International Journal of Computer Sciences and Engineering Open Access
Research Paper Volume-5, Issue-6 E-ISSN: 2347-2693

Text and Emotion Analysis of Twitter Data

Neetu Anand [1*], Tapas Kumar [2]

[1]Maharaja Surajmal Institute, GGSIPU, New Delhi, India
[2]Lingayas University, Faridabad,Haryana,India

**Corresponding Author: neetuanand@msi-ggsip.org, Mob: 9811396950*

Available online at: www.ijcseonline.org

Received: 25/May/2017, Revised: 02/Jun/2017, Accepted: 20/Jun/2017, Published: 30/Jun/2017

Abstract-The intensification of Technology has altered the means of people's communication by means of opinions, views, sentiments and emotions regarding particular product, services, and people on social networking sites .Social Networking sites are defined as a network of reaction, interaction and relations. Many Social Networking sites, like facebook, whatsapp, Twitter, LinkedIn, Google+, YouTube, Pinterest, Instagram, and Tumblr are the medium to convey the user emotions in form of comments for particular topic. But day by day as huge amount of data is generated from these sites. It becomes a challenging task to perform such type of analysis on big data. R is used to perform the analysis of tweets data that are having a size in GBs. Sentiment analysis, subjectivity analysis and opinion mining are the various techniques to process the review .This paper presented an approach to analyze and visualize twitter data with R. Mainly four types of attitudes are connected with each text positive, negative, neutral and uninterested. Each tweet is analyzed for detecting the sentiments attached to it.

Keywords- Twitter, Data Analysis, Sentiments, Social Media, Emotion Analysis.

I. INTRODUCTION

It is an era of Information Technology. Social Media is a very powerful medium through which a person can give their views. Various social networking sites gain popularity day by day. Facebook is a social networking site started in year 2004 and has 1.86 billion users. It is primarily used by youth, to upload photos, videos, send messages to friends or to give comments. People or company used Facebook page for advertising and getting views of their customer.

Twitter is a micro-blogging tool which allows its users to share, whatever they wish to share with rest of the world in maximum of 140 characters. [1] It has 319 million users, as of fourth quarter of 2016. Post or message sent on twitter is termed as a Tweet. [2]Twitter helps people in getting news from all over the universe, and gives liberty to its users to tweet anything, right from just an emoticon or any thought coming to their mind, to some sort of advertisements, to which any other people could react or comment. People express their views and comments on events including elections, music concert, sports tournaments, educational tasks, earthquakes, spot fixing, bomb-blast etc. Twitter not only allows sharing of textual data, but also enables us to share photos; videos or links to some websites or blogs.Table-1 shows the comparison between various features of Twitter and Facebook [3].

Table-1: Comparison between Facebook and Twitter

Features	Facebook	Twitter
Cost	Free of cost	Free of cost
Number of users	1.86 billion	319 million
Age group that use it	Mainly under 30	35-55 year old
Maximum message length	1000 characters	140 characters
Protection mechanism	Personal data is only accessible by friends. Other pages are open to all.	Account ->Protected – only followers can view ->Unprotected – All internet users can view

In social networking sites twitter has gain Maximum popularity in recent years as political groups, Businesses and Internet users all want to view and further review the public sentiments. They could also improvise the current services rendered by them, by considering the suggestions and negative feedbacks given by the customers. On the other hand, customers will also be able to judge the type of product by analyzing the various responses of the other users towards it. In section II, we describe the term Emotion Analysis. Section III discusses the Literature survey on Analysis of Twitter data. Section IV deals with the Data Analysis Task. In section V, we illustrate the methodology used. Section VI presents the complete experiment process. Section VII illustrates conclusion and future scope and finally references are displayed.

II. EMOTION ANALYSIS

"Emotion" in a layman language is the sentiments or feelings through which one undergoes. [4]An emotion of a person refers to the state of mind he or she has at that point of time. Emotions could be broadly categorized as positive, negative and neutral. Happiness, cheerful, soothing, excitement, etc represents a positive emotion. Grief, anger, anxiety, depression, etc are the signs of negative emotion. There could be a statement or fact which is neither positive nor negative, and such a statement has a neutral emotion. Emotion analysis is measuring the people's opinions, attitudes, views, emotions and classifying them mainly as positive or negative, and sometimes neutral as well. With the advancement of social media and increase in the people connected with it, there is a great need of analyzing the emotions. Emotion analysis of twitter data involves analyzing views, sentiments, thoughts, attitudes, opinions, etc of a person from his or her tweet towards any other individual, product, organization, topic, services, etc. Therefore, Emotion Analysis is becoming a trend nowadays.

