SENTIMENT IDENTIFICATION IN TWEETS ON SOCIO-ECONOMIC EVENTS

Thesis submitted in the Partial Fulfillment of the Requirements

for the Degree of **M.Tech in Computer Technology**

in the Faculty of Engineering and Technology

Jadavpur University

2019

by

SUKANYA BASU

Examination Roll No.: M6TCT19024

Registration No.: 137127 of 2016-2017

Under the Guidance of

Dr. Chitrita Chaudhuri

Associate Professor

Department of

Computer Science and Engineering

Jadavpur University

Kolkata-700 032

COMPUTER SCIENCE AND ENGINEERING
DEPARTMENT
FACULTY OF ENGINEERING AND TECHNOLOGY
JADAVPUR UNIVERSITY

I hereby forward the thesis entitled "**SENTIMENT IDENTIFICATION IN TWEETS ON SOCIO-ECONOMIC EVENTS**" prepared by Sukanya Basu under my supervision to be accepted in partial fulfillment of the degree of **Master of Technology in Computer Technology** in the Faculty of Engineering and Technology of Jadavpur University,Kolkata.

_____
(Dr. Chitrita Chaudhuri)
Associate Professor
Thesis Supervisor
Dept. of Computer Science and Engineering
Jadavpur University
Kolkata-32

Countersigned:

_____
(Prof. Mahantapas Kundu)
Head of the Department
Dept. of Computer Science and Engineering
Jadavpur University
Kolkata-32

_____
(Prof. Chiranjib Bhattacharjee)
Dean
Faculty of Engineering and Technology
Jadavpur University
Kolkata-32

JADAVPUR UNIVERSITY

FACULTY OF ENGINEERING AND TECHNOLOGY

CERTIFICATE OF APPROVAL*

The foregoing thesis, "**SENTIMENT IDENTIFICATION IN TWEETS ON SOCIO-ECONOMIC EVENTS**" is hereby approved as a creditable study of an engineering subject carried out and presented in a manner satisfactory to the degree for which it has been submitted. It is understood that, by this approval the undersigned do not necessarily endorse or approve any statement made, opinion expressed or conclusion drawn therein, but approved the thesis only for the purpose for which it has been submitted.

_____

Final Examination for

evaluation of the thesis.

_____

(Signature of Examiners)

*Only in case the thesis is approved.

DECLARATION OF ORIGINALITY AND COMPLIANCE OF ACADEMIC ETHICS

I hereby declare that this thesis contains literature survey and original research work by the undersigned candidate, as part of her Master of Computer Technology studies.

All information of this document has been obtained and presented in accordance with academic rules and ethical conduct.

I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name	**:** Sukanya Basu
Roll Number	**:** M6TCT19024
Registration Number	**:** 13127 of 2016-2017
Thesis Title	: Sentiment Identification in Tweets on Socio-Economic Events
Signature with Date	:

ACKNOWLEDGEMENT

I am pleased to express my gratitude and regard to my Project Supervisor, Dr. Chitrita Chaudhuri of CSE Department, J.U. for her invaluable guidance, constant encouragement and inspiration during the period of my project.

I am thankful to respected teacher, Mrs.Chhanda Roy, my senior, Anupam Baidya and my friend, Rishi Dey for their co-operation and help.

My deepest gratitude is to my family to their kind support and encouragement.

Date:

(Sukanya Basu)

Examination Roll No: M6TCT19024

Registration No.:13127 of 2016-2017

Index

Chapters	Page No.
1. Introduction	1
2. Previous Research Works	3
3. Basic Concepts	5
3.1 Natural Language Processing	5
3.2 Sentiment Analysis	5
3.3 Tokenization and POS Tagging	7
3.4 Stop Words Removal	8
3.5 Correlation within Text Documents	9
4. Methodology	10
4.1 Tweet Extraction	11
4.2 Stop words Removal	12
4.3 POS Tagging	12
4.4 Generation of Sentitokens	13
4.5 Computation of Sentiment Score	14
4.6 Finding average scores per word	15
4.7 Calculation of Positive and Negative Scores per tweet	16
4.8 Normalization of Positive and Negative Scores	16
4.9 Evaluation of Resultant Score per tweet	17
4.10 Indication of Sentiment of Tweet	17
4.11 Generation of both Scores and Counts of Total Positive, Negative and Neutral Tweets	18
4.12 Correlation among events	20

5. Experimental Configuration 21
5.1 Datasets 21
5.2 TwitterWebsiteSearch-master 21
5.3 Natural Language Toolkit 22
5.4 SentiWordNet 23
5.5 CSV Module 23
5.6 Harware Configuration 24
5.7 Software Requirements 24
6. Results and Performances 25
6.1 Sentiment Score Graph and Tweet Count Graph 25
6.2 Cosine Similarities 31
7. Conclusion and Future Scope 33
8. Reference 34

1. Introduction

Sentiment Analysis in Tweets on some Socio-Economic Events is carried out to find out the overall impact of each of the event on public over time.

The increment of the number of Tweets with Positive Sentiment compared to the number of Tweets with Negative Sentiment indicates that a higher percentage of the public is supporting the implementation process of the event. Whereas, if the number of Tweets with Negative Sentiments is more than number of Tweets with Positive Sentiment then it indicates that maximum number of people are not happy with the outcome of the event.

Sentiment Analysis is utilized here to develop insights of social trend.

Constant monitoring of the conversations taking place in twitter also helps to detect the ill-effects of instigations and provocations. Preventive measures do not fall into the purview of the present work. But awareness can be generated by depicting and predicting probable outcomes from existing trends.

Sentiment Analysis of tweets can also help in finding out people's inclinations and needs. To be more precise, an overall assessment of the positive and negative sentiments can indicate the turn of the event.

Sentiment Analysis is being currently put to many uses such as 'reputation management', 'opinion mining', 'customized marketing' etc. Majority of the tweeters observe trends before taking their personal decision. For example, negative reviews of a particular product make people be more cautious. Social

media monitoring as well as public sentiment analysis thus prevent one from being taken for false rides by unscrupulous merchants.

The goal of the present work is to utilize sentiment analysis of tweets in order to broadcast the voice of public over events which are of paramount significance in today's context. It helps to assess their views during and after such an event has occurred.

It is required to find out how people's sentiment towards an event varies with time. Thus, a temporal relationship automatically evolves. Positive comments signify that the population is satisfied with the event whereas negative ones indicate people's dissatisfaction. It is also observed that the peak response time often coincides with the commencement of the event and decays with time, sometimes reaching a crescendo due to certain incidents. This automatically reflects the mood of the public which has been captured in our present work. Based on the sentiment outcome, results of public events such as Elections, Government Policies, Economic Trends can be predicted with a certain degree of accuracy.

