Identifying Tags and Trends by Opinion Analysis of Social Media Data about current Indian Economy: Text Mining Approach using Word Cloud

Roshan R. Karwa, *Computer Science & Engineering, Prof Ram Meghe Institute of Technology & Research, Badnera,* Amravati, India, rrkarwa@mitra.ac.in

Dr. Sunil R. Gupta, *Computer Science & Engineering, Prof Ram Meghe Institute of Technology & Research, Badnera,* Amravati, India, srgupta@mitra.ac.in

Article Info
Volume 82
Page Number: 8863 - 8870
Publication Issue:
January-February 2020

Article History
Article Received: *5 April 2019*
Revised: *18 Jun 2019*
Accepted: *24 October 2019*
Publication: *08 February 2020*

Abstract:
Social media which includes Twitter, Facebook, Whatsapp, Instagram, Linkedin etc. facilitate users to create and share content. In today's era, People expressed their opinions on Social Networking that comprises the use of the internet to unite social media user with their friends, family and Peers. Ease of Access in Internet brings large quantity of users engages in the conversation in form of Comments, Posts or tweets. This paper mainly focused on post which is posted on Social media, mainly Twitter. Analysis of these data is very important and crucial to get aware about recent trends. Analysis of opinions helps in to predict future trend which is helpful in several application such as election, reviews, feedback. If analysis is visual representation instead of number then it becomes easier to understand trend.

Keywords: *Opinion, Social Media Analytics, Word cloud, Indian Economy*

I. Introduction

Social media users express their sentiment on Social media platforms in many ways. Twitter users express their opinion by means of Tweets/Retweets. Instagram users express their stand or choice through Follow or unfollow the user. Facebook users express sentiment through comments, adding/removing friends, liking post. One can comment on particular post to express opinion, one can like particular post to show his opinion whether the user is agree or disagree. There are three issues with respect to it.

a) Analyzing one opinion is easy, hundred opinions are easy but at the same time lakh of opinion is difficult task.

b) On Social Media, visibility of one opinion at one place is easy, ten opinions at one place are easy but lakhs of opinion at one place is very difficult.

c) Extraction of one opinion is easy, ten opinions are easy but lakhs of opinion at once is very difficult job.

These three issues create room for one platform which will help to extract big amount of data at one place.

This paper deals with extraction of opinions related to Indian economy since 2017 until 2019. Sentiment analysis is one field of Natural Language Processing. Sentiment/Opinion Analysis is a method to sense positive and negative opinions in the direction of specific subjects (like organization, product, and reviews on Politics etc.) by collecting/extracting data from media.

It presents influential functionality for all type of analysis [1]. There are several algorithms for sentiment analysis which works as shown in following figure [2].

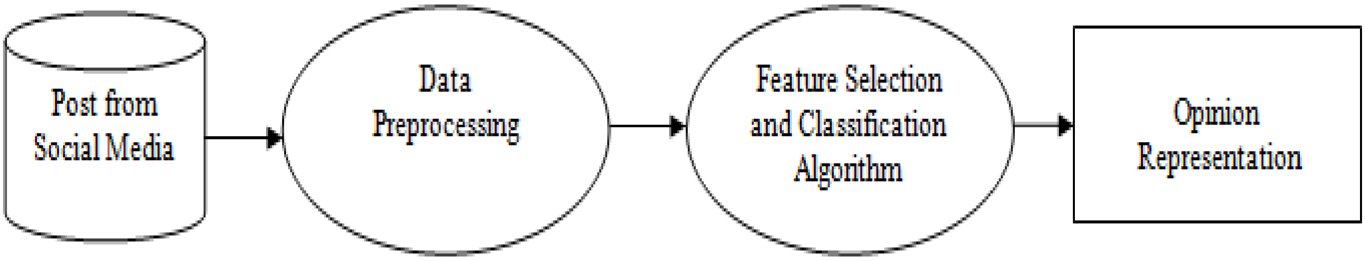


Fig.1. Opinion Representation of Post (Data)

