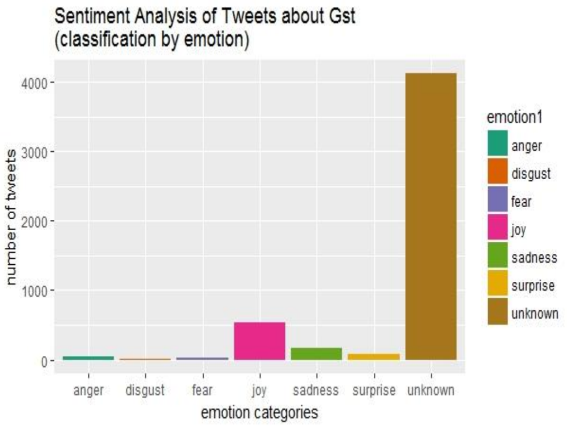

Figure 4. Classification by emotion using naïve bayes

Classification by sentiment score

Classification of tweets with polarity method having the SentiStrength scales from 1 to 5 for both positive (+1 weak positive to +5 extreme positive) and negative (-1 weak negative to -5 extreme negative) sentiments[8].The sentiment score is a more precise numerical representation of the sentiment polarity. Figure 5 shows that the sentiment score on GST tweets.

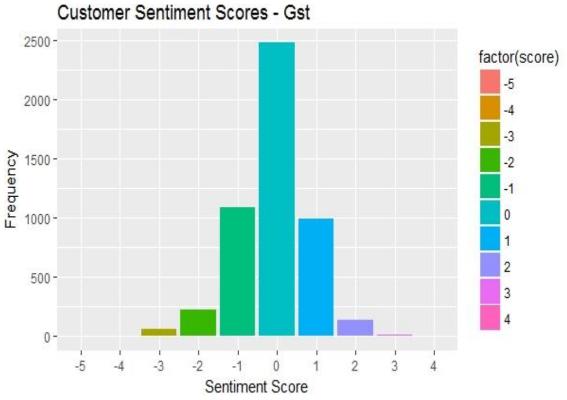

Figure 5. Sentiment score on GST tweets

V. CONCLUSION

Sentiment analysis is the efficient technique to analyze the user behavior. Tweets are the samples of people's opinion. Sentiment analysis is finding the aspects and their polarity of the schemes which helps for implantation of government schemes effectively to take decision to upcoming scheme regarding public opinion. In this paper sentiment scores are calculated and counted number of positive, negative and neutral tweets and classifies the public opinion of particular Scheme. Using sentiment analysis sentiment score was calculated and plotted also sentiment based on polarity as well as emotions are plotted. From this peoples and government can find out the peoples opinion behind that declared scheme. Taking survey from peoples is very expensive and time consuming. So this sentiment analysis is very useful for evaluation of government scheme and monitoring the growth of the scheme from people's perspective.

VI. REFERENCES

[1]. G. Vinodhini and RM. Chandrashekharan, "Sentiment Analysis and Opinion Mining: A Survey" – International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 6, June 2012.

[2]. K.Arun, A.Srinagesh, M.Ramesh, "Twitter Sentiment Analysis on Demonetization tweets in India Using R language" International Journal of Computer Engineering In Research Trends, Volume 4, Issue 6, June-2017, pp. 252-258.

[3]. Bharat R.Naiknaware, Seema Kawathekar, Sachin N.Deshmukh "Sentiment Analysis of Indian Government Schemes Using Twitter Datasets", IOSR Journal of Computer Engineering, e-ISSN: 2278-0661, p-ISSN: 2278-8727, PP 70-78.

[4]. Karthik Ganesan, Akhilesh P Patil, Srinidhi Hiriyannaiah, "Predictive Analysis of Tweets on Goods and Services Tax(GST) in India using Machine Learning", International Journal of Engineering Technology, Management and Applied Sciences, August 2017, Volume 5, Issue 8, ISSN 2349-4476.

[5]. Sajin. S. Chandran, Murugappan S., "A Review on Opinion Mining from Social Media Networks", European Journal of Scientific Research, pp.430-440, 3rd October, 2012.

[6]. Ayesha Rashid1, Naveed Anwer2, Dr. Muddaser Iqbal3, Dr. Muhammad Sher4 "A Survey Paper: Areas, Techniques and Challenges of Opinion Mining", IJCSI International Journal of Computer Science Issues, Vol. 10, Issue 6, No 2, November 2013.

[7]. Kirti Huda, Md Tabrez Nafis, Neshat Karim Shaukat, "Classification Technique for Sentiment Analysis of Twitter Data" , International Journal of Advanced Research in Computer Science, Volume 8, No. 5, May-June 2017 ISSN No. 0976-5697

[8]. M. Govindarajan, Romina M, "A Survey of Classification Methods and Applications for Sentiment Analysis" – International Journal of Engineering and Science (IJES), Volume 2, Issue 12, 2013.