International Journal of Mechanical Engineering and Technology (IJMET)
Volume 9, Issue 1, January 2018, pp. 320–326, Article ID: IJMET_09_01_034
Available online at http://www.iaeme.com/IJMET/issues.asp?JType=IJMET&VType=9&IType=1
ISSN Print: 0976-6340 and ISSN Online: 0976-6359

© **IAEME** Publication **Scopus** Indexed

A COMPREHENSIVE ANALYSIS ON SENTIMENTAL DATA SET USING MACHINE LEARNING TECHNIQUE

P. Pavan Kumar, N. Lakshmi Chandana, A. Sai Venu Madhav Reddy,
Prof. D. Rajeswara Rao
Department of CSE, K L E F, Andhra Pradesh, India

ABSTRACT

The Internet changed the way through which people communicate and also express their opinions and their emotions. Be that as it may, the information via web-based networking media isn't extremely justifiable and cannot be easily understood for effective data mining, this is because these days individuals on the web utilize diverse dialects and web slang to express their emotions and feelings. In this survey, we are making use of "Naïve Bayes" [7] the algorithm to classify our Twitter data. Using this algorithm we are going to do opinion mining on "Goods and Service taxes (GST)" which is currently most discussed topic on social media websites. In this paper, we associate our twitter record to R tool to extricate information from the twitter and with the utilization of "Naive Bayes" calculation we perform sentimental analysis on about 10,000 tweets and show our outcomes graphically for client understandability.

Key words: Twitter data, Sentimental analysis, Opinion Mining, Machine learning, Naïve Bayes, Goods and Service Tax, Data Mining.

Cite this Article: P. Pavan Kumar, N. Lakshmi Chandana, A. Sai Venu Madhav Reddy, Prof. D. Rajeswara Rao, A Comprehensive analysis on Sentimental Data Set Using Machine Learning Technique, *International Journal of Mechanical Engineering and Technology* 9(1), 2018, pp. 320–326.
http://www.iaeme.com/IJMET/issues.asp?JType=IJMET&VType=9&IType=1

1. INTRODUCTION

The Sentimental analysis [10] is additionally called as opinion mining which focuses on the utilization of regular dialect handling, and furthermore message investigation of tremendous amounts of data. This sentimental analysis sometimes alluded as emotion Artificial Intelligence. It is the study of affective states and emotions present in the given information. The sentimental analysis is applied to various aspects of the web such as client audits and reviews.

In general terms, sentimental analysis aims to find the attitude of a customer expressed in a website. It finds out the contextual polarity or emotions of a customer stated in social networking content. The attitude such determined may be the opinion or a judgment based on the text mentioned by the customer on the social media websites such as twitter, facebook [15], Amazon, Flipkart, youtube comments etc.

As the social media plays an most significant role in a person's life, it comprises of information regarding their opinions and feelings. The extraction of constructive information from an enormous amount of unstructured data, also well known as sentiment analysis, is a task that has recently gained more regard for purposes such as social marketing and advertising. Such tasks, however, are to a great degree intense in nature as online networking locales today are impeccably reasonable for human utilize yet they remain barely available for machines to process.

One of the most popular social networking sites is Twitter through which data is being extracted for analysis in this paper. In this paper sentimental analysis [10] is conducted on nearly 10,000 tweets about the topic "Goods and Service taxes". The text that is available in the form of a tweet may contain various forms of data such as images, links, internet slang words, re-tweets etc. The data is processed t to find the emotion of the writer that is the customer or writer with the help of R tool and displayed the results in the graphical manner.

2. LITERATURE SURVEY

The Authors (Ana C. E. S. Lima, 2012) [1] has proposed three distinctive methodologies for the classification of the emotions automatically. The first among the stated approaches is emotion based approach and the Second one is based on words and the third one is the hybrid approach. In the first approach that is the emotion-based approach, they incorporated the sentiment present in the emotions expressed by the customer as the criteria to classify the textual data automatically. The word-based approach uses the words as the criteria to assess and evaluate the text.

In word-based approach, words like good, nice, marvellous, bad and worst etc will show the sentiment and the opinion can be inferred based on these words. And in the hybrid approach both the Emoticons and words are used in determining the sentiment. They classified the tweets based on the hybrid approach as a whole because this approach yields better results when compared to the other two.