III. LITERATURE SURVEY

Various researches have done work in the field of emotion analysis of twitter data. This section discussed some of their work.

Agarwal, B. Xie, I. Vovsha, O. Rambow and R. Passonneau [5] investigated three models, i.e., Unigram model, Tree Kernel model and Feature Based model. They developed a model that grouped the emotions into three different categories, which are positive, negative and neutral. Unlike the before researchers, they considered neutral emotion as well. After the experiment, they drew the conclusion that both the Tree Kernel model and the Feature Based model gave efficient results, when compared with the Unigram model.Po-Wei Liang and Bi-Ru Dai [6] collected sample data that lies in one of the three categories, i.e., camera, movie and mobile, with the help of a Twitter API. The tweets were broadly categorized as opinions and non-opinions. Tweets that had opinions were first filtered and then they were analyzed using Unigram Naive Bayes model with an assumption that Naive Bayes simplified independence. They discarded the unwanted features with the help of Chi Square and Mutual Information feature extraction methods. Therefore, the opinion tweets are classified as positive or negative. Vishal A. Kharde and S.S. Sonawane [7] surveyed and compared the already existing techniques of opinion mining i.e., they compared lexicon based approaches and machine learning algorithms such as Naive Bayes, Support Vector Machine(SVM) and Maximum Entropy techniques of analysis. They concluded that the results of Naive Bayes and SVM models were highly accurate.

IV. DATA ANALYSIS

There are various data analyzing techniques which are used for sentiment analysis, like Naive Bayes Classifier, Support Vector Machine, Lexicon Based Approach, etc [8]. These techniques are briefly described as follows:

a) Naive Bayes Classifier: It is based on counting the frequency of words of various sentiments in the collected log. Accordingly, the tweets are classified. Moreover, nodes weights are adjusted according to its importance and thus the result generated is more accurately represented.

b) Support Vector Machine: It is basically used for categorizing text. It gives the better output comparative to Naive Bayes. Vector of various classes are formed on the hyper plane. Classes can be of negative, positive and neutral tweets.

c) Lexicon Based Approach: In this approach, the dictionary of sentiment words is used to give the score to the opinion words. It basically depends on already compiled sentiment words, phrases and idioms. It is further classified into

- Dictionary based. Eg:- Word net.
- Corpus Based. Eg:- Latent semantic analysis or use of synonyms and antonyms.

V. METHODOLOGY

The complete steps implemented to uncover the emotions of a particular tweet are shown below:

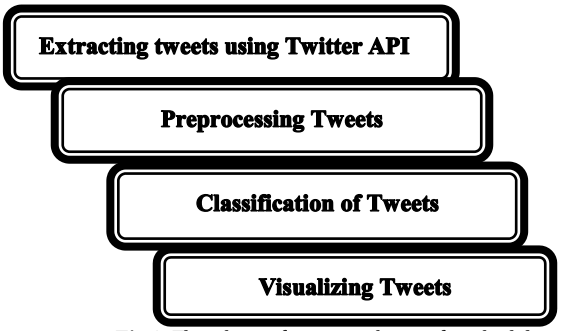

Fig.1. Flowchart of various phases of methodology.

Extracting Tweets Using twitter API: Extracting tweets require creating the twitter API and then storing tweets in a data frame. R studio was used along with the following packages and libraries.

Twitter R-Used for creating Twitter API

RO Auth –for user Authenticate

Word cloud –It creates a cloud depending upon the frequency of words.

RCurl and XML –To download and parsing of web pages

GGplot2-Used to build plots in R

RColourBrewer-Provides colour palettes

GGmap- For plotting on maps

Tm-It's a framework for text mining applications in R

Stringr-Its make R string functions more consistent, simpler and easy to use

***Preprocessing Tweets*:** Preprocessing of tweets includes keeping the essential data and by removing the irrelevant data like URLs, hash (#) tags, @ symbols, stop words, special characters etc.

Classification of Tweets: The preprocessed data is analyzed for categorization of tweets in the form of positive, negative and neutral emotions.

