# Sentiment Analysis on GST Using Naive Bayes and Score Based Approaches on Twitter Data

Rohini Kancharapu,[1] Dr. A. Sri Nagesh,[2]

[1]*Ph.D. Scholar,CSE Department, ANU College of Engineering and Technology,Acharya Nagarjuna University,Nagarjuna Nagar,*
*Guntur-522 510, AP, India.*
[2]*CSE Department, RVR&JC CoE, Chowdavaram, Guntur-522019, AP, India.*
*(E-mail: rohini541@gmail.com, asrinagesh@gmail.com )*

*Abstract*— Sentiment Analysis (SA) extracts and analyses people's opinion about an entity. SA in twitter tackles the problem of analyzing the tweets. Twitter is one of the popular micro blogging platform in which users can publish their thoughts and opinions. It has been used as a forum to understand the opinions of public towards recently launched Goods and Services Tax (GST) by Indian Government. The tweets originated have been analyzed using supervised learning. This type of learning uses unlabelled data to complement the information provided by the labeled data in the training process. SA agrees with the human judgment and determines at sentence level whether the opinions arrives at a decision and perform classification on emotion and polarity using naive bayes and score based algorithms to find the effect of GST on public.

Keywords— Sentiment analysis, Sentiment Polarity, Emotion detection, Twitter, GST

## I. INTRODUCTION

### *A. Problem Definition*

Implementing how sentiment analysis can help to improve the user experience over:

The recent availability of huge amounts of "user-generated content" called "tweets" on the web, produce a need to mine user's opinion on GST.

Familiarize with classification and emotion that enable to apply processing and machine learning with textual data using Naive Bayes and score based algorithm.

### *B. Motivation*

In the past decade, new forms of communication, such as micro blogging and text messaging have emerged and become ubiquitous. While there is no limit to the range of information conveyed by tweets and texts, often these short messages are used to share opinions and sentiments that people have about what is going on in the world around them.

Another aspect of social media data such as Twitter messages is that it includes rich structured information about the individuals involved in the communication. For example, Twitter maintains information of who follows whom and re-tweets and tags inside of tweets provide discourse information.

### *C. What is GST?*

Goods and services tax (GST) is a comprehensive indirect tax, which to be implemented in India from April 2016. It is an indirect tax on sales, manufactures, services, and consumption of goods throughout India. Previously there were different types of tax levied by both central government and state government. It was bit complicated and very difficult to understand. GST is a step to replace those complicated taxes and now it would be only one tax throughout all over India at each stage of sale or purchase of goods or services based on input tax credit system.

### *D. When GST started?*

Actually this proposal was given in the time of our former Prime Minister Atal Bihari Vajpayee Govt. In 2000 government set up a committee to design an appropriate model for GST and it was headed by the Union Finance minister of West Bengal Mr Asim Dasgupta.

### *E. What would be the Tax rate?*

Under the proposed GST the tax rate would be less but the number of assesses would be increased by 5 to 7 times. Tax collection would be also go up due to tax buoyancy.

### *F. What is the reaction of People towards GST?*

People's reaction plays a very important role in implementing this amendment. Thus we tried to find out the sentiment of public. Now day's social media is the primary and important medium of sharing thoughts and opinions.

## II. GST SENTIMENT ANALYSIS

The mounting habit of social media has elevated the prospect of exploring and tracking the response of new reforms and policies in India. Social media has been used profoundly all over the world for analysis of political campaigns, stock market data, new product launch, movie release etc. Recently, Government of India implemented Goods and Services Tax (GST) which replaced many cascading indirect taxes levied by central and state governments. The GST, India's biggest tax reform in 70 years of independence, was launched on the midnight of 30 June 2017 by the Prime Minister of India Narendra Modi. The launch was marked by a historic midnight (June 30-July 1, 2017) session of both the houses of parliament convened at the Central Hall of the Parliament.

In this paper, we present a simple and robust work to gather, analyze and graphically represent people's opinion about India's new taxation system using Naive Bayes and score based algorithm. Here current 10,000 tweets which have been posted till today will be analyzed. The data collected is normalized by removing punctuations, special characters, and emojis, wherever needed. The data is then analyzed through a statistical software R. A word cloud is generated from the set of 10,000 tweets focusing on GST. Furthermore, the words are compared against negative and positive words list from which frequency bar plot is generated to get the generalized view.