The rest of the document is organized as follows: Chapter 2 contains discussions on previous research work. Chapter 3 provides basic concepts on which the system is modeled. Chapter 4 illustrates the methodology used to implement the system. In Chapter 5 we provide the details of the software and hardware tools utilized. The results are discussed in Chapter 6. The conclusions drawn on the subject appear in Chapter 7. It also includes some future scope of the work. The references cited in the work are placed at the end.

2. Previous Research Work

Cristina Ioana Muntean in her conference paper [1], discovered that subject of tweets could be found through hashtags and the terms present in tweets. In this case, distribution of hashtags in each and every tweets were taken into account. Here various datasets were collected through the Twitter Streaming API for a period of three days.

Hidenao Abe in his journal [2] focused on the temporal behaviors of the Twitter service known as “retweeting”. Users’ tagged retweets are affected by the content of the received tweets and their history of tweets. In order to predict such targeted tweeting behavior of followers, a model is constructed, for which one should set up more proper features to consider the history of their tweets.

Zhao Jianqiang in his journal [3], discussed the effects of six text pre-processing method on sentiment classification performance using feature models and classifiers on Twitter datasets. To identify the sentiment polarity, most existing approaches apply text pre-processing to reduce the amount of noise in the tweets. This improved the performance of the classifier and speeded up the classification process.

Pang B. and Lee L. has elaborated on sentiment classification of two types – one in terms of either objective or subjective - and the other categorized as positive, negative or neutral [4].

A. Pappu Rajan in [5] determined the positive or negative sentiment of text which extended to strength of polarity. This included data set collection, reading of opinion dataset, removal of noise data, splitting of opinion sentences into opinion word, finding the positive and negative opinion. Number of positive and negative opinions from multi set are compared. Score of opinion is measured as the difference between the number of positive words and the number of negative words.

Shailendra Kumar Singh in [6] stated that in sentiment analysis process, negation words and negative prefixes have potential to reverse the sentiment of sentences. Part-of-Speech (POS) tagging information and opinion words and phrases are used for sentiment extraction. The opinion words and opinion phrases are used to extract positive / negative sentiments. There are two approaches – one lexicon-based and the other statistical-based.

Anusha K S in [7] discussed about the sentiment analysis of twitter data. This involves data collection, data pre-processing, feature extraction, sentiment analysis, polarity classification into positive, negative and neutral.

3. Basic Concepts

3.1 Natural Language Processing (NLP)

NLP involves computational treatment of human language. It teaches computer the method to understand and generate human language.

NLP is used for massive management of textual information sources that is required for human use which is essential for human use.

NLP is used for text classification which is an essential part in many applications, such as web searching, information filtering and sentiment analysis.

3.2 Sentiment Analysis

Sentiment Analysis is the computational study of people's opinions, appraisals, and emotions toward entities, events and their attributes.

Sentiment Analysis involves subjectivity analysis of a statement and then emotion identification that get expressed through the statement.

The process is depicted in the following figures 1 and 2

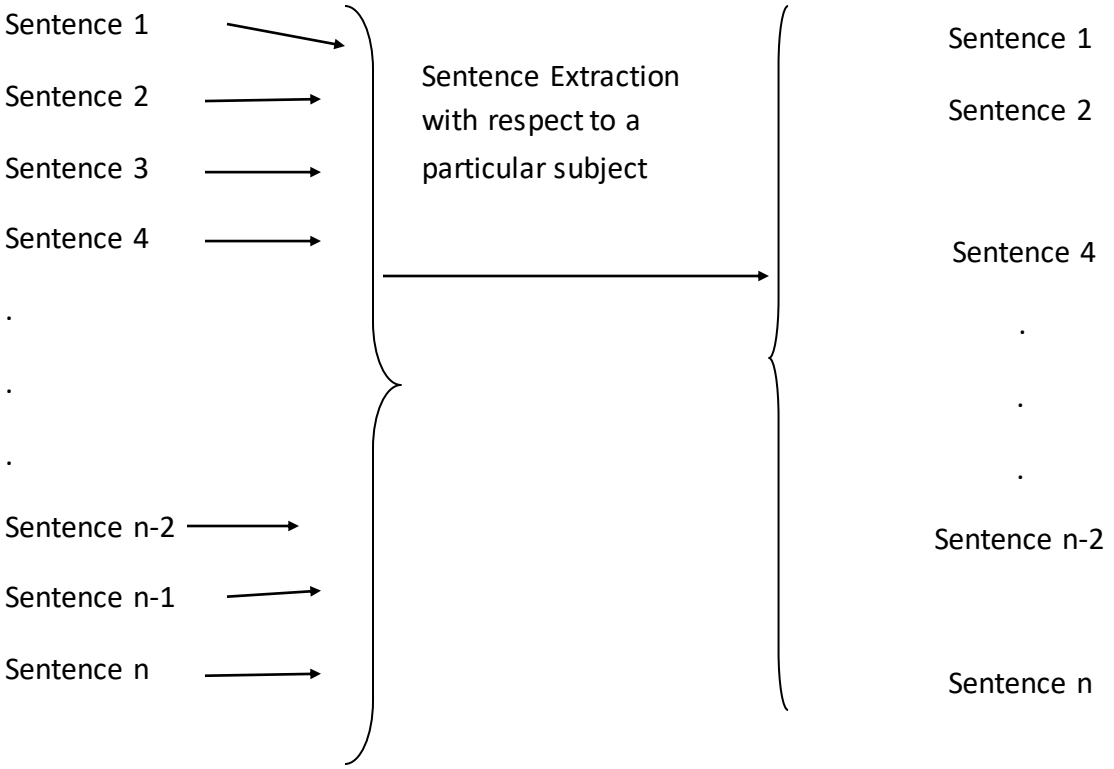


Fig.1 Subjectivity Extraction

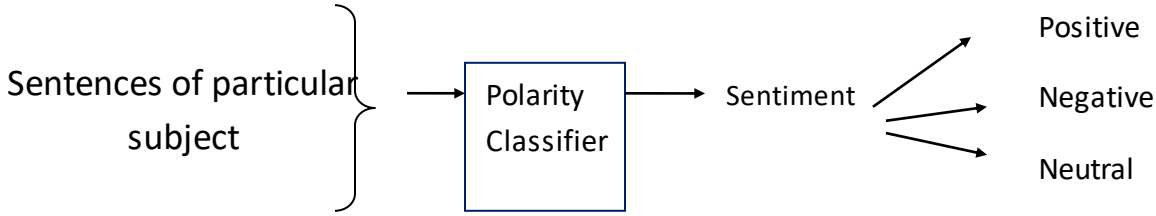


Fig.2 Sentiment Analysis

Subjective Sentences express people's beliefs.