In their paper, (H. Sinha, 2016) [2] they have compared distant algorithms as a part of performing sentimental analysis and they have also rated them based on their accuracy in giving results. They performed opinion mining on training data set containing reviews regarding products in an online site and scrutinized them using multiple analysis algorithms and have compared them for their exactness.

Authors (Pravin Keshav Patil, 2015) have resolved the mostly faced challenge that emerges while performing text analysis. The challenge is, a word may have distinctive synonyms yet same meaning. They used lexicon [3] based approach that is the dictionary based mechanism in solving this challenge by using [8] Naïve Bayes classification. This approach is proved to give better outcomes when compared to various other methodologies that are as of now available in the market.

(Huma Parveen, 2016) used Hadoop [4] Framework to perform sentimental analysis on movie data set. They have done text analysis on data such as reviews, feedbacks and also user comments. Their paper provides us with the user opinions in the form of categories like positive, neutral and negative sentiments. They provided us with the fastest downloading approach for efficient Twitter Trend Analysis.

Naïve Bayes classifier is mainly used for filtering [5] spam emails. But sometimes it may not perform completely well in identifying spam emails because of dependency between the data. So, in order to resolve this issue (Weimiao Feng, 2016) have used Support Vector Machine to reduce dependency between the data variables. This approach is proved to be the best in lessening the dependency among data and produced classification results with higher accuracy.

(A. M. Abirami, 2016) compared different approaches [6] and methods that are applied for performing Opinion Mining. Naïve bayes algorithm, [13] Support vector Machine, Maximum Entropy are some of the known machine learning algorithms that are compared in this paper. They represented their results in a manner such that any user can understand and compare the accuracy of the methods just through an eyescan of the paper.

In this world of massive growing communication mediums, social networking site plays a key role. Nowadays among 100 people 90 people are communicating and expressing their views on the web by accessing social networking sites. Author (Yousukkee, 2016) presented a survey which focuses on the process of classifying these sites in order to analyze customer/user's [9] behavior in those sites.

3. METHODOLOGY

In this survey, [13] Naïve Bayes algorithm to classify our twitter data and to find customer opinion. There are two variant methods used in classifying data such as supervised and unsupervised learning. Naïve Bayes classifier works very well under supervised learning. This algorithm aims to find maximum likeliness of occurrence of words by finding the probability.

This algorithm works well even under complex classifications. The main advantage in using this algorithm lies in its requirement of needing only small amount of training data in order to estimate the necessary parameters forclassification. Naïve Bayes [8] is a condition based probability model.In order to classify the data, we need to consider a vector n=(n1,n2,n3….nx) where x represent number of

Independent variables.

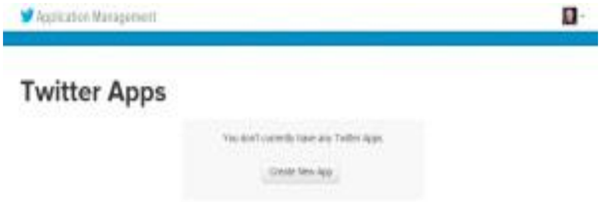
$$\underset{\text{Posterior Probability}}{P(c\,|\,x)} = \frac{\overset{\text{Likelihood}}{P(x\,|\,c)}\,\overset{\text{Class Prior Probability}}{P(c)}}{\underset{\text{Predictor Prior Probability}}{P(x)}}$$

$$P(c\,|\,\mathrm{X}) = P(x_1\,|\,c)\times P(x_2\,|\,c)\times\cdots\times P(x_n\,|\,c)\times P(c)$$

Here in this paper unstructured information will be extracted from twitter account using naïve bayes theorem and we display our results graphically with the help of R tool.

3.1. Process Steps

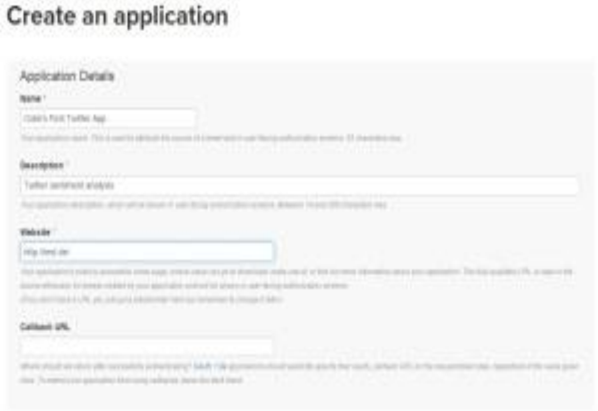
1. Log in to your twitter account using the following link [11] https://apps.twitter.com/. The interface looks like this



Twitter Apps

You don't currently have any Twitter Apps.

