

Sentiment Analysis of #MakeInIndia & #Demonetization using R

CH.R.Vinodkumar¹, Prof.D.Lalitha Bhaskari²

¹Research Scholar, Department of CSSE, AU College of Engineering, Andhra University,

²Professor, Department of CSSE, AU College of Engineering, Andhra University

Abstract: With the advancement of web technology and the availability of digital devices, studies reveal that there is a huge volume of data available in the web for internet users. Internet has become a platform for online learning, exchanging ideas and sharing opinions through Social networking sites like Twitter, Facebook, Google+, LinkedIn etc. thus contributing to huge amount of data every second. The availability of voluminous data has led to an interesting research in the field of sentiment analysis. This paper focuses on the sentiment analysis of Twitter data based on the tweets made for #Make in India, a prestigious initiative by the Government of India in the year 2015and #Demonetization tweets. The main aim of this work is to analyze the opinions of the citizens based on the generated tweets which can help the government authorities with regard to implementation of the program in future and to access the success rate of the implementation of the programs with the help of technology. A brief introduction on Data analysis in presented in section 1 followed by literature review in section 2. The procedure for tweet classification and sentiment analysis is done in section 3. The analysis of the two tweets using R has led to some interesting results which are discussed in the experimental section 4 in this paper followed by conclusions and references.

Date of Submission: 30-10-2017

Date of acceptance: 25-11-2017

I. Introduction

Data Analysis is a process in which data is collected, inspected, transformed and organized so that it can be modelled with the goal of extracting useful information. This extracted data is useful in suggesting modifications and supports decision-making. Data analysis has multiple aspects, approaches and uses. It has diverse techniques which can be applied in different businesses, science and social science domains. The process of data analysis is a process for obtaining raw data and converting it into information useful for decision making by users. Fig 1.explains the different stages in data analysis.

Initially raw data is collected from a variety of sources through interviews, downloads from online sources, reading documents, sensors, cameras, videos, satellites, recording devices, etc. On the available data,

Data preprocessing is to be performed so as to organize the available data into a structured format. After preprocessing the data into the required format, data cleansing must be done so as to further prepare it for data analysis. After data cleaning there are several methods and algorithms available which can be applied on to the data. In this work, R programming packages are used for data analysis and visualization.

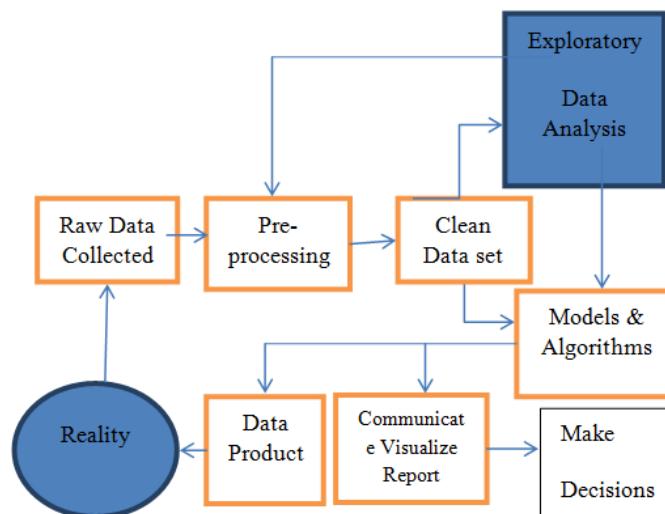


Figure 1: Data Analysis Process

Based on the analysis decisions can be made by the competitive authorities or users. The above process is used to analyze the sentiments of two most important tweets which has moved the entire country in the recent years. The next section deals with the literature review and the work done in the area of sentiment analysis.

II. Overview Of Related Work

This work is inspired by ‘A Survey of Data Mining Techniques for Social Network Analysis’[1] proposed by Mariam Adedoyin – olowie, Mohammed MedhatGabbar. The key learnings from this paper are about the techniques of data mining for social networks and their effective usage.