Components needed for identifying sentiments:

- Text containing the attitudes (sentence or entire document)
- Emotional expressions (eg. Positive, Negative, Neutral)

3.3 Tokenization and Part-Of-Speech(POS) Tagging

Tokens are individual words in a text. Tokenization is a process where text is broken into its individual words.
POS Tagger is a piece of software assigns parts of speech to each word in a text. The English taggers use the Penn Treebank tag set. The following figure 3 details the POS abbreviations adopted by the tagger.

1. CC	Coordinating conjunction	25. TO	*to*
2. CD	Cardinal number	26. UH	Interjection
3. DT	Determiner	27. VB	Verb, base form
4. EX	Existential *there*	28. VBD	Verb, past tense
5. FW	Foreign word	29. VBG	Verb, gerund/present participle
6. IN	Preposition/subordinating conjunction	30. VBN	Verb, past participle
7. JJ	Adjective	31. VBP	Verb, non-3rd ps. sing. present
8. JJR	Adjective, comparative	32. VBZ	Verb, 3rd ps. sing. present
9. JJS	Adjective, superlative	33. WDT	*wh*-determiner
10. LS	List item marker	34. WP	*wh*-pronoun
11. MD	Modal	35. WP$	Possessive *wh*-pronoun
12. NN	Noun, singular or mass	36. WRB	*wh*-adverb
13. NNS	Noun, plural	37. #	Pound sign
14. NNP	Proper noun, singular	38. $	Dollar sign
15. NNPS	Proper noun, plural	39. .	Sentence-final punctuation
16. PDT	Predeterminer	40. ,	Comma
17. POS	Possessive ending	41. :	Colon, semi-colon
18. PRP	Personal pronoun	42. (	Left bracket character
19. PP$	Possessive pronoun	43.)	Right bracket character
20. RB	Adverb	44. "	Straight double quote
21. RBR	Adverb, comparative	45. ‘	Left open single quote
22. RBS	Adverb, superlative	46. “	Left open double quote
23. RP	Particle	47. ’	Right close single quote
24. SYM	Symbol (mathematical or scientific)	48. ”	Right close double quote

Fig.3 Words with their POSTaggings

POS tagging is called grammatical tagging or word-category disambiguation. It is the process of marking up a word in a text as corresponding to a particular part of speech, based on its relationship with adjacent and related words in the text.

Following is an example of a parsed text and its Pos Taggings:

Text: I would like to inform you that I performed badly in the Mathematics test
Tokens:{“I”, “would”, “to”, “inform”, “you”, “that”, “I”, ”performed”, ”badly”, ”in”, ”the”, “Mathematics” “test”}
POSTags: [('I', 'PRP'), ('would', 'MD'), ('like', 'VB'), ('to', 'TO'), ('inform', 'VB'), ('you', 'PRP'), ('that', 'IN'), ('I', 'PRP'), ('performed', 'VBD'), ('badly', 'RB'), ('in', 'IN'), ('the', 'DT'), ('Mathematics', 'NNS'), ('test', 'NN')]

3.4 Stop words Removal

Stop words usually refer to the most common words in a language. They would appear to be of little value in text.Using a stop list significantly reduces the number of words that a system has to use for some analysis. Stop words are used to reduce the noise of textual data.

With Stop Words	Without Stop Words
/growing-up-with-hearing-loss/	/growing-hearing-loss/
/coming-to-terms-with-hearing-loss/	/coming-terms-hearing-loss/
/the-world-of-being-hearing-impaired/	/world-being-hearing-impaired/
/echo-in-ears-from-talking-people/	/echo-ears-from-talking-people/
/listening-is-exhausting/	/listening-exhausting/
/living-with-hearing-loss/	/living-hearing-loss/
/what-is-hearing-loss/	/what-hearing-loss/

Fig. 4 Stop words Removal

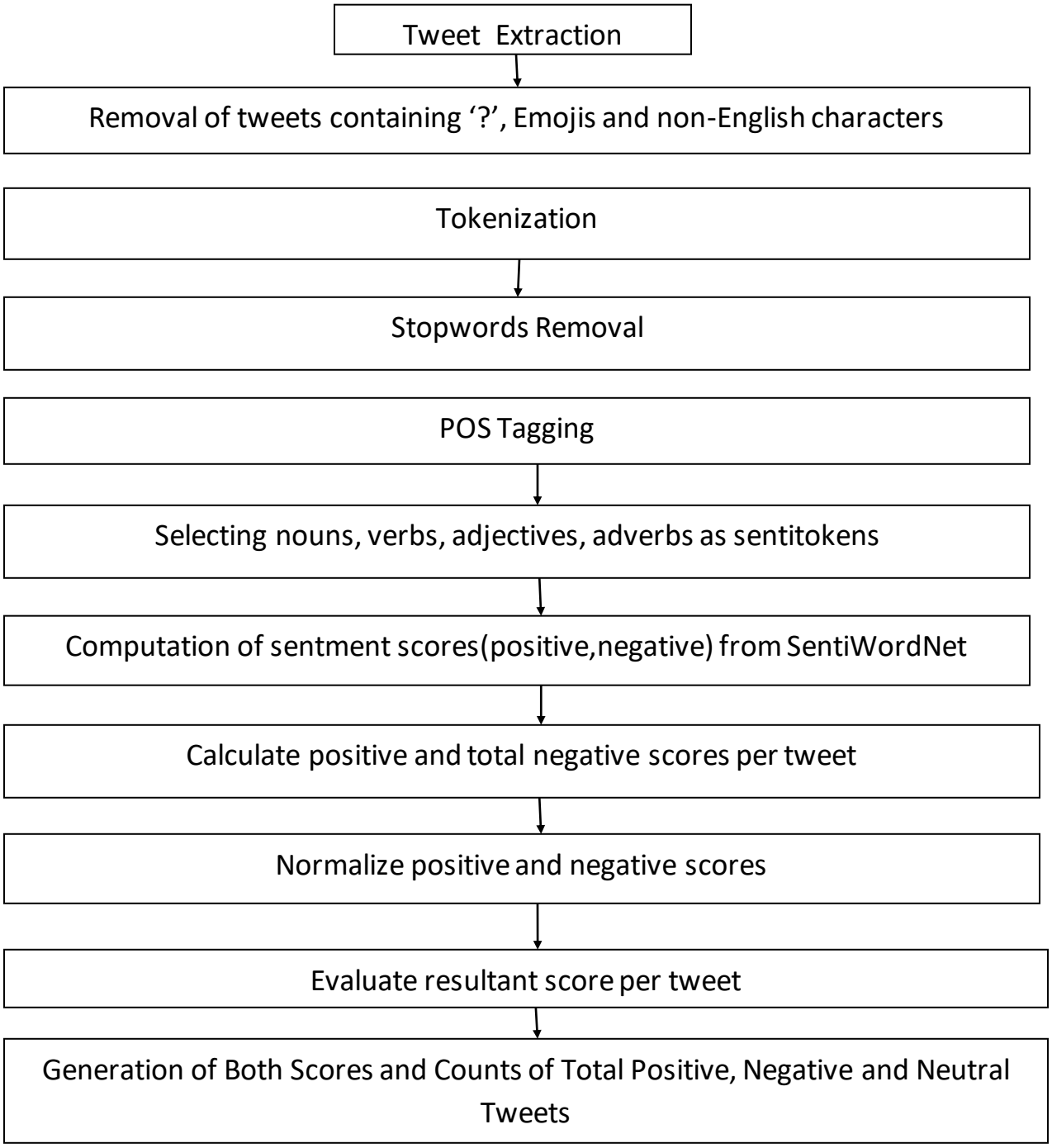
3.5 Correlation Analysis:

J. Han in [8] said that frequent pattern mining is required to find recurring relationships in a given data set. It leads to the discovery of the correlations between items in large transactional or relational data sets. While mining frequent itemsets, problem arises when a huge
number of itemsets are generated satisfying the minimum support threshold. A number of shorter, frequent sub-itemsets may be present in a long itemset.
C. C. Aggarwal in [9] stated that the words are typically correlated with one another in a large corpus of documents. Number of principal componenets is much smaller that the feature space. This necessitates to find word correlations. Document frequency is used to filter out irrelevant features. Very infrequent terms contribute the least to the similarity calculations. A term-document matrix may be viewed as in which the (i, j)th entry is the frequency of the jth term in the ith document. Bursty features can be identified depending its underlying frequency. A pair of documents can have a relation if their cosine similarity is above a user-defined threshold. Considering A = (A_1…A_n) and B = (B_1...B_n)) as the normalized frequency term vector in two different documents A and B, the cosine similarity between the two documents can be defined in Equation(i) below

cosine similarity $$= \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}}$$ -Eq. (i)

4. Methodology

Tweets of a particular event for a specific time range are collected using the package TwitterWebsiteSearch-master. Tweets containing no '?', no emojis and containing only English words are analyzed.



4.1 Tweet Extraction

Tweets on a specific event within a certain time period are collected and those containing emojis, '?' and non-English characters are removed. The tweet text is produced along with Tweet Id and creation time.

Algorithm : Tweet_Extraction

Input: Event name, Time period, Language

Output: Tweet ID, Tweet Created At, Tweet Text

Method:

[1] Download TwitterWebsiteSearch-master package

[2] Extract the package

[3]Give the Event Name, Time Period, Language

[4] Check whether tweet contains emojis,'?' and non-English characters

[5] Remove tweets with '?', emojis and non-English characters.

[6]tweet_id=tweet['id_str']

[7]tweet_createdat=tweet['created_at']

[8]tweet_text=tweet['text']

[9]Record tweet_id,tweet_createdat,tweet_text in TweetDetails.txt

4.2 Stop Word Removal

In this section, Stopwords are removed from Tweet Text.

Algorithm : Stopword_Removal

Input:Tweet Text

Output: Tweet Text without stopwords

Method:

[1] Download NLTK

[2] Import stopwords module from nltk

[3] stop_words = set(stopwords.words('english'))

//Consider English stopwords

[4] Tokenize the Tweet Text into words

[5] Remove stop words

4.3 POS Tagging

All words (except stopwords) in tweet text are assigned with their part of speech

Algorithm: POS Tagging

Input: Tweet Text words other than StopWords

Output: Words with their parts of speech

Method:

[1] Download NLTK

[2]Tag each word of a tweet text with its respective part of speech

4.4 Generation of Sentitokens

Here nouns, verbs, adverbs, adjectives present in a tweet are taken into account for sentiment analysis.

Algorithm: Generation of sentitokens

Input: Words with their pos taggings

Output: Nouns,Verbs, Adverbs, Adjectives

Method:

[1] if postag(word)=='NNP' /'NNS'/'NN'/'NNPS' then word=noun

[2] if postag(word)=='JJ'/'JJR'/'JJS' then word=adjective

[3] if postag(word)=='RB'/'RBR'/RBS' then word= adverb

[4] if postag(word)=='VB'/'VBD'/'VBN'/'VBP'/'VBG'/'VBZ' then word= verb

4.5 Calculation of Sentiment Score

Words of a tweet text are matched with the words with hash in SentiWordNet. If pos tagging of the word according to NLTK and SentiWordNet are matched then the positive and negative scores of each word are taken.

Algorithm : Generation of Sentiment Score

Input: SentiWordNet, Words with their pos taggings

Output: Positive and Negative Scores of Word(s)

Method:

[1] Check if a word in tweet text is present with hashtag in SentiWordNet

[2] If the pos taggings of the word in SentiWordNet and that according to NLTK match or not

[3] If match occurs

take positive and negative scores for further analysis

4.6 Finding average scores per word

A word with its different scores may be present more than one time in SentiWordNet. Here the average of positive and negative scores of each word are taken for analysis.

Algorithm: Average Score Generation per word

Input: Positive, Negative Score of each word

Output: Average of positive and Negative Scores of each word

Method:

[1] Count the Number of times a word occurs

[2] Assign this as Word Count

[3] Sum up all Positive and Negative Scores of the word individually

[4] Divide both the summations by the Word Count separately to produce

Average of Positive and Negative Scores of the word

4.7 Calculation of Positive and Negative Scores per Tweet

The positive and negative average score of all the words are added to find the Total of Positive and Negative Sentiment Scores(Positive and Negative) with respect to a tweet text

Algorithm: Generation of Positive and Negative Scores of a tweet

Input: Average Positive and Negative Scores of all Words in a tweet

Output: Total of the Positive and Negative Scores of all the words in a tweet

Method:

[1] For each word in the tweet

[2] Add Average Positive Score to Total Positive Score

[3] Add Average Negative Score to Total Negative Score

[4] End For

4.8 Normalization of Positive and Negative scores of a Tweet

The positive and negative Sentiment scores of a tweet text are normalized by dividing the total by the number of words in the tweet.

