Sentiment Analysis on GST using Polarity Classification

MS. Y. Vineela Sravya[1], T. Jaya[2], R. Sravani Sandhya[3], M. Pravalika[4], S. Snigtha[5]

[1]Assistant Professor, Department of Computer Science and Engineering, Vignan's Institute of Engineering for Women, Visakhapatnam, India
[2-5]Department of Computer Science and Engineering, Vignan's Institute of Engineering for Women, Visakhapatnam, 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 people can 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 and 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, Opinionmining, Sentiment analysis.

1. Introduction

Now a day internet is flooded with several social network sites like Twitter, Facebook and many more, which provide various ways to connect with other active users worldwide. Twitter[1-4] is one of the popular and vastly used micro blogging services on the internet today. It connects people from diverse areas and helps them to, communicate with each other, participate in online group discussions, share an opinion on a product, service or some social issue in a real and quick manner. It enables users to write short text messages, i.e, up to 140 characters long called "tweets". Twitter provides many interactive features like following, retweet, mentions, etc. that facilitate a way to connect with other like-minded people and this way a group of similar people is formed. The tweets exchanged in such groups can be analyzed to my opinion of the people about the discussed topic. Millions of people from all over the world are active on Twitter, and this count of active users is increasing every day. So, it becomes a rich source of information and hence a good area of research for many researchers[5,6].

1.1 *GST in India*

The Goods and Services Tax (GST) bill was passed in the parliament on 29th March 2017. The Act came into effect on 1st July 2017 and had replaced many Indirect Taxes in India[2,7]. The GST is India's most significant tax reform and majorly discussed the topic in online networking groups such as Twitter. Public opinion towards this uniform tax has an essential role in its acceptance.

1.2 *Sentiment Analysis*

The sentiment is an attitude or perception driven by feeling. Sentiment Analysis (SA) is the process of determining and measuring the emotional state of response made by opinion[8-11]. SA is a tool that enables companies to analyze what their customers are thinking about their products and services and identifies the trends in the behavior or attitude of their customers toward the products and services they provide concerning their competitors. The Opinion mining of a product is performed by identifying the sentiments of opinion holder and then classifying their polarity. This work is based on the analysis of Twitter data related to GST to predict people's opinion. Sentiment analysis of Twitter textual data(called tweets) is done with an event and time-specific training dataset, which is semi-manually annotated, to capture the sentiments of Indian people towards GST implementation. As our dataset is manually annotated (i.e., labeled as positive or negative) we opt for a supervised machine learning approach to perform SA.

Sentiment Analysis can be performed at the following three levels[12, 13]:

i. *Document-level:* Sentiment Analysis at this level is used to identify the polarity of a single entity in the given document. Document-level sentiment analysis assumes that each document expresses opinions only on a separate body, e.g., a single product or service and is expressed by a single opinion holder.

ii. *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.

iii. *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 level.

1.2.1 *Techniques of Sentiment Analysis*

Sentiment analysis is widely used in variety of applications like classifying, summarizing and aggregating reviews from the massive volume of unstructured data

that may be available from customer comments, blogs, feedback and reviews on any product or social issue. The SA can be performed by using the following two approaches[12]:

1) Lexicon based Approach

2) Machine Learning Approach

The Lexicon-based Approach relies on a sentiment discovery, a collection of known and precompiled sentiment terms. It is divided into the dictionary-based approach and a corpus-based approach which uses statistical or semantic methods to find sentiment polarity. The Machine Learning Approach (ML), applies the famous supervised or unsupervised machine learning algorithms and uses linguistic features extracted from the opinion text. Fig. 1 illustrate the different approached that can be used to perform sentiment analysis.

