Impact Factor Value: 4.029 ISSN: 2349-7084
International Journal of Computer Engineering In Research Trends
Volume 4, Issue 6, June-2017, pp. 252-258 www.ijcert.org

Twitter Sentiment Analysis on Demonetization tweets in India Using R language

K.Arun *[1], A.Srinagesh **[2], M.Ramesh**[3]

[1]Research Scholar, Department of Computer Science, Acharya Nagarjuna University, Guntur, India
[2]Associate Professor, Dept of CSE, RVR & JC College of Engineering, Guntur , India
.[3] M.Ramesh, Associate Professor, Dept of IT, RVR & JC College of Engineering, Guntur, India.

karun014@gmail.com [1], asrinagesh@gmail.com[2],mrameshmailbox@gmail.com[3],

Abstract :- In this global village social media is in the front row to interact with people, Twitter is the ninth largest social networking website in the world, only because of microblogging people can share information by way of the short message up to 140 characters called tweets, It allows the registered users to search for the latest news on the topics they have an interest, Lakhs of tweets shared daily on a real-time basis by the members, it has more than 328 million active users per month , Twitter is the best source for the sentiment and opinion analysis on product reviews, movie reviews, and current issues in the world. In this paper, we present the sentiment analysis on the current twitters like Demonetization, Indians and all over the world people are sharing their opinions on Twitter about current news in the country.

The sentiment analysis extracts positive and negative opinions from the twitter data set, R Studio provides the best environment for this Twitter sentiment analysis. Access Twitter data from Twitter API, data is written into txt files as the input dataset. Sentiment analysis is performed on the input dataset that initially performs data cleaning by removing the stop words, followed by classifying the tweets as positive and negative by polarity of the words. Generate the word cloud. Finally, that generates positive and negative word cloud, comparison of positive and negative scores to get the current public pulse and opinion

Keywords: Twitter Data, Text Mining, Sentiment Analysis, NLP, R-Studio.

1. Introduction

Twitter[4] is the best online platform for the sharing information and opinions, twitter is ninth largest social network in the world, now it has 328 millions of active users per month, and Lakhs tweets per day, sharing of information, opinion, feelings, likes very fast, only registered users can tweets and re-tweets, unregister users can only read the tweets. Users are mostly celebrities like presidents, prime ministers, politicians, film industry celebrities, sports stars and the common people registered as followers; users can tweet a current information in the form of text, video, audio or any format, the rest of the users can react on that tweets. In this way, information and reaction convey to the top level to the bottom level of the people at very fast. Resent year's feedback requirement is Increases, about the new product, about the government policy executions, and international talks, by the tweets and re-tweets it is the best platform for the feedback, opinion retrieval system in the social networks. Sentiment analysis using twitter data, which is classifying the positive and negative opinions from the tweets, and some deep mining about the positive and negative words. With this analysis, we can quickly identify positives and negatives of that tweets. In this paper, a system is proposed for the sentiment analysis on Demonetization twitters data using R programming language in step by step approach.

Here, in this work, we employ an open source approach for sentiment analysis and text mining using a set of packages supported by R language to mine the Twitter data and to carry out the sentiment analysis for Demonetization India. R-studio is free and open source IDE for developing and deploying R applications which can be installed on top of Linux/Macintosh/Windows. R language is a scripting language for conducting statistical data computing and big data analytics; it has more than 10000 packages[5].

2. Literature review

In Sentiment analysis is the recent trend in text mining [1]. Usage of text data in the social media has drastically increased; interesting patterns about people can be mine from the data. Sentiment analysis can be used to find the customer response to a product and news, it measures like, dislike sentiments of the people. Sentiment[6] analysis is a research and developing a stream of Natural Language Processing methods of the artificial intelligence. These types of researching techniques ranging from document-level classification to the polarity of sentiment words and phrases. Classifying the sentiment of Twitter messages is most similar to sentence-level sentiment analysis[11] for the limited sized tweets and the paragraph level sentiment analysis for more than one sentence. However, The Twitter follows micro-blogging nature, small tweets and also supports different natural languages, so Twitter is the best source for the sentiment analysis. Classification of tweets with polarity method having the SentiStrength scales from 1 to 5 for both positive (+1 weak positive to +5 extreme positive) and negative (-1 weak negative to -5 extreme negative) sentiments[2].