From the Module Twitter Analysis [2] by KaifyRais, different ways to analyze the tweets from Twitter using R programming language, the techniques to connect to twitter and retrieve tweets, preprocessing of tweets and analysis of the trends are learnt.

From the paper, An Approach to Harvesting, Cleaning, and Analyzing Data from Twitter Using R [3] by Stephen Hill and Rebecca Scott, We have drawn inspirations on preprocessing and analysis of tweets that were tweeted on Twitter.

From the white paper Social Networks Overview: Current Trends and Research Challenges published [4] by European Commission, we learned how to analyze the current trends of the social network analysis and the challenges that were present while doing the analysis of social network analysis.

From the paper Opinion Mining on Social Media Data [5] by Po-Wei Liang, Bi-Ru Dai, we will determine the data’s category first, because we assume that different domains are associated with different customary terms and expressions, which will affect the accuracy of sentiment analysis.

From the paper Twitter as a Corpus for Sentiment Analysis and Opinion Mining [6] by Alexander Pak, Patrick Paroubek, We use a dataset formed of collected messages from Twitter. Twitter contains a very large number of very short messages created by the users of this micro blogging platform. The contents of the messages vary from personal thoughts to public statements. As the services of micro blogging sites are growing rapidly every day and so is the userbase of these platforms, data can be mined from these sources. This data plays an important role in opinion mining and sentimental analysis tasks.

From the paper Opinion Mining on Twitter Data using Unsupervised Learning Technique [7] by MuqtarUnnisa, Syed Raziuddin& Ayesha Ameen, sentiment analysis refers to the application of natural language processing, computational linguistic and text analytics to identify and extract subjective information in source materials.

From the paper Twitter Sentiment Analysis [8] by EfthymiosKouloumpis, Theresa Wilson, Johanna Moore Sentiment analysis is a growing area of Natural Language Processing with research ranging from document level classification to learning the polarity of words and phrases. Given the character limitations on tweets, classifying the sentiment of Twitter messages is most similar to sentence-level sentiment analysis. Based on the above works, the next section deals with the procedure adopted for tweet analysis.

III. Sentiment Analysis

Sentiment analysis is a process of evaluating the use of natural language - either spoken or written - to determine whether the expression is positive, negative or neutral. Sentimental analysis can be applied to evaluate text, computational linguistics and biometrics. It is used to extract, quantify and study affective states and subjective information. Sentiment analysis is widely applied to voice of the customer materials such as reviews and survey responses, online and social media, and healthcare materials for applications that range from marketing to customer service to clinical medicine.

Generally speaking, Sentiment analysis helps to determine what the customers like and dislike, attitude of a speaker, writer, or other subject with respect to some topic or the overall contextual polarity or emotional reaction to a document, interaction, or event. The attitude may be a judgment or evaluation or the intended emotional communication by a large group of people. Next section deals with the implementation process of #MakeInIndia and #Demonetization tweets.

3.1 Sentiment analysis on #MakeInIndia and#Demonetization tweets:

Twitter is one of the largest social networking platforms and has a wide range of users who tweet regularly on various issues, concepts, performance of or initiatives taken by an organization or Government etc. A tweet is a message sent on Twitter which are often very small in terms of the length but convey the total meaning of a context they are tweeted for. Usually people use hash-tags(# to tweet on a particular context or issue. Tweets are very easy to create and publish. They have a wide reach as they can be accessed through computers as well as smartphones through apps.

TwitterR offers an easy way to extract tweets containing a given hash-tag, word or term from a user’s account or public tweets which is shown in Fig 2. However, before loading TwitterR library and using its

functions, developers should create an app on dev.twitter.com and then run their code. The developers can use programming languages like R, Ruby, python to code their apps.