*A. SA using Naive Bayes Algorithm*

The Naive Bayes classifier is the simplest and most commonly used classifier.Naive Bayes classification model computes the posterior probability of a class, based on the distribution of the words in the document. The model works with the BOWs feature extraction which ignores the position of the word in the document. It uses Bayes Theorem to predict the probability that a given feature set belongs to a particular label.

Positive sentiments did seem to have a upward effect on GST. This was found by prediction method: data was divided into 2 sets training and testing (70%–30%). The positive sentiments were more predictive than negative.[9] P(label) is the prior probability of a label or the likelihood that a random feature set the label. P(features|label) is the prior probability that a given feature set is being classified as a label. P(features) is the prior probability that a given feature set is occurred. Given the Naive assumption which states that all features are independent.

An improved NB classifier was proposed by Kang and Yooto solve the problem of the tendency for the positive classification accuracy to appear up to approximately 10% higher than the negative classification accuracy. This creates a problem of decreasing the average accuracy when the accuracies of the two classes are expressed as an average value.

*B. Algorithm for Naive Bayes*

*Input* - A document mytweet_text  
A fixed set of classes C={c1,c2,...,cj}  
Output - Classification of tweets on polarity and emotion.

Steps:

1. Pre-processing
 i. About 10,000 reviews were extracted using twitter API.
 ii. Emotions and polarity in the sentence reviews were broken and those were appended to lists created .
 iii. 3⁄4 of these sentences were kept in the dictionary for training while the 1⁄4 were kept for testing.
2. The classifier was trained using the dataset just prepared.
3. Labeled sentences were kept correctly in reference sets and the predicatively labeled version in test sets.
4. Metrics were calculated accordingly.

*Emotion detection*

Sentiment analysis is sometimes considered as an NLP task for discovering opinions about an entity; and because there is some ambiguity about the difference between opinion, sentiment and emotion, they defined opinion as a transitional concept that reflects attitude towards an entity. The sentiment reflects feeling or emotion while emotion reflects attitude.

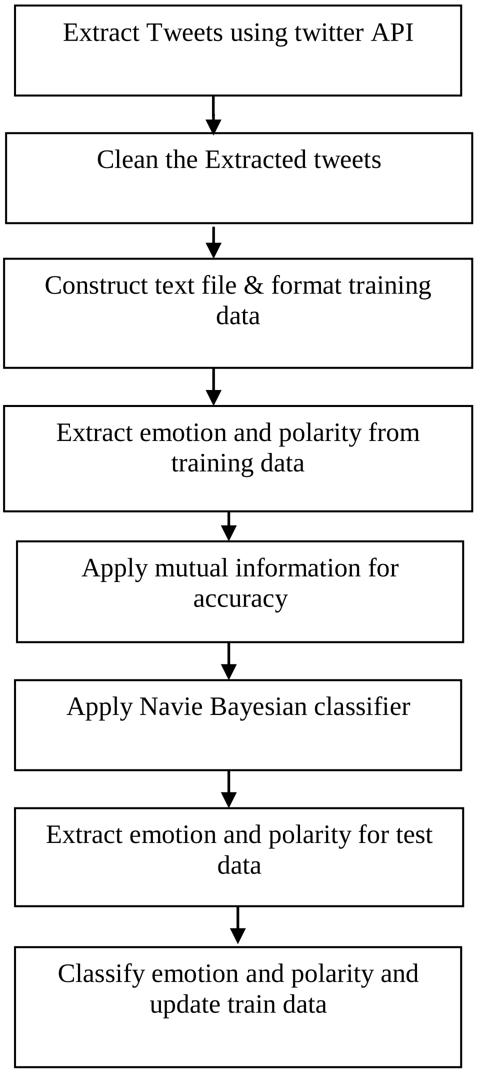
Extract Tweets using twitter API
Clean the Extracted tweets
Construct text file & format training data
Extract emotion and polarity from training data
Apply mutual information for accuracy
Apply Navie Bayesian classifier
Extract emotion and polarity for test data
Classify emotion and polarity and update train data

FIG.1. DATA FLOW OF NAVIE BAYES

### *C. Example of Naive Bayes*

Ken wants to buy IPhone10. She don't know if it is a great phone to buy it or not.

