International Journal of Psychosocial Rehabilitation, Vol. 23, Issue 04, 2019
ISSN: 1475-7192

Dictionary Based Opinion Classifier and Sentiment Analyser of Social Media Data

[1]*K.Jayamalini, [2]Dr.M.Ponnavaikko

ABSTRACT– *Sentiment analysis or opinion mining is a method of NLP that is, used by organizations to find and categorize the expressive nature behind a body of text, which are given by, their customers as reviews about their products and services. This is very popular research area, which uses latest techniques like machine learning (ML), deep learning (DL) and artificial intelligence (AI) to find the insights of user text for finding user sentiment and subjective information behind it.*

The massive volume and variability of the data produced by online social media assistances various businesses in decisionmaking. This paper, explains a dictionary based approach to apply the sentiment analysis technique on twitter data. Real time user timeline tweets had extracted and stored as corpus. This corpus had used for determining user sentiment and opinion polarity. This paper also deals with the classification technique to classify the text into positive, negative or neutral based on the polarity value.

Keywords- *Sentiment Analysis(SA), Opinion Mining(OM), Dictionary Based Approach, Social Media Data.*

I. INTRODUCTION

Twitter is a hugely popular social media platform for expressing our emotions, opinions and activities. Twitter consists of massive amount of information that is available on the web. Today, the structured as well as unstructured data on the internet [1] is increasing at a rapid pace. Industries, companies and businesses are using this tremendous amount of structured, unstructured and semi structured data available on social media to gain insight into people's views and opinions about their products and services. This would be a powerful tool to keep the company ahead of competition. Social media is a critical source to extract the opinions of end users, and more importantly, the emotions attached to how users perceive the products and services. Analysis of this data would be virtually impossible if done manually.

Twitter is one of the most popular online platforms and social media network that allows people to share information. It allows registered users to view and post their own tweets. Tweets are restricted to 140 characters. Registered users can express their opinion by posting tweets or by uploading photos or short videos. They can also re-tweet the tweets posted by other users.

In recent years, tweets and re-tweets posted by users have become a tool to gauge user opinions about products, people, government policies, and national & international issues. Top government officials and many celebrities

[1]* *Research Scholar, Computer Science Engineering, Bharath University, Chennai, India, malini1301@gmail.com*

[2]*Provost, Bharath University, Chennai, India. ponnav@gmail.com*

Received: 22 Aug 2019 | Revised: 13 Sep 2019 | Accepted: 15 Nov 2019

in the fields of politics, film industry, and sports use Twitter data to find public opinion about them. Performing sentiment analysis on stream of Twitter data finds the polarity value for each tweet in the stream and classifies them into positive, negative and neutral. This paper deals in depth with the sentiment analysis on Twitter Data to discover user opinions about GST as a case study.

GST [2] is a taxation procedure for goods and services transported from one destination to another. It is an indirect tax levied in India on the sale of goods and services. Goods and services are divided into five tax slabs for collection of tax - 0%, 5%, 12%, 18% and 28%.The tax came into effect from July 1, 2019, through the implementation of One Hundred and First Amendment of the Constitution of India by the Government of India. The tax replaced existing multiple cascading taxes levied by the central and state governments. Taxation and its associated governing laws, plays a significant role in the life of business, its impact on the individual and on government policies for social good.

This paper experimentally shows user opinions on the impact of GST after its implementation. Section II of this paper gives literature review; Section III of this paper explains overview of the System Architecture, Section IV explains implementation details and Methodology, and Section V contains Result Analysis.