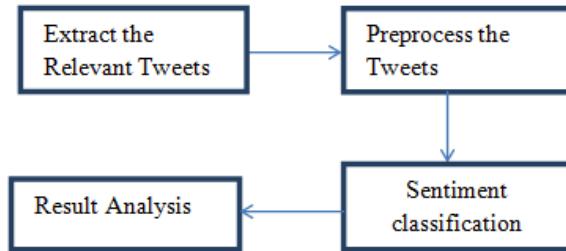


Figure 2: Process of Tweet Analysis

3.1.1 Extract Relevant tweets

In real world, the tweets containing opinion is more valuable. So in this part we want to filter out the tweets without opinion. To extract the tweets containing opinions we use “twitteR” package. It will explain how the system can classify the tweets into opinion class and non-opinion class. Then the system passes the opinion part into the next step.

To demonstrate sentiment analysis, we analyzed tweets relating to #MakeInIndia and #Demonetization. In order to extract specific tweets relating to initiations, developers should query twitter for tweets with the hashtag #MakeInIndia and #Demonetization.

The extracted tweets are saved in .csv file which contains all the information about the tweets. A snapshot of the csv file is given below:

makeinindia.csv - Microsoft Excel																			
A15	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	text																		
2	WhatevIndia has been	FALSE	47	NA	NA	NA	732617 NA	< a href="https://t.co/	replyToId	statusSource	screenName	retweeted	retweetCount	longitude	latitude				
3	India is the agenda	FALSE	59	NA	NA	NA	732617 NA	< a href="https://t.co/	replyToId	statusSource	screenName	retweeted	retweetCount	longitude	latitude				
4	India is on the path	FALSE	24	NA	NA	NA	732617 NA	< a href="https://t.co/	replyToId	statusSource	screenName	retweeted	retweetCount	longitude	latitude				
5	India is on the path	FALSE	45	NA	NA	NA	732617 NA	< a href="https://t.co/	replyToId	statusSource	screenName	retweeted	retweetCount	longitude	latitude				
6	RT @WhatevIndia: makein	FALSE	6	NA	NA	NA	732617 NA	< a href="https://t.co/	replyToId	statusSource	screenName	retweeted	retweetCount	longitude	latitude				
7	RT @WhatevIndia: The	FALSE	58	NA	NA	NA	732617 NA	< a href="https://t.co/	replyToId	statusSource	screenName	retweeted	retweetCount	longitude	latitude				
8	RT @WhatevIndia: Or	FALSE	6	NA	NA	NA	732617 NA	< a href="https://t.co/	replyToId	statusSource	screenName	retweeted	retweetCount	longitude	latitude				
9	RT @WhatevIndia: Most	FALSE	0	NA	NA	NA	732617 NA	< a href="https://t.co/	replyToId	statusSource	screenName	retweeted	retweetCount	longitude	latitude				
10	RT @WhatevIndia: We	FALSE	0	NA	NA	NA	732617 NA	< a href="https://t.co/	replyToId	statusSource	screenName	retweeted	retweetCount	longitude	latitude				
11	RT @WhatevIndia: 1	FALSE	0	NA	NA	NA	732617 NA	< a href="https://t.co/	replyToId	statusSource	screenName	retweeted	retweetCount	longitude	latitude				
12	RT @WhatevIndia: 2	FALSE	0	NA	NA	NA	732617 NA	< a href="https://t.co/	replyToId	statusSource	screenName	retweeted	retweetCount	longitude	latitude				
13	RT @WhatevIndia: 3	FALSE	0	NA	NA	NA	732617 NA	< a href="https://t.co/	replyToId	statusSource	screenName	retweeted	retweetCount	longitude	latitude				
14	RT @WhatevIndia: 5	FALSE	0	NA	NA	NA	732617 NA	< a href="https://t.co/	replyToId	statusSource	screenName	retweeted	retweetCount	longitude	latitude				
15	RT @WhatevIndia: 6	FALSE	0	NA	NA	NA	732617 NA	< a href="https://t.co/	replyToId	statusSource	screenName	retweeted	retweetCount	longitude	latitude				
16	RT @WhatevIndia: 7	FALSE	0	NA	NA	NA	732617 NA	< a href="https://t.co/	replyToId	statusSource	screenName	retweeted	retweetCount	longitude	latitude				
17	RT @WhatevIndia: 8	FALSE	0	NA	NA	NA	732617 NA	< a href="https://t.co/	replyToId	statusSource	screenName	retweeted	retweetCount	longitude	latitude				
18	RT @WhatevIndia: 9	FALSE	0	NA	NA	NA	732617 NA	< a href="https://t.co/	replyToId	statusSource	screenName	retweeted	retweetCount	longitude	latitude				
19	RT @WhatevIndia: 10	FALSE	0	NA	NA	NA	732617 NA	< a href="https://t.co/	replyToId	statusSource	screenName	retweeted	retweetCount	longitude	latitude				
20	RT @WhatevIndia: 11	FALSE	0	NA	NA	NA	732617 NA	< a href="https://t.co/	replyToId	statusSource	screenName	retweeted	retweetCount	longitude	latitude				
21	RT @central bank of	FALSE	51	NA	NA	NA	732617 NA	< a href="https://t.co/	replyToId	statusSource	screenName	retweeted	retweetCount	longitude	latitude				
22	UK Business Minister	FALSE	89	NA	NA	NA	732617 NA	< a href="https://t.co/	replyToId	statusSource	screenName	retweeted	retweetCount	longitude	latitude				
23	EDCBIndia: Consent	FALSE	55	NA	NA	NA	732617 NA	< a href="https://t.co/	replyToId	statusSource	screenName	retweeted	retweetCount	longitude	latitude				
24	After approval from @	FALSE	79	NA	NA	NA	732617 NA	< a href="https://t.co/	replyToId	statusSource	screenName	retweeted	retweetCount	longitude	latitude				
25	EDCBIndia: Need to	FALSE	61	NA	NA	NA	732617 NA	< a href="https://t.co/	replyToId	statusSource	screenName	retweeted	retweetCount	longitude	latitude				