2. In the above shown interface you should click the option button named “Create New app”, then the interface changes and displays a form which you should fill in order to continue creating your application [14] in twitter.



Create an application

Application Details

Name *

[illegible]'s First Twitter App

Description *

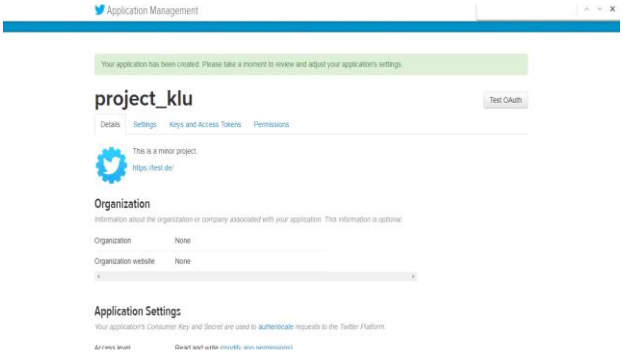
Twitter sentiment analysis

Website *

http://test.de/

Callback URL

3. After filling the above displayed form next step is to choose your app name and a textbox is given to fill in your app description. This website requires you to enter a valid URL [15]. A default URL http://test.de/ in case if we do not have our own created URL.Then it shows a Developer Agreement, read all the information presented in the agreement and then if you want to proceed then click on “Yes, I accept” button. Then the screen appears like this one shown below.



Your application has been created. Please take a moment to review and adjust your application's settings.

project_klu

This is a minor project.

https://test.de/

Organization

Information about the organization or company associated with your application. This information is optional.

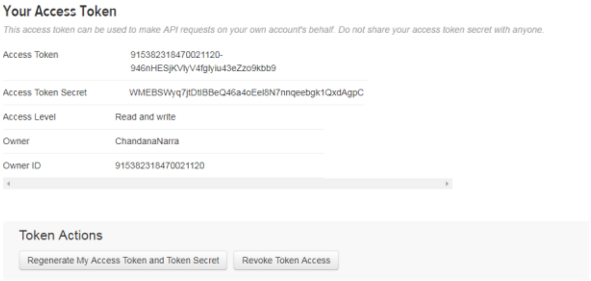
Organization	None
Organization website	None

Application Settings

Your application's Consumer Key and Secret are used to authenticate requests to the Twitter Platform.

Access level Read and write (modify app permissions)

4. In this interface it shows multiple options like Details, Settings, Keys and Access Tokens and Permissions. Next step is to click on Keys and Access tokens, then consumer key, consumer secret key, access level, owner, and owner ID can be seen on your application API screen. After scrolling down further there appears a section named Your [20] Access Token, then in that section click “create my access token” [12] which appears like this



Your Access Token

This access token can be used to make API requests on your own account's behalf. Do not share your access token secret with anyone.

Access Token	915382318470021120- 946nHESjKVlyV4fglyiu43eZzo9kbb9
Access Token Secret	WMEBSWyq7jtDtlBBeQ46a4oEel8N7nnqeebgk1QxdAgpC
Access Level	Read and write
Owner	ChandanaNarra
Owner ID	915382318470021120