Visualizing Tweets: The Result can be represented in the form of various graphs, charts, map and in form of tables. It depicts the total number of tweets in each of the positive, negative and neutral emotions

VI. EXPERIMENT

The detailed description of each step performed in conducting experiment is as follows;

1) Data Collection

We made a Twitter API [9] so as to collect the tweets [10]. All the tweets related to Budget 2017 were downloaded by providing the keyword, "Budget2017" in R Studio. The downloaded file was saved in the .csv format and had a set of 1500 records. Screenshot of the raw data which was collected is shown in Fig-2.

```
[[21]]
[1] "gstnashik: RT @bcs_llp: #GST legislation enters the final lap | @bcs_llp @gstnashik
 #gstcouncil #Budget2017 \n\nhttps://t.co/vEAXtVGYPm https://t.co/4ro…"

[[22]]
[1] "gstnashik: RT @bcs_llp: Tax slabs: Should #GST Bills be tabled as Money Bills? @bcs
_llp @gstnashik #gstcouncil #Budget2017 \n\nhttps://t.co/7NA8eYI6Gk h…"

[[23]]
[1] "sanket215: RT @bcs_llp: Tax slabs: Should #GST Bills be tabled as Money Bills? @bcs
_llp @gstnashik #gstcouncil #Budget2017 \n\nhttps://t.co/7NA8eYI6Gk h…"

[[24]]
[1] "bcs_llp: Tax slabs: Should #GST Bills be tabled as Money Bills? @bcs_llp @gstnashik
 #gstcouncil #Budget2017… https://t.co/v4XlopdvR1"

[[25]]
[1] "sanket215: RT @bcs_llp: #GST legislation enters the final lap | @bcs_llp @gstnashik
 #gstcouncil #Budget2017 \n\nhttps://t.co/vEAXtVGYPm https://t.co/4ro…"
```

Fig-2 : Data collected(raw) from the Twitter API on "Budget2017".

2) Data Preprocessing

The collected data has many inconsistent and redundant elements that are to be filtered so as to perform emotion analysis techniques on the collected tweets [11][12][13] . A number of tasks performed in data preprocessing are as follows:

- Converting to lower characters :- Data is converted into the lower case so that it would become easy to analyze .
- Removing URLs: - URLs were removed from the tweets for effective analysis.
- Removing hash (#) tags :- Hash(#) tags are removed to make analysis process easy.
- Removing @ symbols: -"@" symbols is filtered from the collected tweets.
- Replace emoticons: - Emoticons like ":)" , ":-)" , etc represents a positive sentiment and hence are replaced with the word "happy" so as to count it as a positive emotion. Also, the emoticons like ":(" , ":-(" , etc shows the negative sentiments and is therefore replaced with the word "sad" so that it could be counted as a negative emotion during analysis.
- Remove all non-English words: -All the linguistic words other than English are removed from the data.
- Remove stop words: - Stop words are removed from the data so as to not overcrowd the essential data.
- Expand acronyms: - Various acronyms used in the tweets were expanded so as to perform emotion analysis task efficiently.
- Remove special characters: - Special characters do not play any vital role in emotion analysis and hence should be removed.

The screenshot of data after preprocessing, i.e., after performing all the above steps is as shown in Fig-3.

```
  [1] "changes in income tax laws from apribudget budget capitalgaintax incometaxtaxes i
nvestmentplan taxes"

  [2] "so narendramodi govt promised in the budget saying transparency and accountabilit
y in political fundings\ncontd"

  [3] "whatever said amp done ground reality economy is in mess many businesses struggli
ng to survive demo budget gstfailin…"

  [4] "whatever said amp done ground reality economy is in mess many businesses struggli
ng to survive demo budget gstfailin…"

  [5] "whatever said amp done ground reality economy is in mess many businesses struggli
ng to survive demo budget gstfailing to revive economy"
```

Fig-3 : Data after Preprocessing

A word cloud of the preprocessed data on "Budget2017" is created using RStudio [14][15]. Word cloud presents the frequency of various words used in the collected tweets in a pictorial format. In the word cloud, the most frequently used word has the highest font size and the rarely used word is assigned the lowest font size. The word cloud formed is as shown in Fig-4.

Fig-4 : Word cloud of Preprocessed data on "Budget2017"

3) Emotion Analysis

The processed data was analyzed and the results showing the number and percentage of positive, negative and neutral tweets have been obtained. After studying the results of emotion analysis, we found that 73% of the tweets were positive, which means that out of 1500 people, 1095 of them were in the favour of 2017's budget. 5% of the total tweets were negative, i.e., there were 75 persons who were against the budget report. We also discovered that there is 22% of the entirety who were neither in complete favour nor against the budget, i.e., 330 people showed a neutral response towards budget released for the year 2017. The result of emotion analysis for Budget 2017 is as shown in the Table-2 below.