3. Proposed Work

The methodology to mine the Twitter data and to carry out the analysis is given in the Figure:

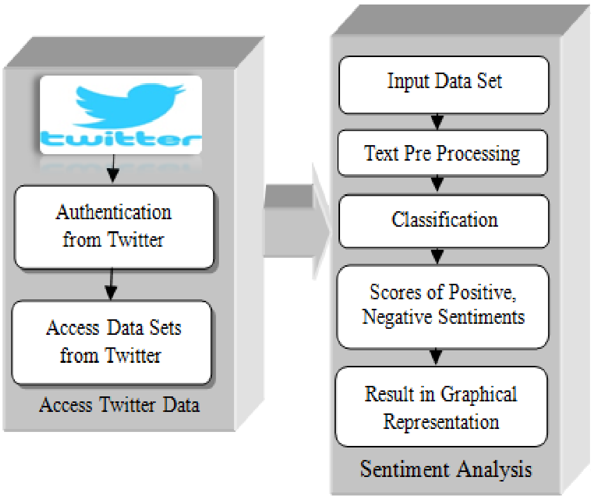
twitter
Authentication from Twitter
Access Data Sets from Twitter
Access Twitter Data
Input Data Set
Text Pre Processing
Classification
Scores of Positive, Negative Sentiments
Result in Graphical Representation
Sentiment Analysis

Figure 1: Proposed Methodology

The various phases involved in the Proposed Methodology are discussed below:

a)Twitter Authentication: Before mining any data from Twitter[3] using APIs, we have to authenticate with Twitter using an application created on Twitter. Once the application is created, we get access to consumer key, consumer secret, access token, access secret using which the API has to authenticate itself with the Twitter Authentication server.

```
consumer_key<-‘xxxxxxxx'
consumer_secret<-'xxxxxxxxxxx'
access_token<-'xxxxxxxxxxxx'
access_secret<-‘xxxxxxxxxxx'
setup_twitter_oauth(consumer_key,consumer_secret,acc
ess_token,access_secret)
```

b)Access twitter data sets: Once API is authenticated with Twitter Authentication service, a token is generated and is made available to API for every transaction with the Twitter server. Using this token, tweets are mined using hashtags. We use searchTwitter() function to access the data, and data set stored in .txt files.

```
searchTwitter('Demonetization',n=max,lang = 'en')
```

c)Text pre-processing:
Clean text: - the clean text is cleaned the unnecessary data from twitter data set, these are HTML Tags, White spaces, Numbers, Special symbols.

Remove stop words: - stop words are the bag of words based on the dictionary, that unnecessary words these are removed form the twitter data set, then the resultant data set contains only required information for the analysis.

d)N-grams: Applying N-grams plays an important role in text and sentiment analysis. These are uni-grams(n=1),
bi-grams(n=2),tri-grams(n=3),....N-grams. Bigram, also called Markov assumption [9], assumes that we can predict the probability of the future word by only looking at the last word encountered. We can classify bigram to trigram (verifying the previous two words in the sentence), and to N-gram (verifying the previous N-1words in the sentence). The general equation of N-grams is the conditional probability of the next word in a sequence would be;

$P(wn| w1n-1) \approx P(wn | wn-N+1n-1)$,where word sequence w1, w2, ... , wn-1 is represented as w1n-1.

$$P(wn| w1n-1) \approx C(wn-1wn)/C(wn-1).$$

e) Sentiment classification scores: - now we apply the sentiment classification on the cleaned data and n-gram features added to the basic twitter data sets.
Here two types of polarity methods are used in this application,

1. Polarity score for sentence level.
2. Polarity score for paragraph level.

Polarity score for sentence level: - The function polarity score is used to sentiment classification, [8] polarity.

scores are defined based on the dictionary which is used for the polarity. This polarity score function is applicable for find the sentiments according to the polarity scores at each sentence level, which is used by Hu, M., & Liu, B. the polarity method depends on the equation P=D_i^2/√(k)
Pre-Processed data that is cluster data D_i is the input to the polarity function. The clusters are tagged as follows:

1.neutral (D_i^0),
2.negator (D_i^N),
3.amplifier (D_i^a), or de-amplifier (x_i^d).

No values in the neural words in the equation, but neural words considered in the word count (k). Moreover, each polarized word is then weighted w based on the weights from the polarity.
In the Twitter, sentiment analysis polarity score for the sentence level is the best method because Twitter follows microblogging nature,
Almost all tweets or re-tweets are only one are two line sentences.