Algorithm : Average Positive and Negative Score per Tweet

Input: Positive and Negative Score of tweet text

Output: Average Positive and Negative Sentiment Score of tweet text

Method:

[1] Count total number of words in tweet and assign in word_countintweet

[2] Divide both Positive and Negative Score of tweet by word_countintweet

[3] Assign the results as Average Positive and Negative Sentiment Score

4.9 Evaluation of Resultant Score per tweet:

The difference between the average of positive and negative sentiments are calculated

Algorithm: Resultant Score per tweet

Input: Average Positive Sentiment and Average Negative Sentiment of tweet

Output: Final _sentimentvalue of tweet

Method:

[1] Sum up Average Positive Sentiment Score and Average Negative Sentiment Score to generate Final _sentimentvalue

4.10 Indication of sentiment of tweet

Finding out whether tweet text contain positive / negative/ neutral sentiment

Algorithm: Ascertain the sentiment polarity of tweet

Input: Final _sentimentvalue

Output: Sentiment Polarity (Positive/Negative/Neutral)

Method:

[1]if Final _sentimentvalue>0 then Sentiment Polarity = Positive

[2] else

[3] if Final _sentimentvalue<0 then Sentiment Polarity = Negative

[4] else

[5] Sentiment Polarity = Neutral

4.11 Generation of Both Scores and Count of Total Positive, Negative and Neutral Tweets

Positive/Negative/Neutral Sentiment Scores of all tweets are taken for analysis. Positive / Negative/Neutral Tweet Counts are evaluated.

Algorithm: Generation of Scores and Count

Input: tweet_createdat,Sentiment of Tweets, Sentiment Scores

Output: Total number of Positive/Negative/Neutral Tweets, Scores of

Positive/Negative and Neutral Tweets per month

Method:

[1] initialization:

positive_Month:=0

negative_Month:=0

neutral_Month:=0

positivescore_Month:=0

negativescore_Month:=0

neutralscore_Month:=0

[2]for year=2016 to 2019

for each Month

if Sentiment_polarity==”Positive” then

positive_Month=positive_Month+1,

positivescore_Month+= sentiment_score

if Sentiment_polarity==”Negative” then

negative_Month=negative_Month+1,

negativescore_Month+= sentiment_score

if Sentiment Score==”Neutral” then

neutral_Month=neutral_Month+1,

neutralscore_Month+=sentiment_score

[3] finalpositivescore_Month= positivescore_Month/ positive_Month,

[4] finalpositive_countMonth=positive_Month

[5] finalnegativescore_Month=negativescore_Month/negative_Month,

[6] finalnegative_countMonth=negative_Month

[7] finalneutralscore_Month=neutralscore_Month/neutral_Month,

[8] finalneutral_countMonth=neutral_Month

4.12 Correlation among events:

Correlation between two events is found by using the cosine similarity of the Proper Nouns in the tweet texts of the events. The cosine similarity is measured based on Eq. (i) provided in the previous chapter. Whenever the cosine similarity between two events exceeded or equalized the value 0.5, the results under each event were taken together for further assessment in order to perform calculation of total positive, negative, neutral sentiment scores and positive, negative, neutral tweet counts of the events. Here all three combinations passed the test of correlation and accordingly the results were combined as depicted at the end of chapter 6.

5. Experimental Configuration

5.1 Dataset

Sl. No.	Tweets on Event	Time Range
1	Demonetization	8th November 2016 to 4th May 2019
2	GST	July 2017 to 4th May 2019
3	Statue of Unity	November 2018 to 4th May 2019

5.2 TwitterWebsiteSearch-master

Twitter website search master is a package to extract tweets older than 7 days from Twittercom.search without using Twitter API.

Here language of tweet to be extracted and the event whose tweets are extracted are placed in particular commands under this package and the tweets get automatically downloaded for further offline uses.

Tweet attributes such as id_str, created_at and text of a tweet are recorded using TwitterWebsiteSearch-master.

5.3 Natural Language Toolkit(NLTK):

NLTK consists of NLP libraries and programs through which various works can be done like tokenisation, postagging, stop word removal and many others.

A sentence or data can be split into words using the method nltk.word_tokenize(sentence).

Words in a sentence can assigned with their parts of speech using nltk.pos_tag(word).

NLTK has a module of stop words to detect stop words from sentences. This can be done using the command

stop_words = set(stopwords.words('english'))

after importing stopwords module of nltk

Sentence	NLTK Tokenization	NLTK POS Tag	NLTK Stopwords
I love to dance	['I', 'love', 'to', 'dance']	[('I', 'PRP'), ('love', 'VBP'), ('to', 'TO'), ('dance', 'VB')]	to

5.4 SentiWordNet

SentiWordNet consists of a positive and negative scores against a number of words.

The words whose positive and negative scores are mentioned in a line of SentiWordNet is denoted preceding by a hash.

Positive and Negative Scores are denoted as Pos Score and Neg Score

A word is present for a number of times with it's positive and negative scores in SentiWordNet. This scores vary depending on the situation when the word is used.

SentiWordNet also assigns scores to words depending on their parts of speech.

Word	POS Score	Neg Score
Able	0.125	0
Haze	0.125	0.25

5.5 CSV Module

CSV Module implements classes to read and write tabular data in CSV format.

Comma Separated Value file is a delimited text file that uses comma to separate and enter values in different cells of Excel sheet. Each line of the file is a data record.

5.6 Hardware Configurations

i. 2GB RAM
ii. A computer
iii. Web Browser

5.7 Software Requirements

i. Python of version greater than 2.6
ii. Integrated Development Environments like IDLE of Python, where Python Programming is done

6. Results and Performance

6.1 Sentiment Score Graph and Tweet Count Graph

Sentiment Score graphs are drawn based on the value obtained by dividing sum of all the positive/ negative/ neutral sentiment scores by the total number of positive/negative / neutral sentiments per month, on Event Topic basis. There are three topics, and their related graphs are depicted in Figures … below.

Tweet Count graphs are obtained from the collected data by finding out the number of positive/negative/neutral tweets per month, producing three such graphs for the three chosen topics as shown in following Figures … .