TABLE 1. SAMPLE DATA SET

| Documents | Words | Class |
|---|---|---|
| 1 | Don't Buy | Negative |
| 2 | Phone got Hanged | Negative |
| 3 | Battery Drains Fast | Negative |
| 4 | Durable Phone | Positive |
| 5 | Great Camera | Positive |
| 6 | Great Phone buy it | ------- |

She wants to know if the review "Great Phone buys it "class is positive or Negative.

Solution:
Step-1: Determine the Train Set and Test Set in your Data Set

TABLE 2. TRAIN SET

| Documents | Words | Class |
|---|---|---|
| 1 | Don't Buy | Negative |
| 2 | Phone got Hanged | Negative |
| 3 | Battery Drains Fast | Negative |
| 4 | Durable Phone | Positive |
| 5 | Great Camera | Positive |

TABLE 3. TEST SET

| Documents | Words | Class |
|---|---|---|
| 6 | Great Phone buy it | ----?--- |

Step-2 : Convert the dataset into frequency table

TABLE 4. FREQUENCY TABLE

| Words | Positive | Negative |
|---|---|---|
| Don't | 0 | 1 |
| Buy | 0 | 1 |
| Phone | 1 | 1 |
| Got | 0 | 1 |
| Hanged | 0 | 1 |
| Battery | 0 | 1 |
| Drains | 0 | 1 |
| Fast | 0 | 1 |
| Durable | 1 | 0 |
| Great | 1 | 0 |
| Camera | 1 | 0 |

Step-3: Compute the Prior Probability
Pr(c) = $N_c$/N
Where P(c) – Probability of the class
$N_c$ – Total Count of a particular class in the training set
N – Total Count of a class in the Training Set
<u>Prior</u>: Pr(Positive) = 2/5 =0.4
Pr(Negative) = 3/5 = 0.6

Step-4 :Compute the conditional probability/likelihood of each word attribute
P(W/C)= (Count(W,C)+1)/(Count(c)+|v|)
Where
P(W/C) – Conditional Probability
W – Word Attribute
C – Class
Count(W,C) – Total Count of word attribute occurs in class C
+1 – Laplace Smoothing
Count(c) – Total Count of word attribute in a particular class occurs in a training set
|v| - Vocabulary – The total Count of different word attribute in the training set

Conditional Likehood Probability (P(W/C))
P(Great/Positive) = (1+1)/(4+11) = 0.13
P(Phone/Positive) = (1+1)/(4+11) = 0.13
P(Buy/Positive) = (0+1)/(4+11) = 0.07
P(It/Positive) = (0+1)/(4+11) = 0.07
P(Great/Negative) = (0+1)/(8+11) = 0.05
P(Phone/Negative) = (1+1)/(8+11) = 0.11
P(Buy/Negative) = (1+1)/(8+11) = 0.05
P(It/Negative) = (0+1)/(8+11) = 0.11

Step-5:Compute the posterior probability
$C_{map}$ = argmax P($X_1$,$X_2$,………,$X_n$) Pr(C) where c E C
Where P($X_1$,$X_2$,………,$X_n$) - Conditional Probability
Pr(C) – Prior Probability of the Class
posterior probability
P(Positive)=(P(Great/Positive)*P(Phone/Positive)*P(Buy/Positive)*P(It/Negative))*Pr(Positive)
= (0.13 * 0.07 * 0.07 * 0.13) * 0.4
= 0.00003
P(Negative)=(P(Great/Negative)* P(Phone/Negative) * P(Buy/Negative) *
P(It/Negative)) * Pr(Negative)
= (0.05 * 0.11 * 0.11 * 0.05) * 0.6
= 0.00002
Step-6:Determine the class of the Test Set
If the posterior probability of a class is greater than that class will be a result
P(Positive) > P(Negative)
Thus, Class of the Test Set is Positive.
Therefore, Ken Decided to buy IPhone10.