An economy is an area of the manufacturing, supply and trade, as well as utilization of goods and services by different representatives. The economy of India is described as a budding market wealth [3][4]. As per the Wikipedia states with respect to references, Indian Economy is the world's fifth-largest economy by nominal GDP. Also, Indian Economy is the third-largest by (PPP) purchasing power parity. According to the International Monetary Fund, on a per capita income basis, India is positioned on 142nd by nominal GDP. Also, International Monetary Fund states that Indian economy is 119th by GDP (purchasing power parity) per capita in 2018[5]. On social media, people express their opinion in natural languages such as English, Hindi, Marathi, Bengali, and Telgu etc. in large part of India. Analyzing, Processing is what makes this task difficult. This task is Natural Language Processing. In Natural Language Processing, Natural language means language that is used for everyday communication by humans; languages such as English, Hindi, or Portuguese.

Natural language evolved from generation to generation and is difficult to write down with unambiguous rules as compare to artificial languages like mathematical notations and programming language [6].

A word cloud is an innovative visual illustration of text data which is stored in the form of posts. Tags in cloud be typically solo terms, and the significance of every tag be publicized through different typeset dimension or shade [12].

II. Background

Social media Twitter is without a doubt the most sought after micro blogging place these days. It is good place to voice your opinions about certain subject. If these opinions will be analyzed then it can result towards better future for mankind as prediction of sentiment can be done through analysis. Gupta et.al extracted data region wise for analysis of sentiment over the subject of demonetization. [13].

According to sentiments expressed over social media, it has been found that Indian economy had impacted due to many government policies like demonetization, GST, article 370. These government policies have both win-lose side and the load of the lose side is frequently experienced by the common man. P.Singh et.al also analyzed the opinions expressed about demonetization on twitter state wise [14].

Singh P., Sawhney R.S., Kahlon K.S analyzed the posts expressed on social media about the GST implementation by Indian government. They performed the sentiment analysis of the posts expressed on Twitter social media from all states region wise in a stage consisting of Time before GST, Time of GST and Time after GST period using mathematically improvised modeling approach. [15]

Roy, S., Sehgal, S., & Agrawal, S. (2018) described an way to Sentiment Analysis of Twitter Data on the Goods and Services Tax The Goods and Services Tax. GST was pioneered in India as an Indirect Tax on July 1st, 2017. It was introduced with a objective of good motive for the Indian Economy increasing the Gross Domestic Product (GDP) of the nation and reducing prices. [16]

Jyoti Ramteke, Darshan Godhia et.al [20] discussed this opinions affect election and discussed the use of social media for prediction of election results which poses challenges at different stage. Authors tackedl the Scarcity property of Data.

III. Packages Required

We have studied the opinion extraction procedure of some previous researchers and found that some researchers select R language, while some selects Python language.

For the extraction, preprocessing, analysis or other social media related purpose; researchers took help of following packages. Major R packages for Social Media task includes twitteR, bitops, httr, ROAuth, RCurl, RJSONIO, dplyr etc. Supportive packages are stringr, xlsx and others used for processing purpose.

A. twitteR: twitteR provides an interface to Twitter API. This package provides certain functions and parameters which helps researcher to retrieve tweets in particular language such as Hindi, English, and Tamil etc. This is to be done for Data Gathering which is pre requirement for Classification/Analysis etc. [7]

B. bitops: bitops provides functions for bitwise operations. Major datatype for the same is integer vector. It provides the functions like bitAnd, bitFlip, bitShiftL, cksum for performing bit related operations [8]

C. ROAuth: Authentication is required to extract data from another platform. OAuth permits users to authenticate by means of OAuth to the server of users choice and this gateway is provided by ROAuth. [9]. Researchers when extract data from twitter, twitter server was their choice via Twitter API and when they needed data from facebook, facebook server was their choice via Facebook API.

D. RCurl: HTTP requests are needed to establish communication with online platforms. RCurl Package gives various functions with different parameters which allow one to create general HTTP requests as well helps to do operation like get & post forms, fetch URIs etc. Web server gives the output which also process by RCurl. RCurl assists in redirects, cookies & other authentications. [10]

E. RJSONIO: To develop or to deploy applications like extracting tweets requires JavaScript object Notation i.e. JSON format. RJSONIO package helps switch to and fro the information within JavaScript object Notation [11]

With these packages, for visualization of collected data and important keyword from the text, for preprocessing of tweets, we required other packages with above mentioned like word cloud for visual representation [17], tm package for preprocessing data [16], string for replacing string related data [19].