demon(1)-Excel																			
A1	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	text																		
2	Demetri	FALSE	0	NA	NA	NA	< a href="https://t.co/	replyToId	statusSource	screenName	retweeted	retweetCount	longitude	latitude					
3	RT @central	FALSE	0	NA	NA	NA	< a href="https://t.co/	replyToId	statusSource	screenName	retweeted	retweetCount	longitude	latitude					
4	RT @central	FALSE	0	NA	NA	NA	< a href="https://t.co/	replyToId	statusSource	screenName	retweeted	retweetCount	longitude	latitude					
5	RT	FALSE	0	NA	NA	NA	< a href="https://t.co/	replyToId	statusSource	screenName	retweeted	retweetCount	longitude	latitude					
6	RT	FALSE	0	NA	NA	NA	< a href="https://t.co/	replyToId	statusSource	screenName	retweeted	retweetCount	longitude	latitude					
7	@ndtvin	FALSE	0	NA	NA	NA	< a href="https://t.co/	replyToId	statusSource	screenName	retweeted	retweetCount	longitude	latitude					
8	Want to	FALSE	0	NA	NA	NA	< a href="https://t.co/	replyToId	statusSource	screenName	retweeted	retweetCount	longitude	latitude					
9	RT @central	FALSE	0	NA	NA	NA	< a href="https://t.co/	replyToId	statusSource	screenName	retweeted	retweetCount	longitude	latitude					
10	RT @central	FALSE	0	NA	NA	NA	< a href="https://t.co/	replyToId	statusSource	screenName	retweeted	retweetCount	longitude	latitude					
11	RT @central	FALSE	0	NA	NA	NA	< a href="https://t.co/	replyToId	statusSource	screenName	retweeted	retweetCount	longitude	latitude					
12	RT @central	FALSE	0	NA	NA	NA	< a href="https://t.co/	replyToId	statusSource	screenName	retweeted	retweetCount	longitude	latitude					
13	RT @central	FALSE	0	NA	NA	NA	< a href="https://t.co/	replyToId	statusSource	screenName	retweeted	retweetCount	longitude	latitude					
14	RT @central	FALSE	0	NA	NA	NA	< a href="https://t.co/	replyToId	statusSource	screenName	retweeted	retweetCount	longitude	latitude					
15	RT	FALSE	0	NA	NA	NA	< a href="https://t.co/	replyToId	statusSource	screenName	retweeted	retweetCount	longitude	latitude					
16	RT	FALSE	0	NA	NA	NA	< a href="https://t.co/	replyToId	statusSource	screenName	retweeted	retweetCount	longitude	latitude					
17	RT @central	FALSE	0	NA	NA	NA	< a href="https://t.co/	replyToId	statusSource	screenName	retweeted	retweetCount	longitude	latitude					
18	RT @central	FALSE	0	NA	NA	NA	< a href="https://t.co/	replyToId	statusSource	screenName	retweeted	retweetCount	longitude	latitude					
19	RT @central	FALSE	0	NA	NA	NA	< a href="https://t.co/	replyToId	statusSource	screenName	retweeted	retweetCount	longitude	latitude					
20	RT @central	FALSE	0	NA	NA	NA	< a href="https://t.co/	replyToId	statusSource	screenName	retweeted	retweetCount	longitude	latitude					
21	RT @central	FALSE	0	NA	NA	NA	< a href="https://t.co/	replyToId	statusSource	screenName	retweeted	retweetCount	longitude	latitude					
22	RT @central	FALSE	0	NA	NA	NA	< a href="https://t.co/	replyToId	statusSource	screenName	retweeted	retweetCount	longitude	latitude					
23	RT @central	FALSE	0	NA	NA	NA	< a href="https://t.co/	replyToId	statusSource	screenName	retweeted	retweetCount	longitude	latitude					
24	RT @central	FALSE	0	NA	NA	NA	< a href="https://t.co/	replyToId	statusSource	screenName	retweeted	retweetCount	longitude	latitude					
25	RT @central	FALSE	0	NA	NA	NA	< a href="https://t.co/	replyToId	statusSource	screenName	retweeted	retweetCount	longitude	latitude					