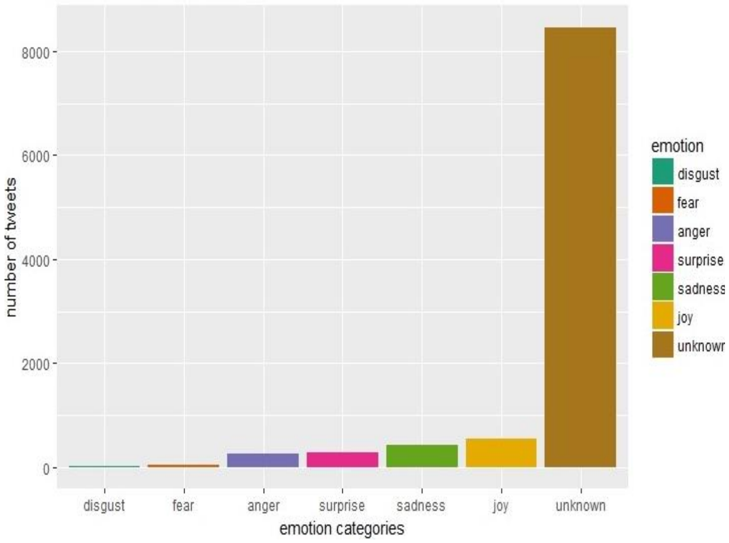

FIG. 2. CLASSIFICATION OF EMOTION ON GST USING NAIVE BAYES

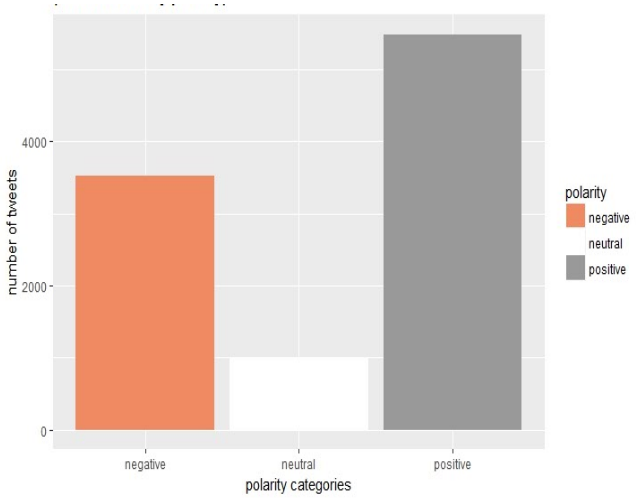

FIG. 3. CLASSIFICATION OF POLARITY ON GST USING NAIVE BAYES

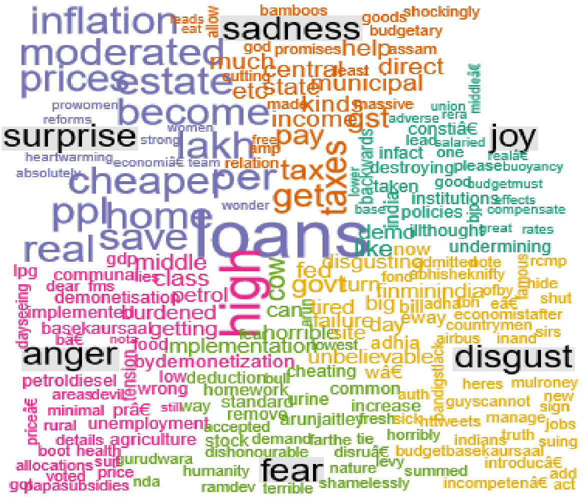

FIG. 4. EMOTION COMPARISON WORD CLOUD

From the sentiment analysis of “GST” tweets, it is very clear that there is a feeling of anticipation for “GST”. There is almost equal joy and sadness emotion for “GST” in India.

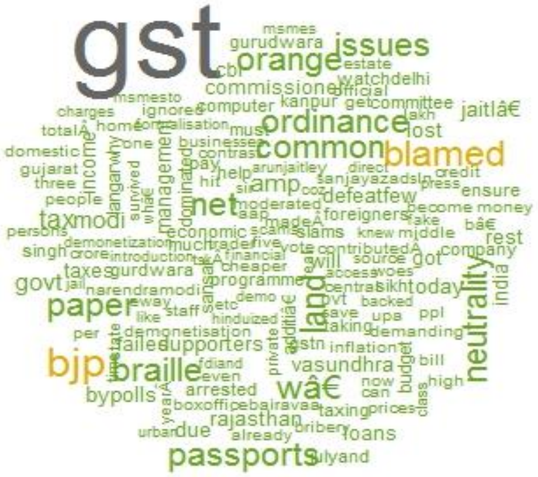

FIG.5. GST WORD CLOUD

*C. SA using Score Based Algorithm*

Sentiment Analysis is the process of determining whether a piece of writing is positive, negative or neutral. It’s also known as opinion mining, deriving the opinion or attitude of a speaker. A common use case for this technology is to discover how people feel about a particular topic.SA is widely applied to reviews and social media for a variety of applications, ranging from marketing to customer service.