Polarity score for paragraph level: - The second method for paragraph level polarity score, method, and formula according to the Jockers, M. L [9]. Each paragraph (prg_i = {sent_1, sent_2, ... sent_n}) collection of sentences, is divided into element sentences (sent_i,j = {wrd_1, wrd_2, ... wrd_n}) where word are the words within sentences. Each sentence (sent_j) is separated into a sequence of a bag of words. By the text pre-processing technique all unnecessary and punctuation makes are removed from the sentence. for the better convenience Here i consider pause words as CW (comma words). Now we can represent these words as an i,j,k notation as wrd_{i,j,k}. For example, wrd_{5,1,3} would be the third word of the first sentence of the fifth paragraph. It is applicable to the conversation of particular talk.

The polarity context cluster input pre-processed twitter data (cc_{i,j,l}. the clusters are tagged as follows:

1.neutral (wrd_i,j,k^0),
2.negator (wrd_i,j,k^N),
3.amplifier (wrd_i,j,k^a), or de-amplifier (wrd_i,j,k^d).
P=p'_i,j,l/√(wrd_i,jn)

And it is similar to the polarity score for the sentence.

f) Graphical Representation : - finally the sentiment analysis can be represented in graphical modes, with the usage of R-studio, there are a rich set of graphical tools are supported by R packages, in this paper, only some methods are used to represent the effective and attractive outcomes of the sentiment analysis using word clouds, ag plots, bar charts.

1.wordcloud(names(v), v, scale , min.words, max.words, colors=brewer.pal(range, "Dark2"))
2.ggplot(test,aes,geom_bar,geom_text,theme)

4. Experiment and Results:

In this work, data collected from Twitter for analyzing Demonetization is the Government decision. Indian people face so many problems till now, Because of the demonetization decision in nov-2016, the main object of the PM Narendra Modi is demolishing the black money in Indian markets. In early days of demonetization decision that is just like emergency days of India, but Indian government takes proper alternates for execution of the Demonetization from Jan-17 to May-17, so Indian people almost step out from that problem, all these feelings are considered in this analysis. Along with this government of India bare good fruits for the nation in different flavours, these are Black money controlling, Digital payments increasing, Income Tax payments increasing, GDP of India drastically increases up to 7.3%,and also the complete study of the demonetization is depended on the following data sets, these are digital payments, Income Tax payments, and Operation Clean Money, in the following table contains twitter tag, number of tweets and size of the Twitter file.

Table No:1 Twitter Data Analyzed

Name of the Twitter	**No. of tweets**	**Size**
Demonetisation	10,000	1.2MB
Digital Payments	2000	255KB
Operation Clean money	2358	287KB
Income tax payments	1983	252KB

Let us first verify Demonetization sentiment details

Table No:2 Twitter Sentiment Classification Calculated Values

Total tweets	**Positive Tweets**	**Negative Tweets**	**Neutral tweets**	**Avg polarity**
12974	2974	4936	5064	-0.09111864

The polarity operation is applied on pre-processed data set, the data which contains cleaned data with bigram features, the polarity function can generate the sentiment scores for each tweet, if it is negative or positive tweets, we need what are the positives an negative from the public. These are represented by word cloud and ag plots in the following.

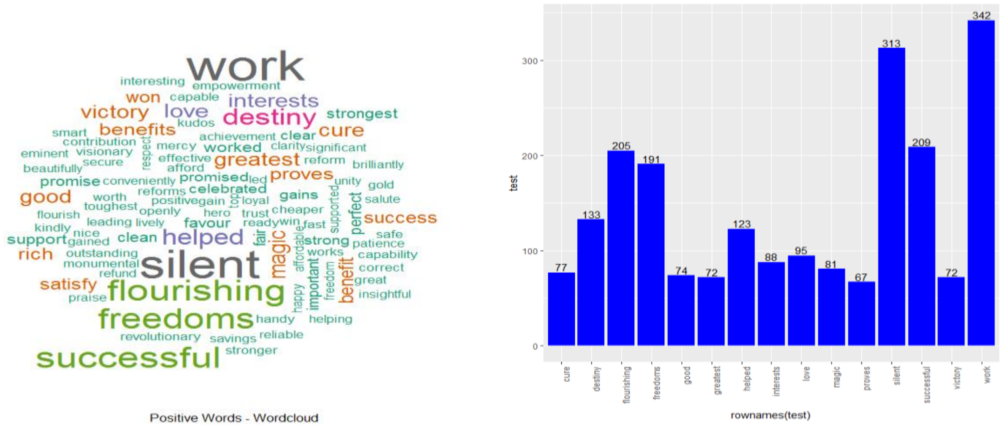

Figure 2: Positive Word Cloud and Plot

The above results are only positive word cloud and positive go plot. The bold and big words in the word cloud and highest poles in the ggplots are the words which contain maximum term frequencies. Similarly, negative wordcloud and bar plot using ggplots are the followings