(i) Demonetization

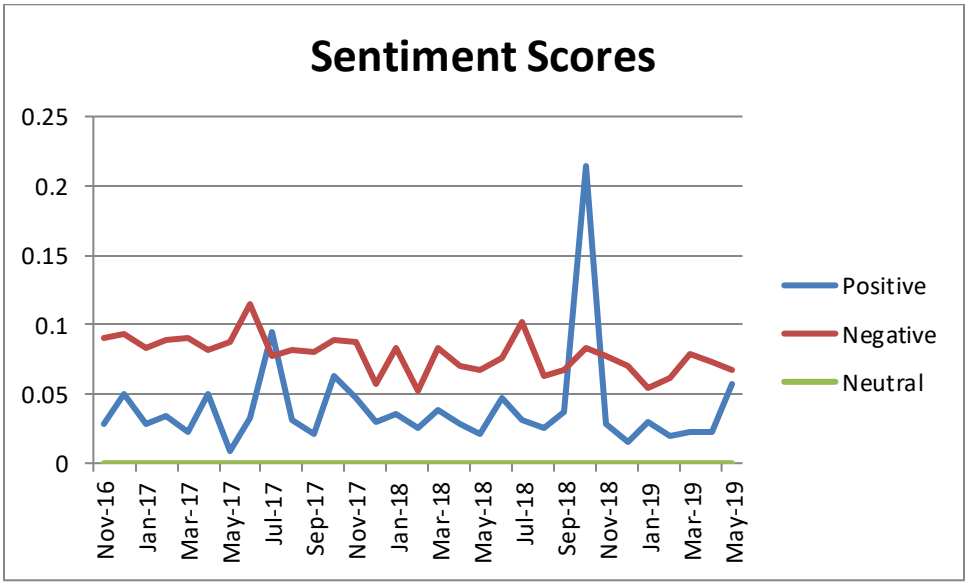


Fig.5: Sentiment Scores on Demonetization

Fig 5 apparently shows that the negative sentiment score is more than the positive sentiment score almost throughout the time period of 8th November 2016 to 4th May 2019. Interestingly, there is a sudden hike of positive score during Oct'18, which possibly indicates changes in Economic or Political policies during that time.

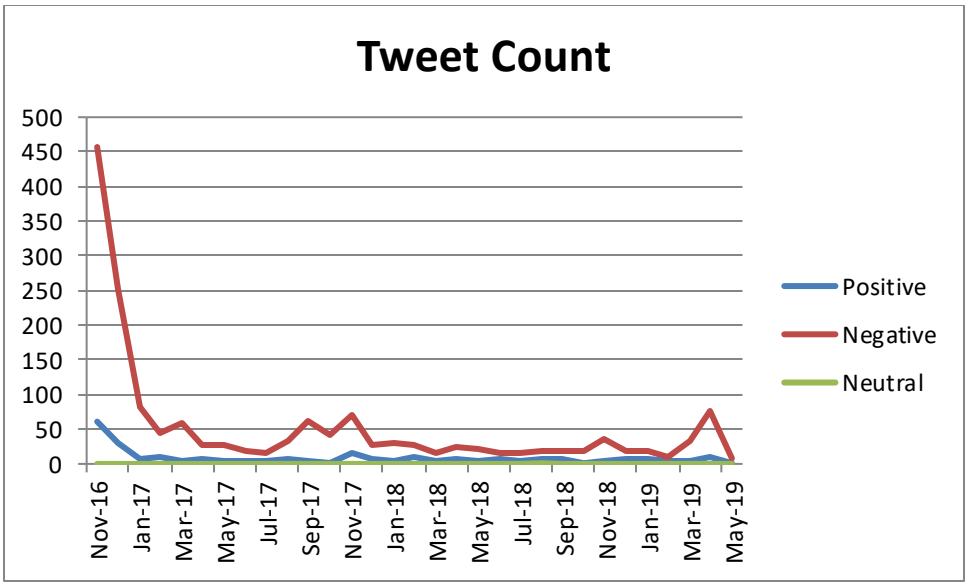


Fig 6: Tweet Count on Demonetization

The Tweet Count graph, on the other hand, shows an initial high spike of negative sentiment when Demonetization started. Very understandably, the repercussion of the incident was felt most strongly at the beginning. The graph also reflects that Negative count exceed positive count throughout the whole time period.

(ii) GST

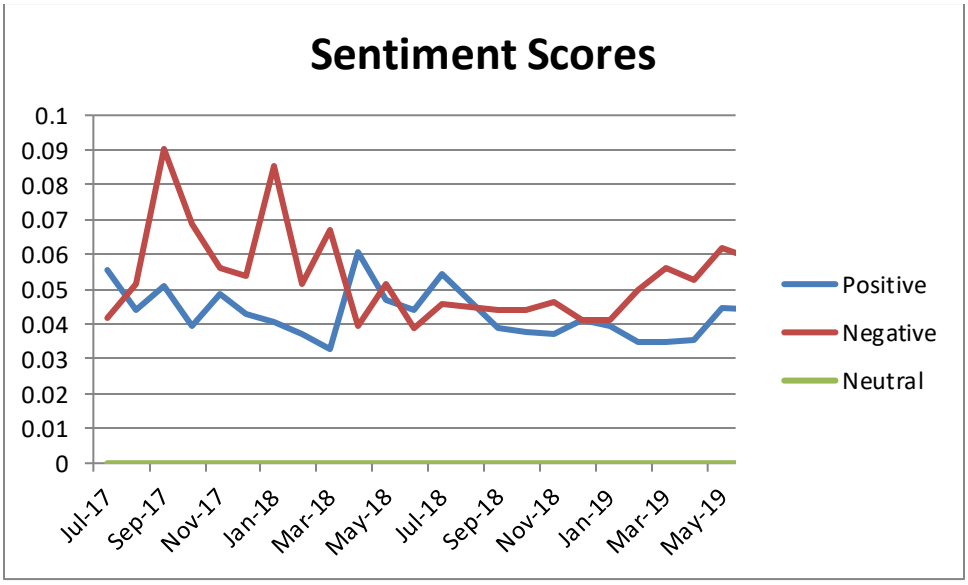


Fig. 7: Sentiment Scores on GST

Sentiment Scores of GST also reflect an overall Negative dominance throughout. Some minor exceptions are noted during April'18 and July'18.