We will classify the sentiment of a tweet based on the polarity of the individual words. Each word will be given a score of +1 if classified as positive, -1 if negative, and 0 if classified as neutral. The total polarity score of a given tweet will result in adding together the scores of all the individual words in a sentence.

The polarity score is not always very accurate. It sometimes misses out on the overall context of the tweet because it focuses on individual words. Sometimes words like ‘ohhh’ can be used as positive or negative.

*D. Sentiment Function*

Once we have the tweets we just need to apply some functions to convert these tweets into some useful information. The main working principle of sentiment analysis is to find the words in the tweets that represent positive sentiments and find the words in the tweets that represent negative sentiments.

Firstly the tweets are compared to positive and negative words. Then the score is calculated using keyword score sentiment for each tweet. A graph is plotted on comparing the score and the polarity of each tweet. The below figure describes the methodology of Score Based Algorithm.

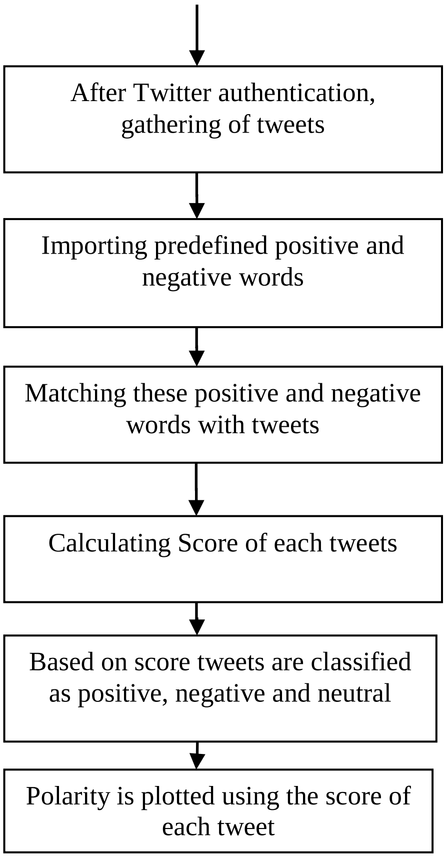
After Twitter authentication, gathering of tweets
Importing predefined positive and negative words
Matching these positive and negative words with tweets
Calculating Score of each tweets
Based on score tweets are classified as positive, negative and neutral
Polarity is plotted using the score of each tweet

Fig. 6. Flow Chart of Score Based Algorithm

*E. Score Based Algorithm*

*Input* - A document mytweet_text

A fixed set of classes C={c1,c2,...,cj}

*Output* - Classification of tweets on polarity.

*Steps:*

1. About 10,000 reviews were extracted using twitter API.
2. Gathering and Importing Predefined positive and negative words
3. Matching these positive and negative words with tweets
4. Calculating Score of each tweets
5. Based on score tweets are classified as positive, negative and neutral
6. Classify polarity using sentiment function.

*F. Example for Score Based Algorithm*

Ken wants to buy IPhone10. She doesn't know if it is a great phone to buy it or not.

TABLE 5. SAMPLE REVIEW DATA

| Documents | Words |
|---|---|
| 1 | Don't Buy |
| 2 | Phone got Hanged |
| 3 | Battery is good and Drains Fast and more |
| 4 | Durable Phone |
| 5 | Great Camera |
| 6 | Great Phone buy it |

Solution

Step-1: Gather the Reviews and Start Cleaning the reviews

Reviews will be:

1. Don't Buy
2. Phone got Hanged
3. Battery is good and Drains Fast and more
4. Durable Phone
5. Great Camera
6. Great Phone buy it

Step-2: Import the predefined positive and negative files and compare with cleaned reviews.

TABLE 6. +VE / -VE WORDS COLLECTION

| Positive Words | Negative Words |
|---|---|
| Good | Bad |
| Fast | Hanged |
| Bright | Drained |
| Durable | Don't |
| Great | Less |
| Better | Worst |
| Smart | |
| More | |

Step-3: Count the no. of positive and negative words in the reviews

TABLE 7. COUNT OF +VE / -VE WORDS IN REVIEW DATA

| Reviews | Positive Words Count | Negative Words Count |
|---|---|---|
| Don't Buy | 0 | 1 |
| Phone got Hanged | 0 | 1 |
| Battery is good and Drains Fast and more | 3 | 1 |
| Durable Phone | 1 | 0 |
| Great Camera | 1 | 0 |
| Great Phone to Buy it | 1 | 0 |

Step–4: Calculate the score of each and every review.