Token Actions

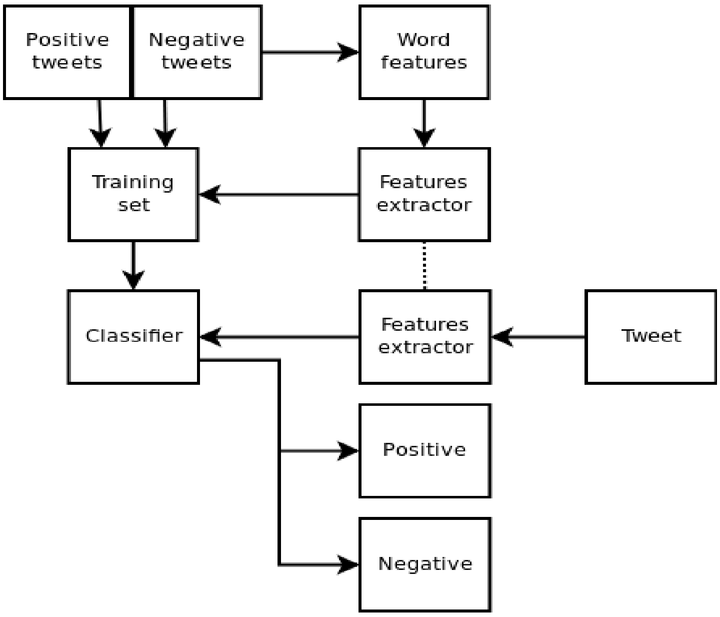
5. Using these access tokens generated by following step 3 authorize R tool to access [12] Twitter. Then loading packages and libraries required in order to analyze tweets[11] present in corpus installed in our R tool.

Some more packages are needed like twitterR which provides us R interface to our twitter application program interface.

Install packages like [20] ROAuth and RCurl which help us in authenticating our web servers and also paves a way to access our http requests and responses returned by our web server.

6. Next step is to substitute our secret password in place of default one in the R script and then the screen will display a page with [15] authorize app and cancel buttons highlighted. Then click on authorize app which then generates a one time use [18] PIN(which may differ for different applications and executions).

7. After this copy the PIN from twitter API and paste this PIN in the [17] R script which gives us permission to run twitter searches. One can generate this random PIN for only once but R language requires you to use your token string and your secret strings again.



8. The final and most important step is to write an effective R script to search our Twitter [11]. This task is very crucial because writing it effectively helps you in getting better results. Finally, results will be displayed in the form of a boxplot in order to provide clear results for the user.

4. RESULTS

The below shown screenshots are the output of the search about the topic GST in twitter data. The Fig 1 represents the output in the form of a [19] boxplot. In Fig 1 a boxplot is shown with X axis representing emotion categories such as joy, fear, anger, surprise, disgust and unknown. Where as Y axis denotes the number of the tweets ranging from 0 to 8000.

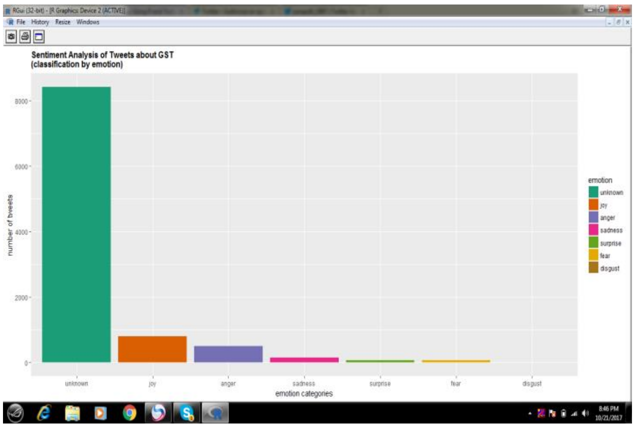


Figure 1

Fig 2 shows boxplot [19] which is the result of classification by polarity. Here in this X axis denotes the polarity categories like [16] positive, negative and neutral and our Y axis shows the number of tweets ranging from 0 to 4000.

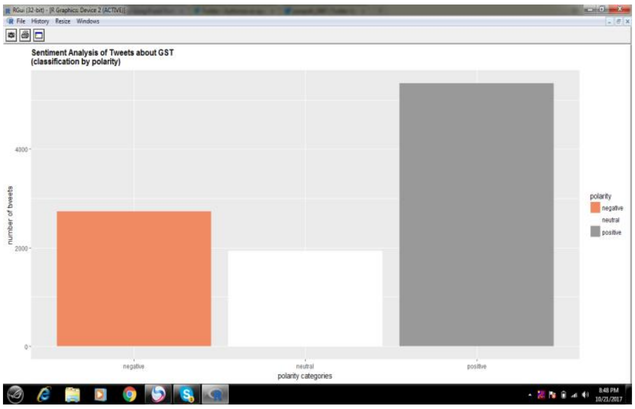


Figure 2

Fig 3 clearly displays the word cloud formed by the words present in our corpus or library.