Type of Tweets	Number of Tweets	Percentage of Tweets
Total	1500	100%
Positive	1095	73%
Negative	75	5%
Neutral	330	22%

Table-2: Emotion Analysis Results

4) Experimental Result

We have represented the results of emotion analysis in the form of various types of graphs. We represent the analysis results in the form of pie chart and bar graph for easy understanding and interpretation of emotions related to the tweets of Budget 2017.

Pie chart provides a pictorial representation which helps in analyzing the sentiments of tweets more effectively. Fig-5 shows the pie chart that gives the percentage of different emotions attached with the Budget 2017 tweets. [16]Green portion of the pie chart represents the positive tweets, which is 77% of the total collected tweets. The negative tweets, which is 5% is shown by the pink part in the chart. The purple portion of the pie chart depicts the neutral tweets, i.e., 22% of all the collected tweets.

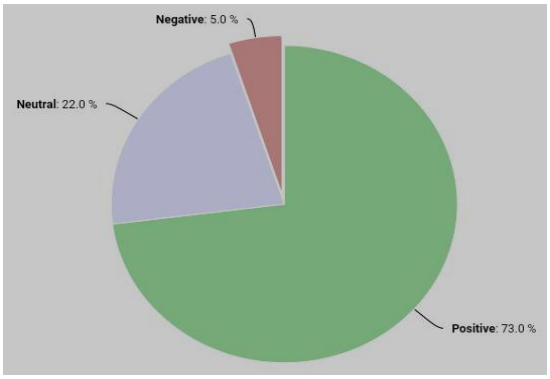

Fig-5 : Pie chart of the Emotion Analysis results.

Bar graph gives a graphical representation of the data which helps us to perform a comparative study of different emotions related to Budget 2017 very easily. [17]The x-axis of the bar graph represents the percentage of the total collected tweets and y-axis shows the three different categories of emotions, i.e., positive, negative and neutral, in which all the tweets had been classified. Fig-6 shows the bar graph that was formed on the basis of analysis results of Budget 2017.

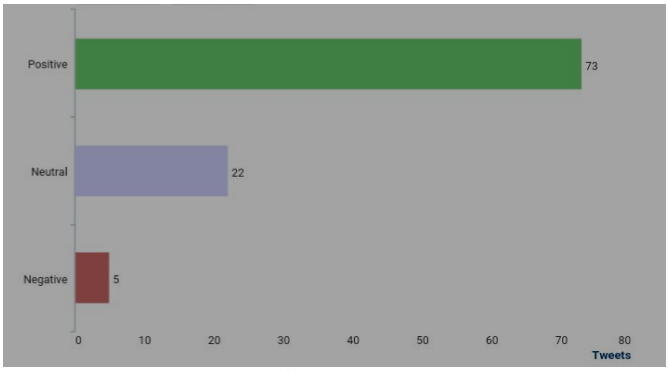

Fig-6 : Bar graph of the Emotion Analysis results.

VII. CONCLUSION AND FUTURE SCOPE

In this paper, we describe and design system for twitter data analysis and visualization using R. We developed a set of analytical representation which helps user to identify about any twitter data and can gain insights from it. We had presented various phases of the experiment that was undertaken in order to analyze the impact of Budget 2017 on inter-linguistic and inter-regional people across India. On the basis of the experiment, we found that there were 77%, 5% and 22% of the total collected tweets which depicted positive, negative and neutral emotions respectively. By interpreting the results of emotion analysis, we conclude that approximately $3/4^{th}$ strength of India is in favour of Budget 2017.In future, we will perform regional level emotion analysis and find out the areas where the people show blissful or aggrieved emotions towards Budget 2017. We will also provide a visual representation of different types of tweets from distinct locations on the map.

ACKNOWLEDGEMENT

I am thankful to my PhD thesis supervisor, Professor Tapas Kumar for guiding me in preparing this paper.

REFERENCES

[1] R. Parikh and M. Movassate, "*Sentiment Analysis of User- Generated Twitter Updates using Various Classification Techniques*", CS224N Final Report, pp. 1-18, 2009