II. REVIEW OF LITERATURE

A. Previous Work

The rapid growth and population of blogs and social media networks [3]make the field of opinion mining and sentiment analysis an interesting field for researchers. Bernard J. Jansen , Mimi Zhang [4] explains the facts, methods, data sources and various algorithms used in the field of opinion mining. Liang Wu, Teng-Sheng Moh, Natalia Khuri [5] describes the algorithmic analysis of the sentiment of e-micro-blogs covering about 50 brands and categorizes them to determine the aggregate characteristics of the brand. In Paper [6], Peiman Barnaghi, et al. describes a computational pipeline for collecting, processing, and analyzing tweets to find signals about adverse drug reactions, (drug side effects caused by a drug at a normal dose during normal use) using NLP. In Paper [7], Lu Lin, et al. presents an objective to use Twitter Streaming API and official world cup hash tags to mine, filter and process tweets, in order to analyze the reflection of public sentiment towards unexpected events using SVM machine learning classifier. Lei Wang and John Q Gan [8] proposes an efficient method to mine and summarize opinions collected by Weibo, a Twitter-like micro blog service in China. In Paper [9], Roshan Fernandes, Dr. Rio D'Souza explains how Twitter sentiment analysis was used as a tool to predict the French General Elections in 2019.

B. R & R Tool

R is a programming language and free software environment for statistical computing and graphics supported by the R Foundation for Statistical Computing. The R language is widely used among statisticians and data miners for developing statistical software and data analysis. Polls, surveys of data miners, and studies of scholarly literature databases show that R's popularity has increased substantially in recent years.

Received: 22 Aug 2019 | Revised: 13 Sep 2019 | Accepted: 15 Nov 2019

R is a programming language and software environment intended for deep statistical computing and graphics. It isopen source and available across different platforms, e.g., Windows, Mac, Linux. It is used in a variety of applications including visualizations and data mining.

III. OVERVIEW OF THE SYSTEM ARCHITCTURE

This section explains the complete approach and various steps used in the Twitter Opinion Miner. The framework ofopinion analyzer to analyze the user opinions towards a specific issue is depicted in Figure 1. The nature of data extracted from Twitter may bein a structured format, semistructured format or unstructured format. The framework comprises of Tweets Extractor, preprocessor, Feature Extractor,
Opinion Value Calculator, Opinion Classifier,

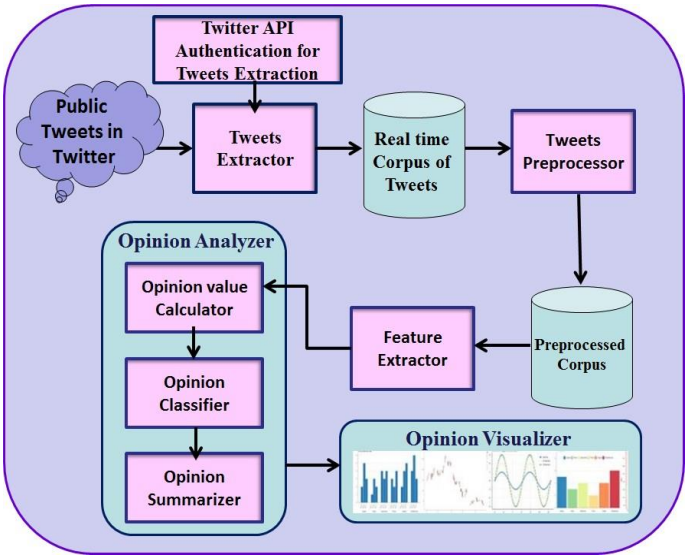

Figure 1: Architecture - Twitter Data Opinion Analyzer

The methods [10] involved in the development of Twitter Data Opinion Miner had explained below:

A. Tweets Extraction& Data Collection
B. Tweets Preprocessing
C. Feature Extraction
D. Opinion Analysis
a. Opinion identification & Value Calculation
b. Opinion Classification
c. Opinion Summarization
E. Opinion Visualization

The Extractor connects to Twitter and extracts specific tweets based on given search keyword. The data collected from Twitter contain irrelevant and noisy information. This irrelevant and noisy information has to be cleaned before it is used by other parts of the system. The preprocessed tweets were categorized into positive, negative, or neutral by the opinion analyzer.