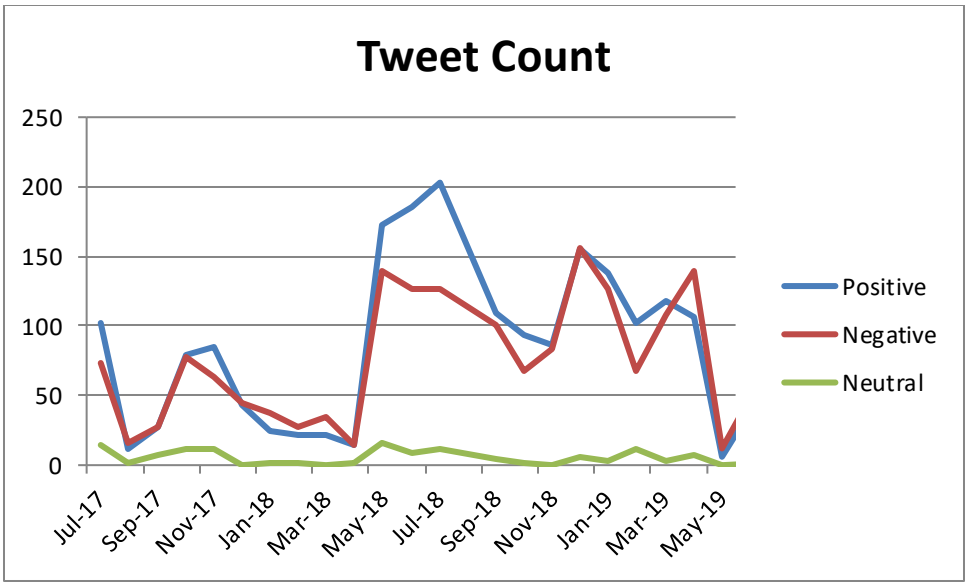


Fig. 8 : Tweet Count on GST

The Tweet Counts of GST show an overall marginal supremacy of Positive sentiment , with slight reversals between Dec 17 and June 18. But there is a sudden positive hike in July 18, followed by slight ups and downs with a negative hike in the pre-election period! The neutral tweets are least in number and may well be ignored in context.

(iii) Statue of Unity

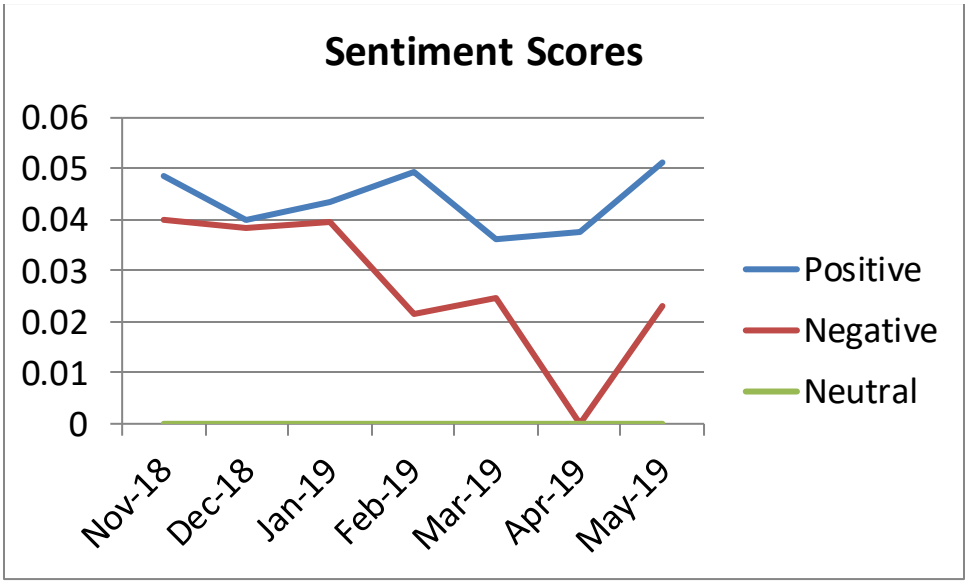


Fig.9: Sentiment Scores on Statue of Unity

The Sentiment scores and the tweet count graph both show a definite positive trend for the building of this statue. The result is justified in a country like ours, where hero-worship is a mandate, and the netizens do not contain the poorer sects in general. So, the economic back lash of this activity does not hurt the population within the Twitterati!

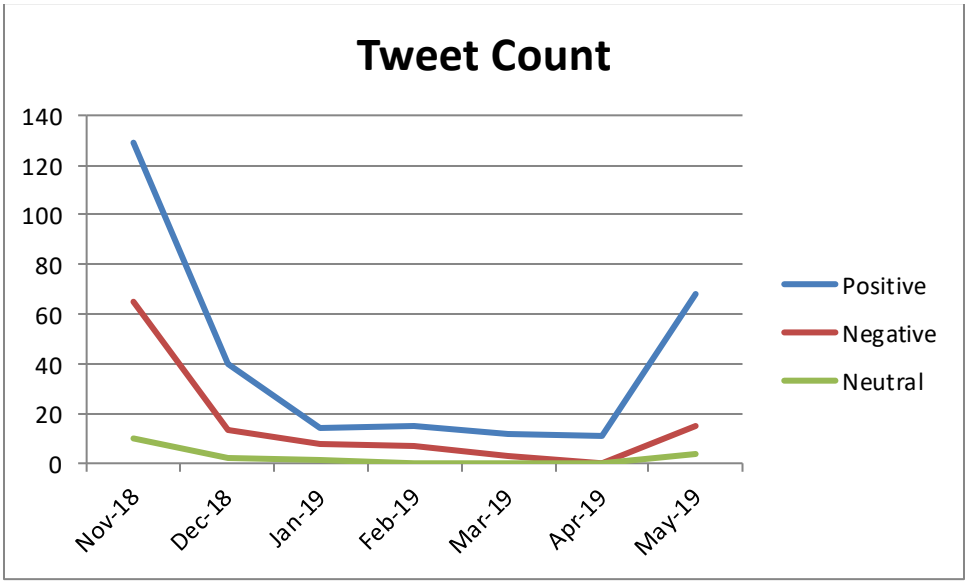


Fig.10: Tweet Count on Statue of Unity

6.2 Cosine Similarities:

Sl. No.	Documents	Cosine Similarity (rounded to two decimal places)
1	GST and Demonetization	0.50
2	Demonetization and Statue of Unity	0.77
3	Statue of Unity and GST	0.69

All the cosine similarity scores(calculated using Eq.(i)) are greater and equal to than 0.5. So the tweet scores and counts of all the events can be combined to give an overall statistics as depicted in the following figure .