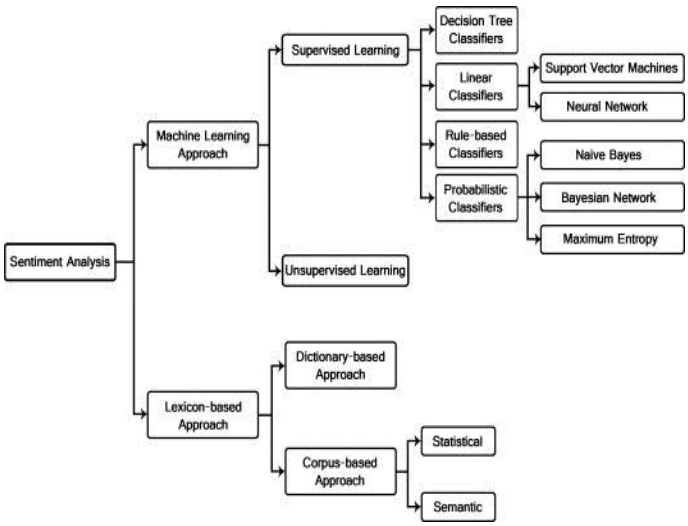

Fig. 1. Sentiment Analysis Techniques

2. Literature Survey

The growth is social media networks has made sentiment analysis a popular research area, in recent years. Sentiment plays a crucial role in the decision-making process. In the field of information retrieval, web analysis, decision making, and product/service reviews opinion mining has become increasingly important.

In [14] unsupervised method is used to perform classification. To determine the semantic orientation of a phrase two seed words are used, i.e., 'poor' and a point wise mutual information method was used. [15] covers techniques and approaches that promise to enable opinion-oriented information-seeking systems directly. The material on summarization of evaluative text and broader issues regarding privacy, manipulation and economic impact that the development of opinion-oriented information-access services gives rise are included in this paper. In [16] part of speech (POS) patterns are used to extract the sentiment from given phase, and unknown sentiment phrase is used as a query term. The top n number of relevant phrases are extracted, and the sentiment analysis is done on them by using the lexicon of nearby known relevant phrases. An Automatic seed Word Selection method[17] was proposed for unsupervised sentiment classification of product reviews in Chinese. The lexical-based and machine learning-based approaches are combined in [13] with the aim of introducing a hybrid architecture with high accuracy than the pure lexical method and provides more structure and increased redundancy than machine learning approach. The Semi-lexical algorithm [9] to find the polarity of a review as positive, negative or neutral is proposed to handle words which have negation effect on the reviews. The role of emotions is also discovered in this paper. In [18] a fuzzy inference system is designed based on experimentally developed fuzzy membership functions and concepts of hedges to standardized and formulate the process of strength quantification of subjective sentences when the strength of opinion word gets modified by the presence of n-gram adverbial modifiers pattern in the sentence.

3. Proposed Method

A. *Steps to extract the tweets*

i. The first step is the Creation of twitter application.

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

iii. OAuth package is used for authentication.

iv. Twitter authenticated credential object such as consumer key, consumer secrete, access token access secrete 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 the 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 Naive Bayes, SVM, Maximum Entropy ,etc… In this work, the classification is carried out using Naive Bayes.

a) *Naive Bayes Algorithm:* Naive Bayes classification model computes the posterior probability of a class is computed in the 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 a bag of words feature extraction technique. Bayes theorem is used to predict the probability where a 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 = positive words – negative words

Polarity types:

1) *Positive polarity –The* Number of positive words is greater than the number of negative words.

2) *Negative polarity –The* Number of negative words is greater than the number of positive words.

3) *Neutral polarity –The* Number of positive and negative words are the 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.

4. Implementation 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 Authenticate server.

Consumer_key<-'xxxxxxxxxxxxxxxxx'

Consumer_secrete<-'xxxxxxxxxxxxxxx'

Access_token<-'xxxxxxxxxxxxxxxxxxx'`

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 search Twitter() function to access the data. In this work, we extract 5000 tweets on GST.

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 sets, 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 shows that the word cloud.