[2] Go, R. Bhayani, L.Huang, "*Twitter Sentiment Classification Using Distant Supervision*"., Stanford University, Technical Paper,2009
[3] A.Pak and P. Paroubek, "*Twitter as a Corpus for Sentiment Analysis and Opinion Mining*", In Proceedings of the Seventh Conference on International Language Resources and Evaluation, pp.1320-1326,2010.
[4] Bifet and E. Frank, "*Sentiment Knowledge Discovery in Twitter Streaming Data*", In Proceedings of the 13th International Conference on Discovery Science, Berlin, Germany: Springer, pp. 1-15,2010.
[5] Agarwal, B. Xie, I. Vovsha, O. Rambow, R. Passonneau, "*Sentiment Analysis of Twitter Data*", In Proceedings of the ACL 2011Workshop on Languages in Social Media, pp. 30-38, 2011.
[6] Po-Wei Liang, Bi-Ru Dai, "*Opinion Mining on Social Media Data*", IEEE 14th International Conference on Mobile Data Management, Milan, Italy, June 3 - 6, 2013, pp 91-96, 2013.
[7] Vishal A. Kharde, S.S. Sonawane, "*Sentiment Analysis of Twitter Data: A Survey of Techniques*". International Journal of Computer Applications, Volume 139 – No.11, pp. 5-15, April 2016.
[8] Jahiruddin, "*Sentiment Analysis of Twitter Data using Statistical Methods*", International Journal of Innovative Research in Engineering & Management (IJIREM), Volume-2, Issue-4, pp. 30-34, July 2015.
[9] Hana Anber, Akram Salah, A. A. Abd El-Aziz3. "*A Literature Review on Twitter Data Analysis*", International Journal of Computer and Electrical Engineering. Volume 8, Number 3, pp. 241-249, June 2016.
[10] I.Hemalatha, G. P Saradhi Varma, Dr. A. Govardhan. "*Preprocessing the Informal Text for efficient Sentiment Analysis*", International Journal of Emerging Trends & Technology in Computer Science (IJETTCS). Volume 1, Issue 2, pp.58-61, July – August 2012.
[11] Adam Crymble, "An Analysis of Twitter and Facebook use by the Archival Community" in Archivaria 70, pp. 125-151, 2010.
[12] Chetashri Bhadane, Hardi Dalal and Heenal Doshi, *"Sentiment Analysis-Measuring Opinions"*, International Conference on Advanced Computing Technologies and Applications (ICACTA), vol. 45, pp. 808–814, 2015.
[13] Xing Fang and Justin Zhan, *"Sentiment Analysis Using Product Review Data"*, Journal of Big Data, pp. 1-14, Springer, 2015.
[14] Xia Hu, Jiliang Tang, Huiji Gao and Huan, Liu, *" Unsupervised Sentiment Analysis with Emotional Signals"*, Proceedings of the 22nd International Conference on World Wide Web, WWW'13, ACM, pp. 607-617,2013.
[15] Santhi Chinthala, Ramesh Mande, Suneetha Manne and Sindhura Vemuri,*" Sentiment Analysis on Twitter Streaming Data"*, Springer International Publishing Switzerland, pp. 161-168, 2015.
[16] Pablo Gamallo, Marcos Garcia, "*Citius: A Naive-Bayes Strategy for Sentiment Analysis on English Tweets"* Proceedings of the 8th International Workshop on Semantic Evaluation, pp: 171–175, Dublin, Ireland, 2014.
[17] Dhanashri Chafale , Amit Pimpalkar, "Review on Developing Corpora for Sentiment Analysis Using Plutchik's Wheel of Emotions with Fuzzy Logic", International Journal of Computer Sciences and Engineering", pp. 14-18, 2014.

Authors Profile

Ms.Neetu Anand pursed Master of Information technology from GJU, Hissar, Haryana. She is currently pursuing Ph.D. and currently working as Assistant Professor in Department of Computer Sciences, Maharaja Surajmal Institute,GGSIP University ,Delhi. She is a member of CSI .She has published more than 15 research papers in reputed International journals and conferences including IEEE,Elsevier and Springer.Her main research work focuses on Web mining, Cloud Security and Privacy, Big Data Analytics and Data Mining. She has 17 years of teaching experience .

Prof. (Dr.) Tapas kumar pursed B.Tech in CSE from Amravati University,Maharashtra ; Master of Computer Science from Guru Jambeshwar University, Hissar, Haryana and PhD (Engineering) on *"An Application of Cellular Automata Paradigm in Image Processing",* BITS, Mesra, Ranchi. He is currently working as Associate Dean & Head - School of Computer Science& Engineering in Lingayas University,Faridabad. He is a member of ISTE,IAENG and CSI. He has published more than 30 research papers in reputed International Journals including scoupus Indexed Journal and Conferences including IEEE, Elsevier, Springer. His main Areas of Interest includes Cellular Automata, Image Processing, Data Sciences and Cloud Computing.He is currently supervising 8 PhD scholars and 2 PhD theses has been submitted. He supervised 14 M.Tech Thesis and guided more than 100 students of B.E in their research based & application based projects. He has 19 years of teaching experience.