IV. Methodology

Proposed Approach for opinion posted on social media about Indian Economy majorly consist four steps which includes:

A. Data Gathering
B. Data Preprocessing
C. Machine Learning
D. Results

A. Data Gathering

This step is to extract data in form of tweets from Social Media here we are considering Twitter. There are some sub steps:

- Creation of Twitter Account
- Creation of Twitter App Developer

- Obtaining Consumer Secret and Access Token Keys
- Fetching data in form of tweets

B. Data Preprocessing

This step is to preprocessed obtained result from first step. Basically Obtained result is not in pure readable format. So there is need of preprocessing like removing of stop words, getting root word, getting its Part of Speech etc. Also this step includes data cleaning which includes removal of #,@ like symbols, removal of white space, removal of irrelevant words like when, where, the, a etc.
After preprocessing steps, word cloud i.e. visual representation of collected and cleaned data can be created which indirectly gives hint or knowledge about the subject.

C. Machine Learning

After Preprocessing & Cleaning, Next step is to identify whether the opinion generated is Positive, Negative, in favor, against or neutral. This gives room for Machine learning algorithm which majorly divides into Classification and Clustering.

Classification algorithm such as Logistic Regression, Naïve Bayes, Support Vector machine, K-Nearest Neighbor, Decision Tree is supervised algorithm. Supervised algorithm requires annotated data. Clustering algorithms such as Graph based, K-means clustering etc. is unsupervised algorithm. Unsupervised algorithm doesn’t require annotated data. Here for the approach of opinion analysis, we require Classification algorithm which will classify data as Positive, Negative or Neutral.

D. Word cloud & Automatic identification of trends

After extraction of tweets, preprocessing is required then after applying algorithms, data will be analyzed. This is the final step which represents the output of the analysis of the data in form of visual representation in creative form. Following figure shows detailed methodology of opinion analysis.

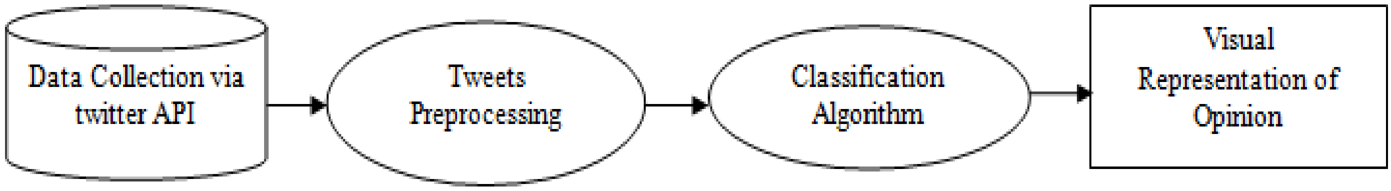

Data Collection via twitter API
Tweets Preprocessing
Classification Algorithm
Visual Representation of Opinion

Fig. 2. Approach of Opinion Analysis about Indian Economy

V. Implementation

a) Packages are having number of Functions and its usability. These are accumulating under directory called “Library” in R environment.

To install Package, the syntax is
install.packages (”package_name”)

For Example:
Install.packages(“twitter”)

The above command gets the particular package from CRAN Website and then install in your R environment. While installing, it gives message to select nearest location. Select India (https) if you are in India or otherwise. For our approach, we are loading following packages after installing it on R Environment.

b) For connectivity between user and server, Twitter API has been used. It comprises four keys Consumer Key, Consumer Secret Key, Access token key and Access token Secret key. Storing keys in variable is important step.