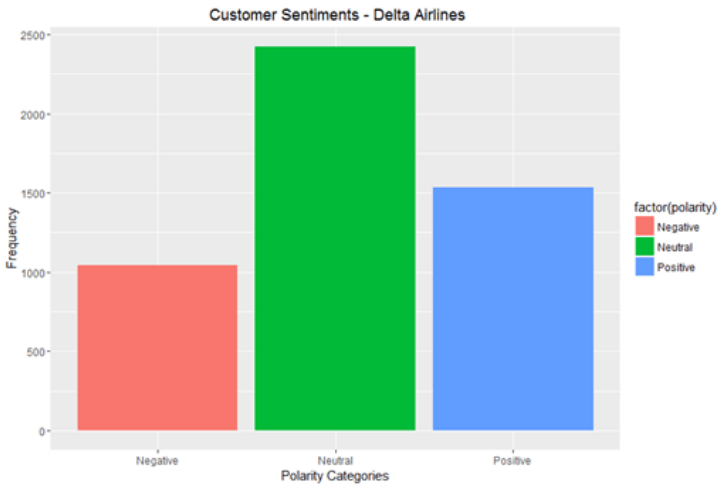

Figure 1. Classification by polarity based on sentiment score

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

Polarity	Number of tweets
Positive	2203
Negative	638
Neutral	2159

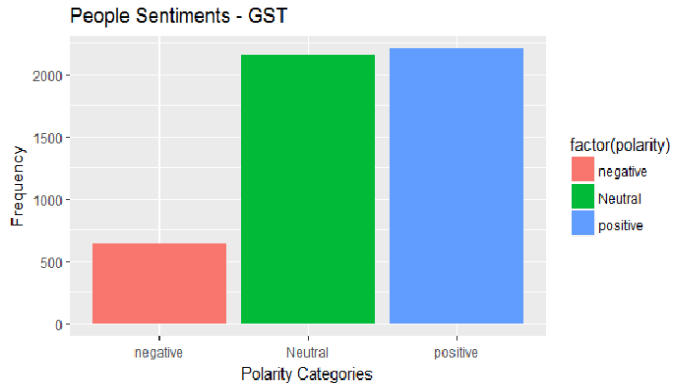

Figure 2. Classification by polarity using naïve bayes

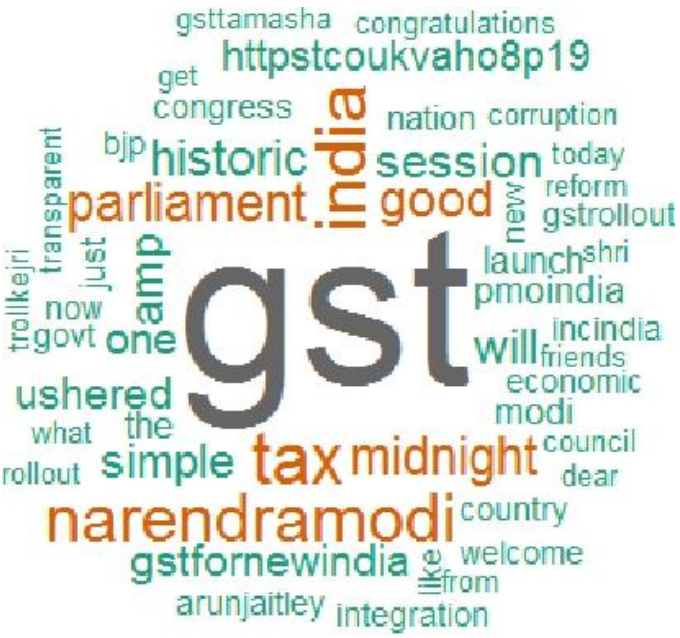

Figure 3. Word cloud of GST tweets

Classification by emotions

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

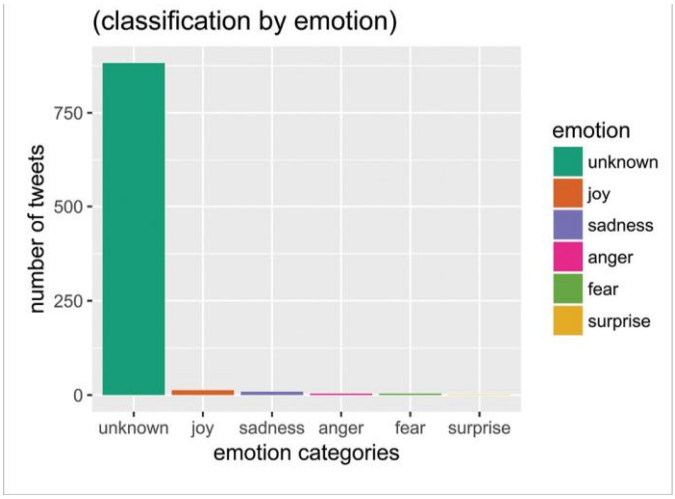

Figure 4. Classification by emotion using naïve bayes

Classification by Sentiment score

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

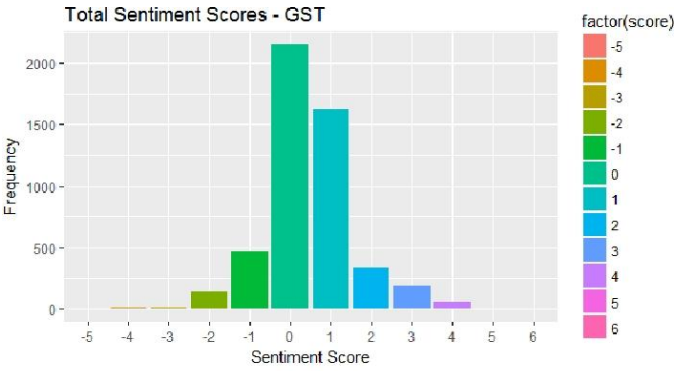

Figure 5. Sentiment score on GST tweets

5. Conclusion

With the rapidly expanding social networks, it is challenging to analyze its large data using existing data mining tools. We have shown that our Architecture to access Twitter and R-studio. Environment analyses large data for decision making. We have shown through our algorithms to do Sentiment Analysis on retrieved "GST" data from Twitter that the number of people has given polarity and emotions. With this, it is advisable to conclude the R Statistical Tool is sufficiently used for the analysis of streaming data.