Received: 22 Aug 2019 | Revised: 13 Sep 2019 | Accepted: 15 Nov 2019

A. Extraction Of Tweets

Twitter is an amazing social media network for text analysis, sentiment analysis, and social web analysis. Different types of software provide different ways to extract data from Twitter. Among all, **R** offers wide-ranging choices to do many interesting things.

Toextract tweets from Twitter, a twitter application should be created, which will use the Twitter API to make the connection. The Twitter application willprovide the authentication to the system and create your application's Consumer Key, Consumer Secret, Access token and Access token secret. Then, these details will be used in coding to connect and extract data from Twitter.

In Twitter, we can extract tweets for specific purposes because it provides search based on specific keywords or hash tags. Using search keyword technique tweets related to GST
were collected and stored for analysis

B. Tweets Preprocessing

The data collected from Twitter contain irrelevant and noisy information [10], which requires a cleanup before further usage.

The Pre-processor is used for:

- ✓ Removal of non-English letters, duplicate tweets, URLs
- ✓ Removal of special characters, punctuation and white spaces and stop words
- ✓ Stemming and Spell Checking
- ✓ Lowercase conversion and Replacement of Slang with Original Words
- ✓ Tokenization and Separation of Folksonomies

The Flowchart of pre-processor is shown below in figure 2:

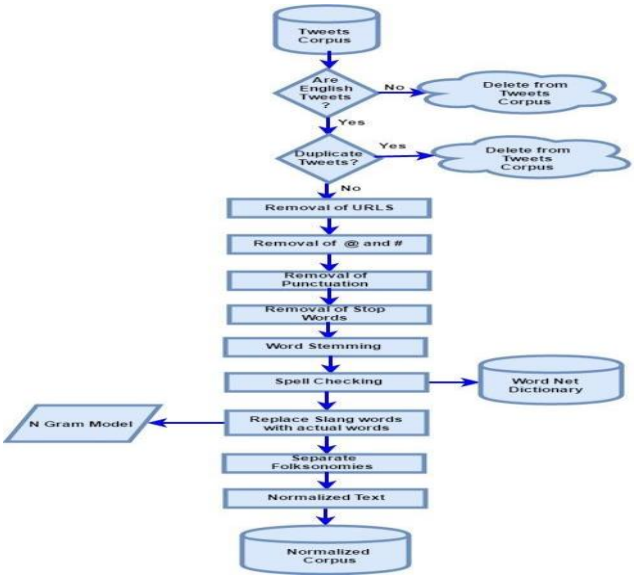

Figure 2: Preprocessor of Opinion Analyzer

Received: 22 Aug 2019 | Revised: 13 Sep 2019 | Accepted: 15 Nov 2019

The extracted Tweets contains huge amount of additional meaningless information which will not add any value to the opinon analyzer. The extracted data contains the unwanted data in following fields:

- createdDate
- favoriteCount
- favorited
- Tweets id
- isRetweet
- latitude
- longitude
- replyToSID
- replyToSN
- replyToUID
- retweetCount
- retweeted
- screenName
- statusSource
- text
- truncated

From the above fields, only "Text" field contains valid information which can be used by opinion analyzer to find the opinon. All other data are considered as noise and those data to be removed before processing. The Preprocessor of opinion Analyzer is used to remove the unnecessary data from the extracted tweets corpus.

C. *Feature Extraction*

The extracted tweets contain the following attributes [11]:

- ✓ Created Date - UTC time (when the Tweet was created)
- ✓ id– a sequence of integers to represent a unique key
- ✓ id_str - string representation of the unique identifier
- ✓ text - actual UTF-8 text posted by the user
- ✓ source - utility used to post the Tweet,
- ✓ truncated - indicates whether the value of the text parameter was truncated or not
- ✓ user - user who posted this Tweet
- ✓ coordinates - represents the geographic location of this Tweet
- ✓ place - indicates the place of Tweet
- ✓ retweet_count - the number of times Tweet has been retweeted

When the Tweets Extractor of the system extract the tweets, all the above attributes are stored in the corpus. Not all the above attributes, will add any value to the opinion analyzer. The feature extractor was used to identify the most relevant attributes that contain valuable information. Only "Text" field contains valid information and the feature extractor extracts that field for further processing.