Figure3: Extraction of relevant tweets

3.1.2 Clean the Tweets:

Before performing data pre-processing we extract data from the twitter. To clean up the sentences with R we use gsub() function. Easiest way to clean the columns of our data frame with regular expression And last we get pre-processed data.

The cleaning of tweets requires the following steps:

- Remove html links from the tweets
- Remove retweet entities
- Remove all hashtags
- Remove all @people
- Remove all punctuation
- Remove all numbers
- Remove all unnecessary white spaces
- Convert all text into lowercase and
- Remove duplicates

3.1.3 Sentiment Classification:

Classification of sentiments can be done using the package 'sentiment' in R. First convert the tweets into a data frame.

In this paper #MakeInIndia and #Demonetization, tweets are analyzed using R language. Nearly 1000 tweets were extracted from the available tweets and the data analysis process was done on the above 1000 tweets.

3.1.4 Sentiment Score Classification:

The sentiment score is calculated as

$$\text{Score} = \text{Positive score} - \text{Negative Score}$$

The sentiment score is generated based on comparison of tweet words with positive and negative words lexicon. There are multiple ways to calculate sentiment scores. The approach followed here is to count the positive, neutral and negative words in each tweet and assign a sentiment score. This way, the tweet can be ascertained how positive or negative it is.

IV. Experimental Analysis

The output of the experimental analysis gives us various graphical representations for easy and better understanding of the #MakeInIndia and #Demonetization initiatives of government of India.

Fig-4 gives the trend of the number of tweets that were tweeted on Twitter over a period of time.