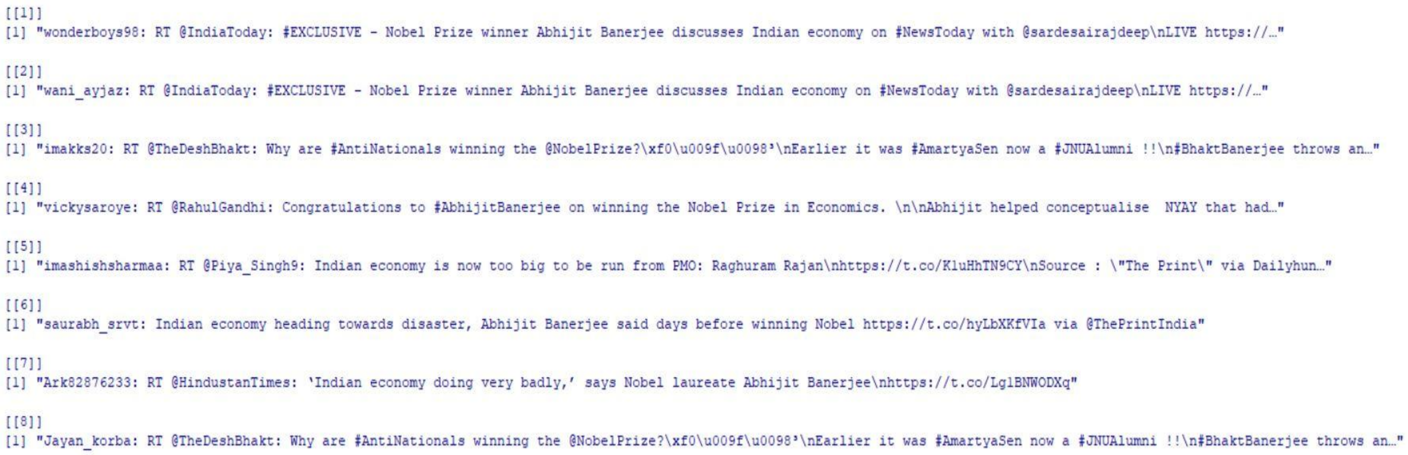
```
ck <-'Consumer Key'
cs<-'Consumer Secret Key'
```

```
at <-‘Access Token key’
as <-'Access Token Secret Key'
```

c) 25000 Tweets have been extracted with the help of twitteR package after authentication of Keys by following command.

```
searchtwitteR(“Indian Economy”, n=25000,
lang=’en’)
```

Here, First parameter is Query for Searching; Second parameter is ‘n’ which represent number of tweets and third parameter is ‘lang’ which stands for language that can be English or Hindi.



```
[[1]]
[1] "wonderboys98: RT @IndiaToday: #EXCLUSIVE - Nobel Prize winner Abhijit Banerjee discusses Indian economy on #NewsToday with @sardesairajdeep\nLIVE https://…"

[[2]]
[1] "wani_ayjaz: RT @IndiaToday: #EXCLUSIVE - Nobel Prize winner Abhijit Banerjee discusses Indian economy on #NewsToday with @sardesairajdeep\nLIVE https://…"

[[3]]
[1] "imakks20: RT @TheDeshBhakt: Why are #AntiNationals winning the @NobelPrize?\xf0\u009f\u0098³\nEarlier it was #AmartyaSen now a #JNUAlumni !!\n#BhaktBanerjee throws an…"

[[4]]
[1] "vickysaroye: RT @RahulGandhi: Congratulations to #AbhijitBanerjee on winning the Nobel Prize in Economics. \n\nAbhijit helped conceptualise  NYAY that had…"

[[5]]
[1] "imashishsharmaa: RT @Piya_Singh9: Indian economy is now too big to be run from PMO: Raghuram Rajan\nhttps://t.co/KluHhTN9CY\nSource : \"The Print\" via Dailyhun…"

[[6]]
[1] "saurabh_srvt: Indian economy heading towards disaster, Abhijit Banerjee said days before winning Nobel https://t.co/hyLbXKfVIa via @ThePrintIndia"

[[7]]
[1] "Ark82876233: RT @HindustanTimes: ‘Indian economy doing very badly,’ says Nobel laureate Abhijit Banerjee\nhttps://t.co/Lg1BNWODXq"

[[8]]
[1] "Jayan_korba: RT @TheDeshBhakt: Why are #AntiNationals winning the @NobelPrize?\xf0\u009f\u0098³\nEarlier it was #AmartyaSen now a #JNUAlumni !!\n#BhaktBanerjee throws an…"
```

Fig. 3. Extraction of tweets about Indian Economy

Here in examples, we got the data like RT which shows the retweet. This retweet can be removed using strip_retweet function of twitter package present in R.