Table 1: Tweets after Noise Removal and Feature Extraction

Text
RT @bhatia_niraj23: He is Suresh Tanna , A small businessman ..has faced great slowdown after note ban & GST. he has faith in @Office Of RG &…

Received: 22 Aug 2019 | Revised: 13 Sep 2019 | Accepted: 15 Nov 2019

RT @ptrmadurai: 6 years of ADMK Govt, capped by a disastrous GST implementation have seriously damaged TN's Industrial Growth. This article…
RT @ashokgehlot51: Five quarters of slowdown in GDP growth are an indication that both #DeMonetisation and Modi government's #GST proved to…
RT @kukk44: This is not GST, this is Gabbar Singh Tax. Its aim is to take away your hard earned money.
RT @livemint: Gujarat elections: Saurashtra traders miffed after GST nixes 'kutcha' transactions https://t.co/PJRFe5J2MC

D. Opinion Analysis

The opinion analyzer of the system is divided into three parts as follows:

- ✓ Opinion Value Calculator: used to calculate the overall value of the sentence by comparing and assigning each word in the sentence, against a dictionary of positive words and negative words (explained in detail in next section)
- ✓ Opinion Classifier: based on the value assigned to the sentences by the opinion value calculator, the sentences are classified into positive, negative or neutral (detailed explanation in next section)
- ✓ Opinion Summarization: used to summarize all the positive tweets into one group, negative tweets into the second group and neutral tweets into the third group. Also, used to count the total number of tweets in each group on a particular date.

E. Data Visualization

The data Visualizer in the system is used to represent the information in the form of a chart, diagram, or picture. These are depicted in the results section.

IV. IMPLEMENTATION DETAILS

The opinion analyzer finds the user opinion hidden inside the sentences using following three steps:

Step 1: Calculation of Opinion Value of each sentence

Step 2: Classification of sentences into Positive, negative and neutral

Step 3: Summarization and date wise counting of total tweets in each group

A. Method of Opinion Value Calculation

A corpus each of positive words and negative words are maintained. Each corpus contains more than 3000 words. Figure 3a and 3b below shows snapshot of positive and negative words corpus.

Received: 22 Aug 2019 | Revised: 13 Sep 2019 | Accepted: 15 Nov 2019

3a) Positive Words 3b) Negative Words

Figure 3: Snapshot of positive and negative corpus

Firstly, every word(tw) in a tweet (t)is compared against each item of the positive words corpus. If the word(tw) matches the words in positive words corpus the count of 'positive words(pw)' increases by1. Similarly, if it matches the negative words corpus, the count of 'negative words(nw)' increases by 1. The process continues for all the words of all the tweets. At the end of this process, system gives total count of positive words and negative words. Finally, the score of the sentence will be calculated using the formula given below:

$$\boldsymbol{Score(t)} = \boldsymbol{Positive\ Words\ Count(pw)} - \boldsymbol{Negative\ Words\ Count}\ \mathbf{(nw)} \ldots\ldots\ldots \quad \mathbf{(1)}$$

At the end, the system stores the scores of each sentence in a corpus called 'Tweet_Scores'.

For Example, consider a tweet :

"He is Suresh Tanna, A small businessman ...has faced great slowdown after noteban & GST..he has faith in it"

In this tweet, "great" and "faith" are positive words and "slowdown" had considered as negative word.