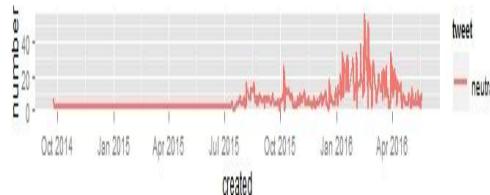


Figure 4: Year wise tweets analysis

Fig-5 provides the opinion of the users on #MakeInIndia initiative. The opinion is either positive or negative or neutral. This is the graph that represents the overall view of the society on the given initiative.

Among 1000 tweets analyzed for #MakeInIndia, 397 tweets were positive, 571 tweets are neutral and 32 are negative.

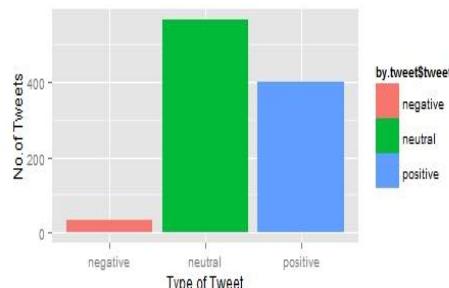
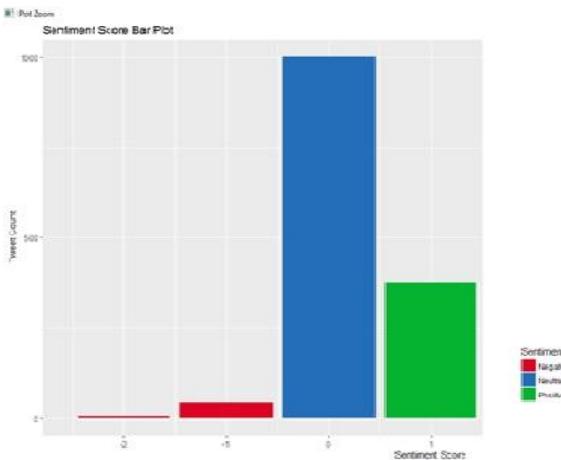


Figure5: Opinions of MakeIndia

Fig-6 provides the opinion of the users on #Demonetization initiative. The opinion is either positive or negative or neutral. This is the graph that represents the overall view of the society on the given initiative. Among 1500 tweets analyzed for #Demonetization, 454 tweets were positive, 1001 tweets are neutral and 45 are negative.



Opinions of Demonetization

Figure6: Opinions of Tweets

Fig-7 is the word cloud generated from the #MakeInIndia & #Demonetization tweets. A word cloud is the list of words used frequently in the tweets related to the topic.

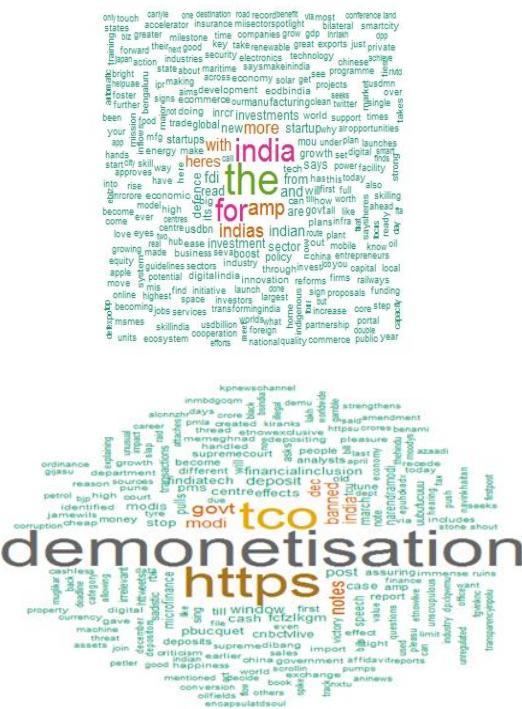


Figure 7: Word cloud on #MakeInIndia and #Demonetization

The above histogram shows the frequency of tweets with respect of scores allotted to each tweets. The x-axis shows the score of each tweet as a negative and positive integer. A positive score represents positive or good sentiments associated with that particular tweet whereas a negative score represents negative or bad sentiments associated with that tweet. A score of zero indicates a neutral sentiment. The more positive the score, the more positive the sentiments of the person tweeting and vice-versa.