Score = No. of Positive words – No. of negative Words

TABLE 8. SCORE CALCULATION OF REVIEW DATA

| Reviews | Score |
|---|---|
| Don't Buy | -1 |
| Phone got Hanged | -1 |
| Battery is good and Drains Fast and more | 2 |
| Durable Phone | 1 |
| Great Camera | 1 |
| Great Phone to Buy it | 1 |

Step-5: Based on score we classify the reviews as positive , negative and neutral.

If score >0 then review is positive

score<0 then review is negative

score = 0 then review is neutral

Step -6: Based on the score the we come up with an decision.
Out of 6 reviews 4 reviews are Positive, 2 reviews are Negative.
So finally we get a positive feedback.
Therefore, Ken decided to buy IPhone10.

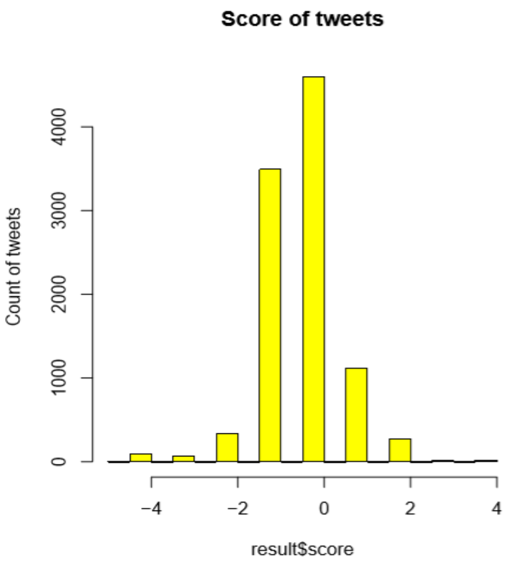

FIG.7. SCORE OF EACH TWEET USING SCORE BASED ALGORITHM

Figure7 describes histogram of Live Twitter Data for GST, depicting sentiments

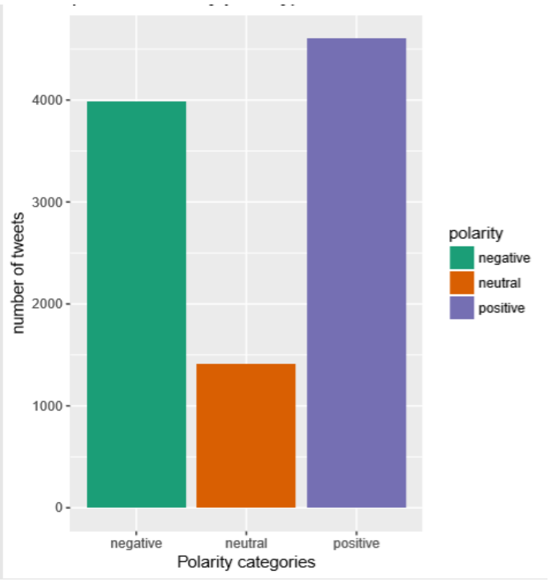

FIG.8. POLARITY VISUALIZATION USING SCORE BASED ALGORITHM

## III. CONCLUSION

With the rapidly expanding social networks, it is challenging to analyze its large data using existing data mining tools. It can be shown through algorithms to do Sentiment Analysis on retrieved "GST" data from Twitter that the numbers of people have given polarity and emotions. With this, it is advisable to conclude R Statistical Tool is sufficiently used for the analysis of Streaming data.

After performing the algorithms, Naive bayes and Score based algorithm there is a high positive polarity in Naive bayes algorithm related to Score based algorithm on retrieving 10000 tweets from Twitter API. So, it can be conclude that GST has a positive response in people point of view. Though there is a difference in the polarity range in both the algorithms, positive effect on GST can be obtained.

## IV. FUTURE ENHANCEMENT

Sentiment Analysis is a field that is still being studied, although not at great lengths due to the intricacy of this analysis. That is this field has functions that are too complicated for machines to understand. The ability to understand sarcasm, hyperbole, positive feelings, or negative feelings has been difficult, for machines that lack feelings. Algorithms have not been able to predict with more than 60% accuracy the feelings portrayed by people. Yet with so many limitations this is one field which is growing at great pace within many industries. Companies want to accommodate the sentiment analysis tools into areas of customer feedback, marketing, CRM, and ecommerce.