Total Number of positive words : 2

Total Number of positive words : 1

Score (t) =2 -1 = 1

Scores for all the tweets in the corpus will be calculated in the same way and stores in GST_Scores corpus. The snapshot of this is shown below in figure 4:

Received: 22 Aug 2019 | Revised: 13 Sep 2019 | Accepted: 15 Nov 2019

score	text
2	RT @bhatia_niraj23: He is Suresh Tanna , A small businessman ..has faced great slowdown after noteban & GST.. he has faith in @OfficeOfRG &...
-2	RT @ptrmadurai: 6 years of ADMK Govt, capped by a disastrous GST implementation have seriously damaged TN's Industrial Growth. This article...
0	RT @ashokgehlot51: Five quarters of slowdown in GDP growth are an indication that both #DeMonetisation and Modi government's #GST proved to...
0	RT @livemint: Gujarat elections: Saurashtra traders miffed after GST nixes 'kutcha' transactions https://t.co/PJRFe5J2MC
-2	@PRSLegislative @arunjaitley Pl increase limit of income for tax on salary class drastically in this...

Figure 4: Snapshot of Scores corpus

B. Opinion Classification

The system reads the scores of each tweet from Tweet_Scores corpus. If the score of the sentence is greater than 0, then the sentence is considered as "Positive Sentence ". If the sentence score is less than zero, then the sentence is considered as "Negative Sentence". Otherwise the sentence is considered as "Neutral Sentence". The following algorithm is applied on 'Tweet_Scores' Corpus and the system classifies the tweets into positive, negative or neutral.

For each Tweet(t) in Tweets Corpus
Calculate Score(t) using (1)
Opinion=null
If score(t) > 0 then
Opinion="Positive"
Else if score(t) < 0 then
Opinion="Negative" Else
Opinion="Neutral"

Figure 5: below shows the corpus with score and its corresponding opinion:

Received: 22 Aug 2019 | Revised: 13 Sep 2019 | Accepted: 15 Nov 2019

score	text	opinion
2	RT @bhatia_niraj23: He is Suresh Tanna , A small businessman ..has faced great slowdown after noteban & GST.. he has faith in @OfficeOfRG &...	Positive
-2	RT @ptrmadurai: 6 years of ADMK Govt, capped by a disastrous GST implementation have seriously damaged TN's Industrial Growth. This article...	Negative
0	RT @ashokgehlot51: Five quarters of slowdown in GDP growth are an indication that both #DeMonetisation and Modi government's #GST proved to...	Nuetral
0	RT @livemint: Gujarat elections: Saurashtra traders miffed after GST nixes 'kutcha' transactions https://t.co/PJRFe5J2MC	Nuetral
-2	@PRSLegislative @arunjaitley Pl increase limit of income for tax on salary class drastically in this...	Negative

Figure 6: Snapshot of corpus with score and Opinion

C. Opinion Summarization

The summarizer in the opinion analyzer counts and clusters the tweets into 3 clusters positive, negative and neutral using the

'Tweets created date' as key. The output will be stored in GST_opin Corpus. The snapshot of opinion summarizer is shown in figure 6:

tweet	created	number
negative	06/12/2017	202
neutral	06/12/2017	394
positive	06/12/2017	177
negative	07/12/2017	421
neutral	07/12/2017	961
positive	07/12/2017	424
negative	11/12/2017	175
neutral	11/12/2017	618
positive	11/12/2017	219

Figure 6: Snapshot of corpus of Opinion Summarizer

V. RESULTS EVALUTION

This section deals with the results obtained by Twitter Opinion

Miner. For the purpose of analysis, we collected tweets about GST for one month. The sample dataset contains around 17000 tweets about GST. Sample Tweets extracted using Twitter API is shown below in Table 2.