After removing retweet, this data is stored in data variable.

To create word cloud from extracted tweets, we first preprocess data/text/tweet such as

a) Removal of unnecessary spaces present in extracted tweet

b) Removal of URLs present in extracted tweet

c) Removal of stop words present in extracted tweet

d) Removal of punctuations and digits

Four word clouds are created for visual representation; only important keywords are necessary and meaningful.

a) Indian Economy (result type: Recent tweets before 15 October 2019)

b) Indian Economy (result type: Popular tweets before 15 October 2019)

c) Demonetization (result type: Recent tweets before 15 October 2019)

d) Opinion of Abhijeet Banerjee (Nobel winner Economist) (result type: Recent tweets before 15 October 2019)

Significance of each sentiment/word/opinion is expressed in different color/font. Major repeated words in tweet are shown in bigger font size. Table 1 consist tags in clouds

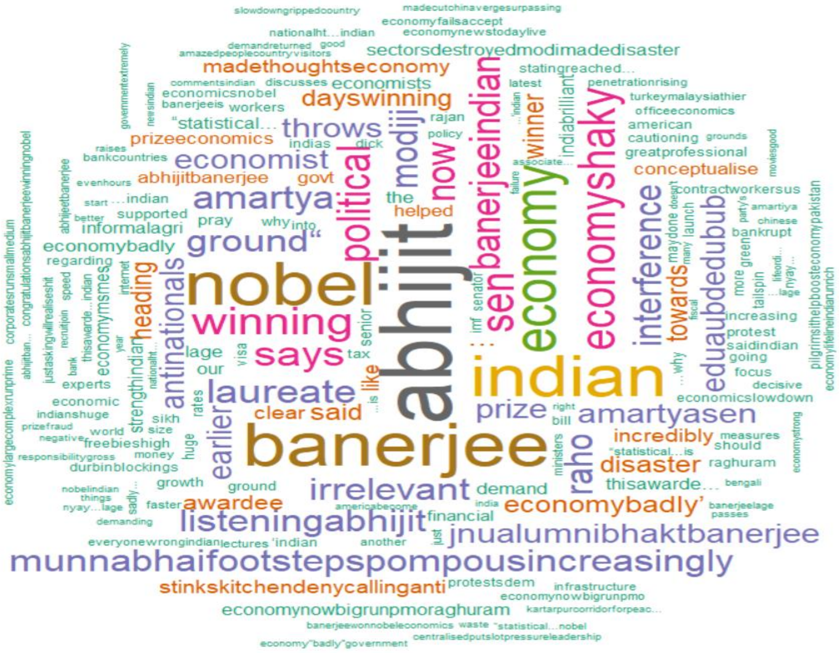
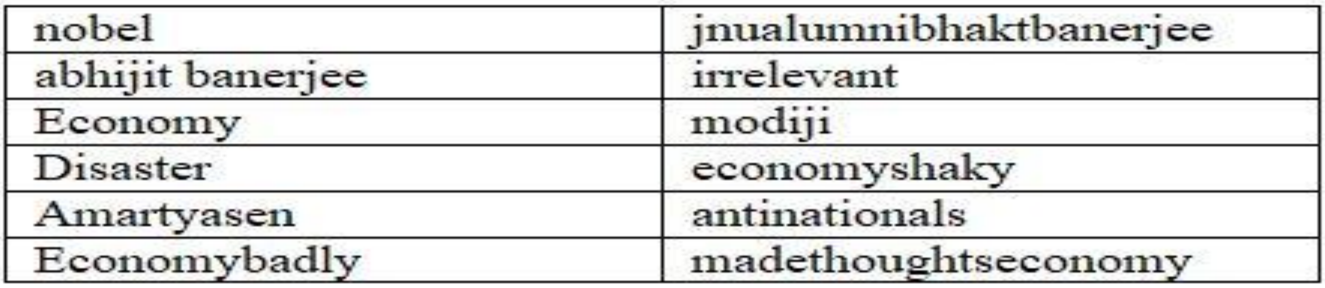