The positive or negative word might mean completely opposite depending upon the context used in the sentence. Then sometimes the sentence ambiguity can be a problem since some positive or negative words might mean nothing in perspective of the sentence and sometimes words with no individual meaning express a lot of sentiment in the sentence. Sarcasm is the biggest challenge that sentiment analysis faces. Machine or algorithms with no emotion will find it extremely difficult to differentiate when users are commenting sarcastically.

The language used throughout social media is different. This makes it hard for any tool to predict the emotion or semantic of the sentence. People also use a lot of slang language and hash tags which makes the accuracy of the algorithms lower. It is difficult for the tool to even understand who the object of the sentence is. So, the future enhancement will be based upon the above constraints and algorithms are developed.

## REFERENCES

[1] Vidisha M. Pradhan ,Jay Vala,Prem Balani "A Survey on Sentiment Analysis Algorithms for Opinion Mining", International Journal of Computer Applications (0975 – 8887) Volume 133 – No.9, January 2016.

[2] Faiza Belbachir, B´en´edicte Le Grand, “Opinion Detection: Influence Factors”, IEEE, 2015.

[3] Reshma Bhonde, Binita Bhagwat, Sayali Ingulkar, Apeksha Pande,"Sentiment Analysis Based on Dictionary Approach", International Journal of Emerging Engineering Research and Technology Volume 3, Issue 1,2015

[4] Diksha Sahni, Gaurav Aggarwal, “Recognizing Emotions and Sentiments in Text: A Survey”, International Journal of Advanced Research in Computer Science and Software Engineering , 2015.

[5] K.V.Kanimozhi1, Dr.M.Venkatesan, “Unstructured Data Analysis - A Survey”
International Journal of Advanced Research in Computer and Communication Engineering, Vol. 4, Issue 3, March 2015.

[6] Walaa Medhat, Ahmed Hassan, Hoda Korashy , "Sentiment analysis algorithms and applications: A survey" Ain Shams engineering Journal (April-2014).

[7] Richa Sharma, Shweta Nigam, Rekha Jain, "Polarity Detection at Sentence Level", International Journal of Computer Applications, Volume 86- No 11, 2014.

[8] Pang B, Lee L. Opinion mining and sentiment analysis. Found Trends Inform Retriev 2008;2:1–135.

[9] Cambria E, Schuller B, Xia Y, Havasi C. New avenues in opinion mining and sentiment analysis. IEEE Intell Syst 2013;28:15–21.

[10] Feldman R. Techniques and applications for sentiment analysis. Commun ACM 2013;56:82–9.

[11] Montoyo Andre´s, Martı´nez-Barco Patricio, Balahur Alexandra.Subjectivity and sentiment analysis: an overview of the current state of the area and envisaged developments. Decis Support Syst 2012;53:675–9.

[12] Nath Banamali, "Goods and services tax: A milestone in Indian economy", IJAR 3.3, pp. 699-702, 2017.

[13] Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, Manfred Stede, "Lexicon-based methods for sentiment analysis", Computational linguistics, vol. 37, no. 2, pp. 267-307, 2011.

[14] Zhu Maoran et al., Identification of Opinion Leaders in Social Networks Based on Sentiment Analysis: Evidence from an Automotive Forum, 2016.

[15] Gautama Geetika, Divakar Yadav, "Sentiment analysis of twitter data using machine learning approaches and semantic analysis", Contemporary computing (IC3) 2014 seventh international conference on. IEEE, 2014.

[16] Cliche Mathieu, BB_twtr at SemEval-2017 Task 4: Twitter Sentiment Analysis with CNNs and LSTMs, 2017.

[17] Pabreja Kavita, "GST sentiment analysis using twitter data", IJAR 3.7, pp. 660-662, 2017.

[18] [online] Available: https://wearesocial.com/special-reports/digital-in-2017-global-overview

Ph.D. Scholar, CSE Dept.,
ANU College of Engineering and Technology, Acharya Nagarjuna University,
Nagarjuna Nagar,
Guntur - 522 510, AP, India.
**Research Interests** :Data mining, Machine Learning

**Designation:** Associate Professor
**Research Interests:** Image Processing and Web Technologies
**Qualifications:**
2015 Ph.D, JNTUH, Hyderabad, Andhra Pradesh, India.
2001-2002 M.Tech, VIT, TamilNadu, India.