Table 2: Sample Tweets extracted using Twitter API

Created	Tweets
06/12/2019	RT @bhatia_niraj23: He is Suresh Tanna , A small businessman ..has faced great slowdown after noteban & GST.. he has faith in @OfficeOfRG &

Received: 22 Aug 2019 | Revised: 13 Sep 2019 | Accepted: 15 Nov 2019

06/12/2019	RT @ptrmadurai: 6 years of ADMK Govt, capped by a disastrous GST implementation have seriously damaged TN's Industrial Growth. This article…
06/12/2019	RT @ashokgehlot51: Five quarters of slowdown in GDP growth are an indication that both #DeMonetisation and Modi government's #GST proved to…
06/12/2019	RT @kukk44: This is not GST, this is Gabbar Singh Tax. Its aim is to take away your hard earned money.

The output of opinion analyzer, after finding the opinion value and classifying the tweets into "Positive, Negative and Neutral", is shown in Table3 below.

Table 3: Output of Opinion Analyzer

Tweets	**Positive Words Count(A)**	**Negative Words Count(B)**	**Score = 1A-B**	**Opinion**
He is Suresh Tanna, A small businessman.has faced great **He has faith in GST**… slowdown after note ban &	2 (Great, Faith)	1 (Slowdown)	1	Positive
6 years of ADMK Govt, capped by a disastrous GST implementation have **seriously damaged TN's Industrial Growth.**	0	2 (disastrous, damaged)	-2	Negative

Received: 22 Aug 2019 | Revised: 13 Sep 2019 | Accepted: 15 Nov 2019

Gujarat elections: Saurashtra traders miffed after GST nixes 'kutcha' transactions.	0	0	0	Neutral

The Results of Opinion Analyzer on sample dataset is shown in Table 4 below.

Table 4: Results of Opinion Analyzer on Sample Dataset

Creation Date	Negative	Neutral	Positive	Grand Total
01/12/2019	166	409	129	704
03/12/2019	183	531	200	914
05/12/2019	87	274	107	468
06/12/2019	202	394	177	773
07/12/2019	421	961	424	1806
11/12/2019	175	618	219	1012
12/12/2019	204	421	198	823
13/12/2019	377	933	320	1630
16/12/2019	38	206	95	339
17/12/2019	137	301	191	629
19/12/2019	199	327	249	775
21/12/2019	277	615	322	1214
22/12/2019	376	1047	348	1771
23/12/2019	106	169	124	399
25/12/2019	113	266	110	489
26/12/2019	127	372	112	611
27/12/2019	154	410	165	729
29/12/2019	149	290	128	567
30/12/2019	157	316	101	574
31/12/2019	110	341	139	590
Grand Total	**3758**	**9201**	**3858**	**16817**

Received: 22 Aug 2019 | Revised: 13 Sep 2019 | Accepted: 15 Nov 2019

The various types of visualization of the results of opinion analyzer are depictedin figure 6 below.In these figures, green color denotes “Positive Opinion”, red color denotes “Negative Opinion”, and yellow color denotes “Neutral Opinion”.

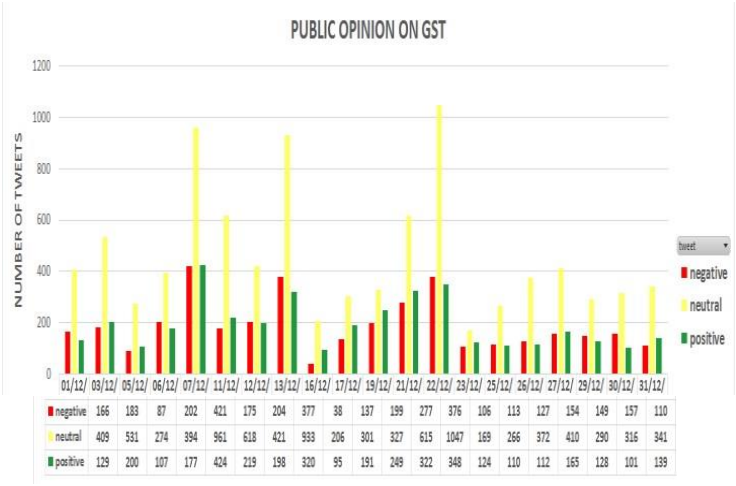

Figure 6: Visualization of Twitter Opinion Analyzer

The data visualition technique on small dataset shows the clear idea about output of Twitter Opinion Analyzer which is shown in Table 5 and also dipicted in figure 7 .