Fig.4. Word cloud of Indian Economy (Until October 15, 2019)

TABLE I. TAGS AND KEYWORDS IN WORD CLOUD OF INDIAN ECONOMY (UNTIL OCTOBER 15, 2019) FIG 4



nobel	jnualumnibhaktbanerjee
abhijit banerjee	irrelevant
Economy	modiji
Disaster	economyshaky
Amartyasen	antinationals
Economybadly	madethoughtseconomy

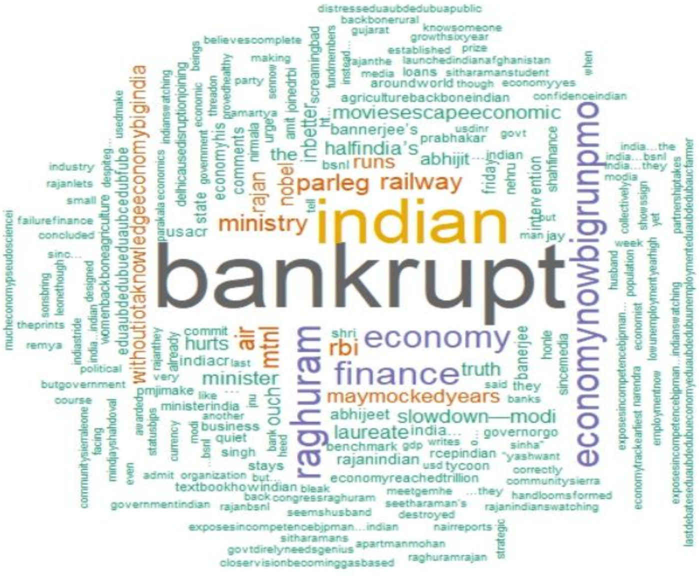


Fig.54. Word cloud of Indian Economy (Until October 15, 2019)

Significance of each tag is shown in different color as well font. Table 2 consists of tags in clouds.

TABLE II. TAGS AND KEYWORDS IN WORD CLOUD OF INDIAN ECONOMY (UNTIL OCTOBER 15, 2019) FIG 5.

indian	economybigrunpmo
Bankrupt	slowdown
Raghuram	bsnl
Moviesescape	finance
Parleg	railway

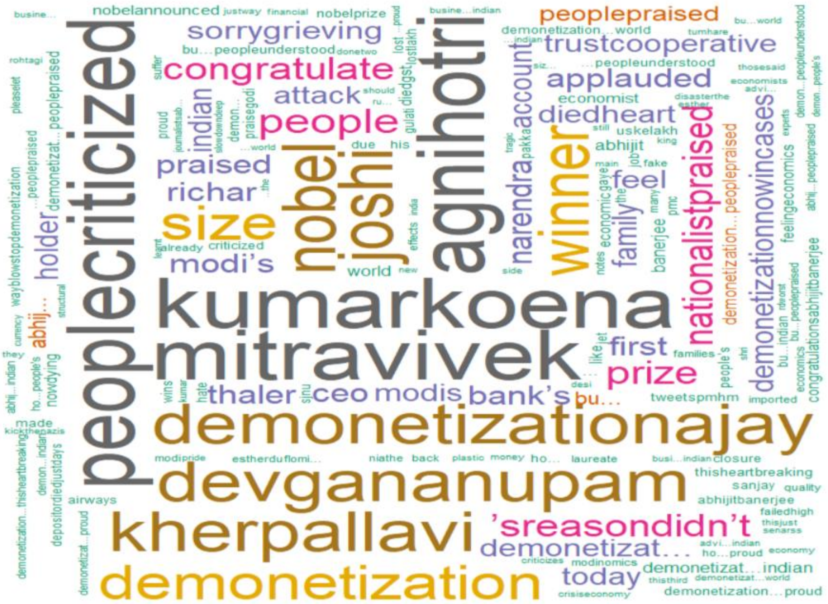


Fig.6. Word cloud of Demonetization (Until October 15, 2019)

VI. CONCLUSION & FUTURE SCOPE

Currently, System is extracting specified 'n' number of tweets by authenticating the keys. Also System is extracting opinion linguistic languages also. While the number of Indians living in urban areas has increased over the last two decades, about 67% of people still live in rural areas and the number of rural population expressing their opinion on social media is less than expected. So its impact with the social media opinions will affect overall result. By creating the word clouds, we found that visual representation is creative way to analyze the data. It doubles the interest of researcher.