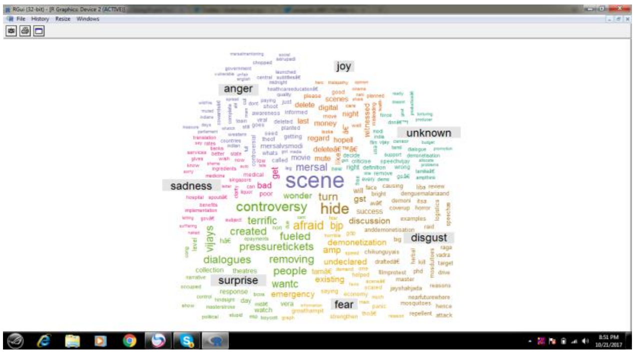


Figure 3

5. CONCLUSIONS

Web contains a lot of data which can be analysed to find patterns. One way to do so is by performing sentimental analysis which finds the opinion of a customer by performing text analysis. In this paper Naïve Bayes algorithm is utilized to perform text analysis in order to find the opinion of the Twitter users about the topic Goods and service tax effect in India. Results here are obtained by using R tool to search for the specific keyword which is GST in our case and found the results which we have displayed with the help of a boxplot. Boxplot is a flexible graphical representation provided by R tool that helps us understand our results clearly without any ambuigity.In case of unsupervised data sets Naïve bayes algorithm works well with results having an accuracy of nearly 85%.

REFERENCES

[1] Ana C. E. S. Lima, L. N. (2Ol2). Automatic Sentiment Analysis of twitter Messages.

[2] H. Sinha, A. K. (2Ol6). A Detailed Survey and Comparative Study of Sentiment Analysis Algorithms. *Ieee*.

[3] Pravin Keshav Patil, K. P. (2Ol5). Automatic Sentiment Analysis of Twitter Messages Using Lexicon Based Approach Naive Bayes Classifier with Interpretation Sentiment Variation. *Ijirset , 4* (9), 10.

[4] Huma Parveen, S. P. (2Ol6). Sentiment analysis on Twitter Data Set using Naive Bayes algorithm. *iCAtccT* .

[5] Weimiao Feng, J. S. (2Ol6). A Support Vector Machine based Naive Bayes algorithm for Spam Filtering. *IPccC* .

[6] A. M. Abirami, V. G. (2Ol6). A Survey on Sentiment analysis methods and approach. *ICoac*.

[7] https://machinelearningmastery.com/naive-bayes-for-machine-learning/

[8] https://en.wikipedia.org/wiki/Naive_Bayes_classifier

[9] Yousukkee, S. (2016). Survey of analysis of user behavior in online social network. *MITicon* .

[10] https://www.lexalytics.com/technology/sentiment

[11] https://developer.twitter.com/en/docs/tweets/data-dictionary/overview/user-object

[12] https://developer.twitter.com/en/docs/tweets/data-dictionary/overview/user-object

[13] https://www.dezyre.com/article/top-10-machine-learning-algorithms/202

[14] http://analyzecore.com/2017/02/08/twitter-sentiment-analysis-doc2vec/

[15] https://www.credera.com/blog/business-intelligence/twitter-analytics-using-r-part-1-extract-tweets/

[16] http://www.evoketechnologies.com/blog/sentiment-analysis-r-language/

[17] https://www.r-bloggers.com/sentiment-analysis-with-machine-learning-in-r/

[18] https://developer.twitter.com/en/docs/basics/authentication/overview/pin-based-oauth

[19] https://sites.google.com/site/miningtwitter/questions/sentiment/analysis

[20] http://rstudio-pubs static.s3.amazonaws.com/283869_04d1ed5678d84af68978cf34661d63cf.html

[21] Magesh G and Dr. P. Swarnalatha, Analyzing Customer Sentiments Using Machine Learning Techniques. International Journal of Civil Engineering and Technology, 8(10), 2017, pp. 182 9 – 1842

[22] V. Sathya and T.Chakravarthy, Automatic Facial Expression Related Emotion Recognition Using Machine Learning Techniques, International Journal of Computer Engineering & Technology , 8(5), 2017, pp. 126–135.