Table 5: Sample Dataset of Opinion Analyzer

Row Labels	negative	neutral	positive	Grand Total
01/12/2019	166	409	129	704
03/12/2019	183	531	200	914
05/12/2019	87	274	107	468
06/12/2019	202	394	177	773
07/12/2019	421	961	424	1806
Grand Total	**1059**	**2569**	**1037**	**4665**

Received: 22 Aug 2019 | Revised: 13 Sep 2019 | Accepted: 15 Nov 2019

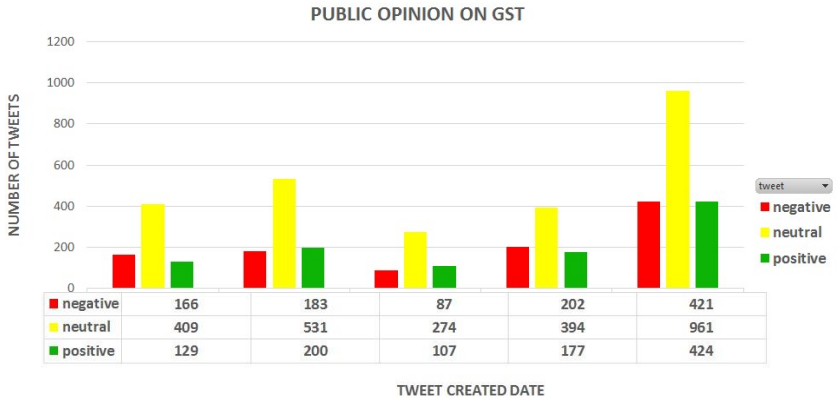

Figure 7: Results of Twitter Opinion Analyzer on small dataset

Table 6: below shows the overall summary of the results of Opinion Analyzer

Table 6: Summary Of Results

Negative	**3758**
Neutral	**9201**
Positive	**3858**
Total Tweets	**16817**

The graph below in figure 8 shows overall result analysis of Opinion Analyzer. The results clearly show that 24 % of users expressed positive opinion about GST. Majority of the respondents (around 54%) had posted neutral comments about GST. 22% of users tweeted negatively about GST.