The above histogram is skewed towards negative score which shows that the sentiments of people regarding #MakeInIndia&#Demonetization

The work in this paper **The sentiment behind #MakeInIndia & #Demonetization tweets** were analyzed explicitly indicates that majority Twitterati 57% responded are neutral towards these two initiatives of the government, while 40% have firmly welcomed it and a meager 3% number stood against the program. This makes the researcher believe that the government can go with these popular programs.

V. Conclusion

The study in this paper found that even though there is no negative opinion on Make in India and Demonetization initiative by the Government of India, majority of the people are neutral towards this program. Considerable amount of positive opinions have been given which shows that this initiative is a success. Such Analysis helps the government to make further improvements and implement any schemes that are useful for the country.

References

- [1]. A Survey of Data Mining Techniques for Social Network Analysis proposed by Mariam Adedoyin – olowie, Mohammed MedhatGabbar.
- [2]. TWITTER ANALYSIS (IN RSTUDIO USING R PROGRAMMING LANGUAGE) Prepared By: KAIFY RAIS in.linkedin.com/pub/kaify-rais/31/346/886/
- [3]. Social Networks Overview: Current Trends and Research Challenges November 2010, Coordinated by the “nextMEDIA” CSA, European Commission, Information society and Media
- [4]. An Approach to Harvesting, Cleaning, and Analyzing Data from Twitter Using R , Stephen Hill hills@uncw.edu , Rebecca Scott scottra@uncw.edu ,2016 Proceedings of the EDSIG Conference ISSN: 2473-3857 Las Vegas, Nevada USA
- [5]. Po-Wei Liang, Bi-Ru Dai, “Opinion Mining on Social Media Data” IEEE 14th International Conference on Mobile Data Management 2014.
- [6]. Alexander Pak, Patrick Paroubek, “Twitter as a Corpus for Sentiment Analysis and Opinion Mining”, 2015.
- [7]. MuqtarUnisa, Syed Raziuddin, PhD & Ayesha Ameen, “Opinion Mining on Twiter Data using Unsupervised Learning Technique” International Journal of Computer Applications (0975 – 8887) Volume 148 – No.12, August 2016.
- [8]. EfthymiosKouloumpis, Theresa Wilson, Johanna Moore “Twitter Sentiment Analysis: The Good the Badandthe OMG!” Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media.
- [9]. Evis DRINI, Dr. Rovena BAHITI, “Anti Money Laundering Reporting in a Cash-Less World”, IJCSN International Journal of Computer Science and Network, Volume 6, Issue 1, February 2017.
- [10]. Prof. Durgesh M. Sharma, Prof. Moiz M. Baig, “Sentiment Analysis on Social Networking: A Literature Review” International Journal on Recent and Innovation Trends in Computing and communication Volume:3 Issue:2 IIRITCC | February 2015.
- [11]. KEN KELLEY,KEKE LAI, AND PO-JU WU, “USING R FOR DATA ANALYSIS”, A Best Practice for Research, 2008.
- [12]. Lo Siaw Ling, “Mining Opinion from Twitter”, IISDV 2012 Nice Copyright 2012 Nanyang Polytechnic.

IOSR Journal of Computer Engineering (IOSR-JCE) is UGC approved Journal with S1. No. 5019, Journal no. 49102.

CH.R.Vinodkumar Sentiment Analysis of #MakeInIndia & #Demonetization using R.” IOSR Journal of Computer Engineering (IOSR-JCE) , vol. 19, no. 6, 2017, pp. 42-47.