References

[1] R. Srivastava and M. Bhatia, "Quantifying modified opinion strength: A fuzzy inference system for sentiment analysis" in Advances in computing , Communications and informatics (ICACCI), 2013 International Conference on, 2013, pp, 1512-1519:IEEE.

[2] K. Pabreja, "GST sentiemnt analysis uisng twitter data", IJAR, vol. 3, no. 7, pp.660-662,2017

[3] G. Vinodhini and RM. Chandrashekharan, "Sentiemnt Analysis and Opinion Mining: A Survey"- Internal Journal of advanced Research in Computer Science and Software Engineering, Volume 2, Issue 6, June 2012.

[4] M. Govindarajan, Romina M, "A Survey of Classification Methods and Apllications for Sentiment Analysis"- International Journal of Enigeering and Science (IJES), Volume 2, Issue 12, 2013.

[5] Kirti Huda, Md Tabrez Nafis, Neshat karim Shaukat, "Classifiaction 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.

[6] https://analytics4all.org/2016/11/25/r-twitter-sentiment-analysis/

[7] John Ross Quinlan. C4. 5: programs for machine learning , volume 1. Morgan Kaufmann, 1993.

[8] https://www.cs.mum.ca/~donald/slug/2005-04-07/slugspam.pdf

[9] Yanchang Zhao RDatamining.com Mining Data analyiss Easier

[10] https://pdfs.semanticscholar.org/837d/2b24fe706054abbe729f86b680360dcf58fc.pdf