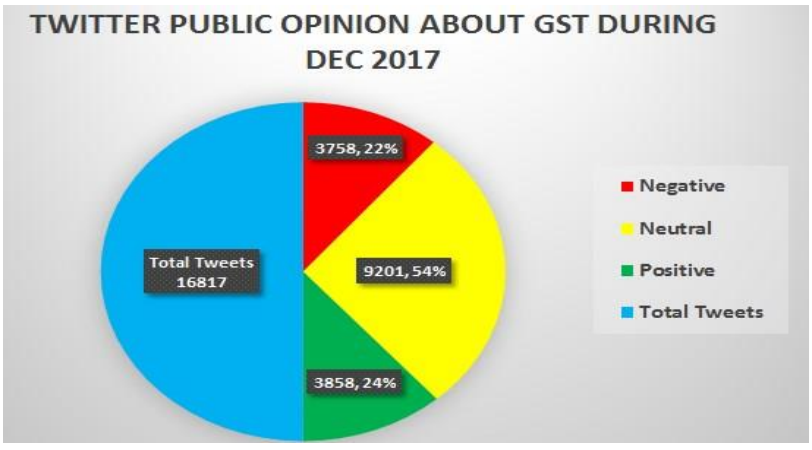

Figure 9: Overall Result Analysis of Twitter Opinion Analyzer

VI. CONCLUSIONS

Received: 22 Aug 2019 | Revised: 13 Sep 2019 | Accepted: 15 Nov 2019

This paper explained the various steps involved in finding user opinion and sentiments on a specific topic or event. This paper explained in detail about the construction of Twitter opinion analyzer. The opinion analyzer extracts real time tweets based on a given search term, removes noise in tweets and constructs the corpus of sample data. Then, the Opinion Analyzer calculates the polarity for each sentence in the corpus. The classifier of the system finally classify the tweets into "Positive", "Negative" and "Neutral" based on the polarity. For this work, sample tweets on GST was collected and processed. In a onemonth span, around 17000 tweets were collected. Opinion Classifier has classified 3758 tweets into "Negative",9201 tweets into "Neutral" and 3858 tweets into "Positive". The results clearly show that 24 % of users expressed positive opinion, around 54% of them had posted neutral comments and 22% of users tweeted negatively about GST.

VII. FUTURE WORK

The rapid growth in volume, velocity, variety of data generated by social media leads to difficulties for the traditional systems with limited storage capacity and computing power. This is one of the main factors for booming of big data. Nowadays, companies use big data analysis to make targeted, real-time decisions to increase profit of the business. Sentiment Analysis with big data framework will help the business to find the appropriate insights and help the business in strategic decision-making.

REFERENCES

1. Lu Lin, Jianxin Li, Richong Zhang, Weiren Yu and Chenggen Sun, "Opinion Mining and Sentiment Analysis in Social Networks: A Retweeting Structure-aware Approach," IEEE/ACM 7th International Conference on Utility and Cloud Computing,2014 pp.890-895.
2. S.Thowseaf, M. Ayisha Millath,"A Study on GST Implementation and its Impact on Indian Industrial Sectors and Export",International Journal of Management Research and Social Science (IJMRSS),Volume 3, Issue 2,June 2016.pp.27-30.
3. Vandana Singh , Sanjay Kumar Dubey, "Opinion Mining and Analysis: A Literature Review", IEEE, 2014 , pp, 232-239.
4. Bernard J. Jansen ,MimiZhang , " Micro-blogging as Online Word of Mouth Branding ," ACM 2009. pp. 15-18.
5. Liang Wu, Teng-Sheng Moh, Natalia Khuri, "Twitter Opinion Mining for Adverse Drug Reactions ," International Conference on Big Data (Big Data)IEEE, 2015, pp. 1750-1753.
6. Peiman Barnaghi, John G. Breslin, "Opinion Mining and Sentiment Polarity on Twitter and Correlation Between Events and Sentiment," IEEE Second International Conference on Big Data Computing Service and Applications 2016, pp. 52-57.
7. Lu Lin, Jianxin Li, Richong Zhang, Weiren Yu and Chenggen Sun, "Opinion Mining and Sentiment Analysis in Social Networks: A Retweeting Structure-aware Approach," IEEE/ACM 7th International Conference on Utility and Cloud Computing,2014 pp.890-895.
8. Lei Wang and John Q Gan,"Prediction of the 2019 French Election Based onTwitter Data Analysis",Computer Science and Electronic Engineering (CEEC), IEEE 2019,pp89 - 93.

Received: 22 Aug 2019 | Revised: 13 Sep 2019 | Accepted: 15 Nov 2019

9. Roshan Fernandes, Dr. Rio D'Souza, "Analysis of Product Twitter Data though Opinion Mining," IEEE Second 2016 IEEE Annual India Conference (INDICON) 2016, pp. 1-5.
10. Mr.SanketPatil, Prof.VarshaWangikar, Prof. K. Jayamalini,"Data
11. Preprocessing, Sentiment Analysis & NER On Twitter Data," IOSR Journal of Computer Engineering (IOSR-JCE), International conference on computing and virtualization (ICCCV-2019),pp 73 -79.
12. https://developer.twitter.com/en/docs/tweets/datadictionary/overview/tweet-object(Accessed: dec, 2019)

Received: 22 Aug 2019 | Revised: 13 Sep 2019 | Accepted: 15 Nov 2019