VII. REFERENCES

[1] Tetsuya Nasukawa , Jeonghee Yi. (2003) "Sentiment analysis: capturing favorability using natural language processing", Association for Computing Machinery(ACM), Proceedings of the 2nd international conference on Knowledge capture, Sanibel Island, FL, USA. DOI: 10.1145/945645.945658

[2] Medhat, W., Hassán, A., & Korashy, H. (2014). "Sentiment analysis algorithms and applications: A survey" Ain Shams Engineering Journal,5(4),1093-1113,doi: https://doi.org/10.1016/j.asej.2014.04.011

[3] Alamgir, Jalal. "India's Open-Economy Policy: Globalism, Rivalry, Continuity" Routledge. p. 176. ISBN 978-1-135-97056-7.

[4] Kanungo, Rama P.; Rowley, Chris; Banerjee, Anurag N. “Changing the Indian Economy: Renewal, Reform and Revival” Elsevier. p. 24. ISBN 978-0-081-02014-2.

[5] World Economic Outlook Database, April 2019". IMF.org. International Monetary Fund.

[6] Bird,S. Klein,E. & Loper,E. (2009). “Natural Language Processing with Python” United States of America: O’Reilly Media.

[7] Package ‘twitteR’ by Jeff Gentry .

[8] Package ‘bitops’ S original by Steve Dutky <sdutky@terpalum.umd.edu> initial R port and extensions by Martin Maechler; revised and modified by Steve Dutky

[9] Package ‘ROAuth’ by Jeff Gentry .

[10] Package ‘RCurl’ by Duncan Temple Lang and the CRAN team https://cran.r-project.org/web /packages/RCurl/RCurl.pdf

[11] Package ‘RJSONIO’ by Duncan Temple Lang https://cran.r-project.org/web/packages/RJSONIO/RJSONIO.pdf

[12] Martin Halvey and Mark T. Keane, “An Assessment of Tag Presentation Techniques”.

[13] Gupta, F., & Singal, S. (2017). “Sentiment analysis of the demonitization of economy 2016 India”, Regionwise. 2017 7th International Conference on Cloud Computing, Data Science & Engineering - Confluence. doi:10.1109/ confluence. 2017.7943240

[14] P. Singh, R.S. Sawhney, K.S. Kahlon, “Sentiment analysis of demonetization of 500 & 1000 rupee banknotes by indian government” ICT Express (2017), http://dx.doi.org/10.1016/ j.icte.2017.03.001

[15] Singh P., Sawhney R.S., Kahlon K.S. (2019) “Twitter Based Sentiment Analysis of GST Implementation by Indian Government” In: Patnaik S., Yang XS., Tavana M., Popentiu-Vlădicescu F., Qiao F. (eds) Digital Business. Lecture Notes on Data Engineering and Communications Technologies, vol 21. Springer, Cham

[16] Roy, S., Sehgal, S., & Agrawal, S. (2018). “An Approach to Sentiment Analysis of Twitter Data on the Goods and Services Tax” 2018 International Conference on Advances in Computing, Communication Control and Networking (ICACCCN). doi:10.1109/icacccn.2018.8748822

[17] Package ‘tm’ by Ingo Feinerer https://cran.r-project.org/web/packages/tm/vignettes/tm.pdf

[18] Package ‘wordcloud’ by Ian Fellows https://cran.r-project.org/web/packages/wordcloud/wordcloud.pdf

[19] Package ‘stringr’ by Hadley Wickham https://cran.r-project.org/web/packages/stringr/stringr.pdf

[20] Jyoti Ramteke, Darshan Godhia et.al, “Election Result Prediction Using Twitter sentiment Analysis” in 2016 International Conference on Inventive Computation Technologies (ICICT), IEEE