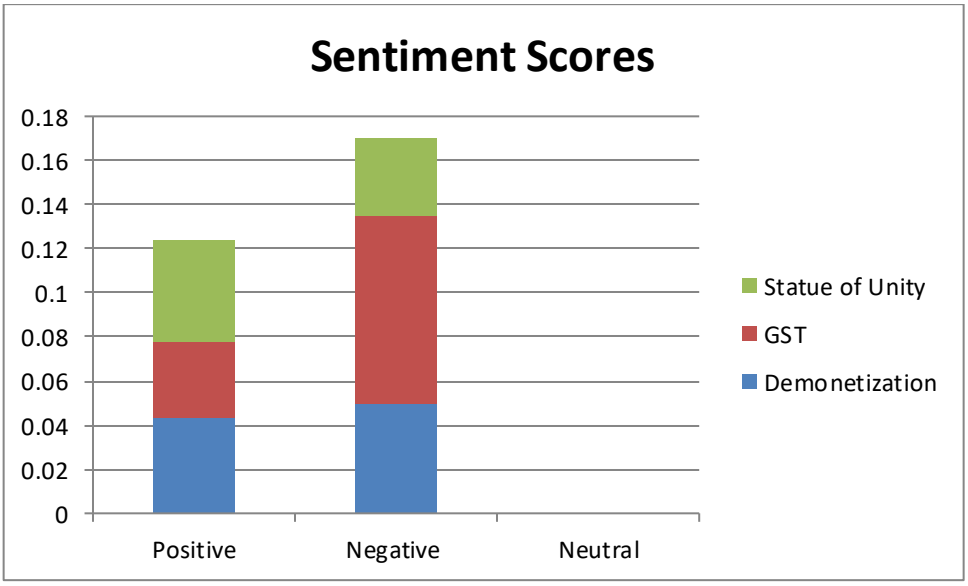


Fig.10: Sentiment Scores of all Events

The combined sentiment score reflects a majority towards negative polarity, with the contribution being most from GST, as it must have been the most burning pain in the neck! The tweet counts in the next figure (reflected in logarithmic terms to fit to scale) also depicts a more vociferous protests against GST, while for Demonetization the protests almost equal the praise. The Statue earns more positive response, again quite understandably.

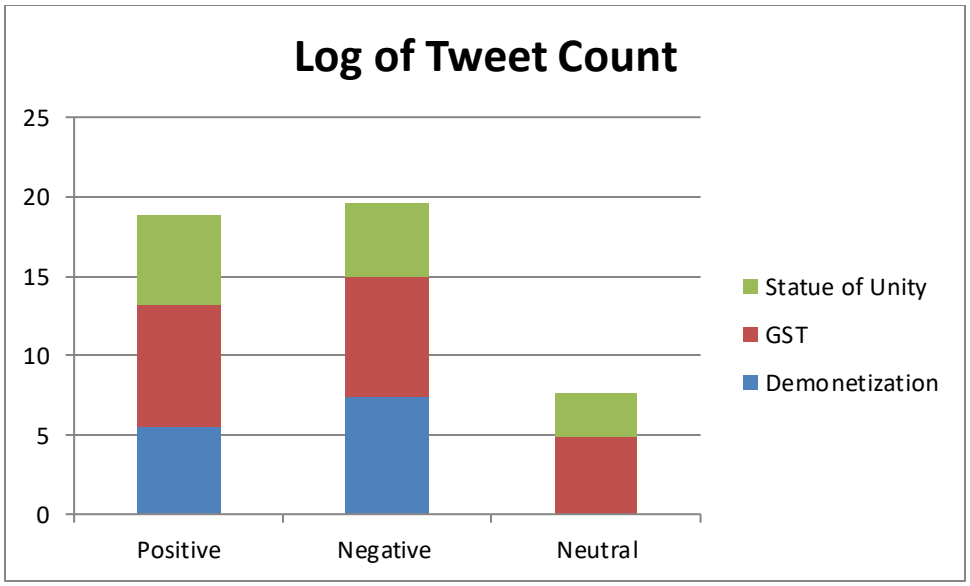


Fig 11. Log of Tweet Count of all the Events

7. Conclusion and Future Scope

In the present work we have considered socio-economic events such as Demonetization, GST and Building of Statues, where common people faced the consequence of actions adopted by the Government. The public view on such subjects were mainly accumulated through tweets collected over an extensive period. The emotional output gathered therein were generated as sentiment scores- positive, negative and neutral.

The actual counts of the positive, negative and neutral tweets also generate a vivid picture of how society rises unanimously in support or protest in a crescendo like sweep at the initial stage. The public interest then decays down mostly over a period, however strong may be the issue.

The most interesting point to be noted here is that these visual outputs can project a verdict of the population, on matters more selective such as election results, policy decisions and trend settings. They can actually represent the true picture if data is collected en masse, like in the recent exit polls.

Better techniques can be envisaged, like capturing more fine-grained emotions, than just positive, negative and neutral. With the recent growth of IoT, the mode of captivating the response may be tuned up to amazing heights. The future endeavors in this domain may expect to incorporate even prediction capabilities by storing past experiences systematically.

8. Reference

[1] Christiana Loana Muntean, Gabriela Andreea Morar, Darie Moldovan: Exploring the Meaning behind Twitter Hashtags through Clustering. Business Information Systems Workshop

[2] Hidenao Abe: Extracting User Behavior-related Words and Phrases using Temporal Patterns of Sequential Patterns Evaluation Indices. Vietnam J Comput Sci, 2017

[3] Zhao Jianqiang, Gui Xiaolin: Comparison Research on Text Pre-Processing Methods on Twitter Sentiment Analysis. Supported by: NSFC under Grant 1472316(in part), Shaanxi Science and Technology Plan Project under Grants 2016ZDJC-05 and 2013ZS16 -Z01/P01/K01 (in part) and Fundamental Research Funds for Ministry of Education of China under Grant XKJC2014008, February,2017.

[4] Pang B. and Lee L. Opinion Mining and Sentiment Analysis. Journal Foundation and Trends in Information Retrieval. 2008; 2(1-2): 1 – 135

[5] A Pappu Rajan and S.P.Victor, " Web Sentiment Analysis for Scoring Positive or Negative Words using Tweeter Data", International Journal of Computer Applications (0975 – 8887) Volume 96– No.6, June 2014

[6] Shailendra Kumar Singh and Sanchita Paul , "Sentiment Analysis of Social Issues and Sentiment Score Calculation of Negative Prefixes", International Journal of Applied Engineering Research · June 2015

[7] Anusha K S , Radhika A D, "A Survey on Analysis of Twitter Opinion Mining Using Sentiment Analysis", International Research Journal of Engineering and Technology (IRJET)

[8] J. Han, M. Kamber, J. Pei, "Data Mining: Concepts and Techniques", 3rd Edition, Morgan Kaufmann Publishers, An imprint of Elsevier, © 2012

[9] C. C. Aggarwal, C. X.Zhai, "Mining Text Data", Springer, 2012