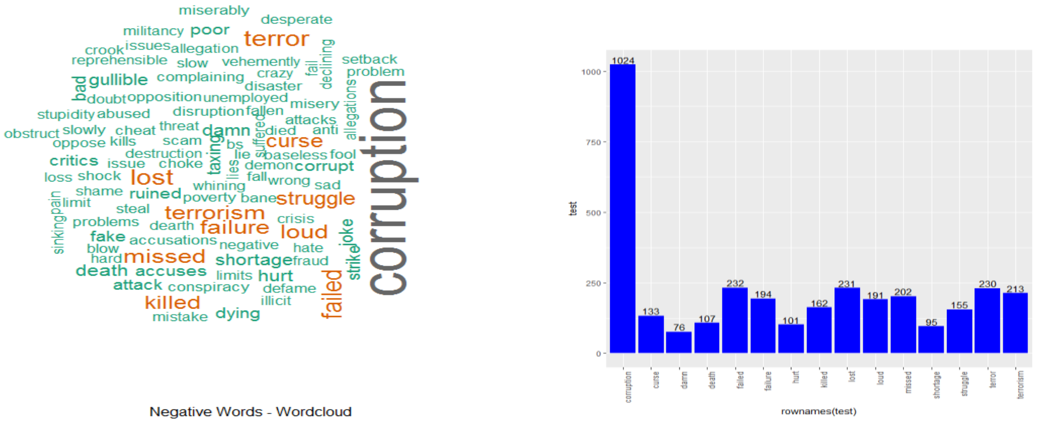

Figure 3: Negative Word Cloud and Plot

Moreover, finally, sentiment scores of the demonetization is in pie chart

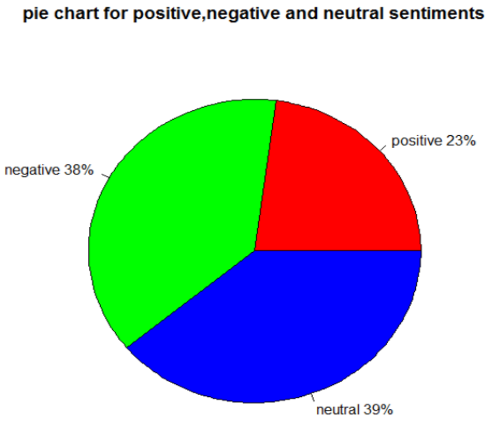

Figure 4:Pie-Chart Visualization for Positive, Negative and Neutral Sentiment Analysis

The pie chart is divided into three parts, red for positive percentage is 29%, green for negative percentage is 38%, and blue for the neutral word percentage is 38% sentiment scores of the Twitter data. Neutral words have the zero polarity score, no effect on the sentiment, only for total word count, but here negative sentiments are (4936 tweets) more compared to the positive sentiments(2974 tweets) in the sentence level polarity score.It seems negative sentiments are more, that means people are still facing problems in different ways, in future these problems may clear with proper alternative executive plans by the government of India. Even though problems are there, But it improves the income sources of India like income tax payments, digital payments, in 2016-2017 GDP rate in India it increases up to 7.3%, and also it is hopeful to get 8% in next coming economic year 2017-2018, it is the best result to the country. This opinion also is taken from the Twitter to supporting for the demonetization, in the following comparison and analysis on the Digital payments, Income payments and operation clean money data sets.

Name of the tweets shortcuts:

DP: Digital Payments,

ITP: Income Tax Payments
OCM: Operation Clean money

Table No:3 Twitter Sentiment Classification Calculated Scores

Twitter	Positive	Negative	Neutral	Total	Avg polarity
DP	769	235	1496	2496	0.072
ITP	779	510	1237	2526	0.03
OCM	454	123	28	605	0.24

Hence these three data sets are getting a more positive score; all average polarities are positive, so the public sentiments or opinion is purely positive for these tweets.DP: Digital payments are increased with the effect of the demonetization, in near future it increases more, in the word cloud there are positive words from the twitter data set, and the pie chart represents sentiment scores, which is 31% positive, only 9% negative, and finally 60% neutral words, here positive is more.

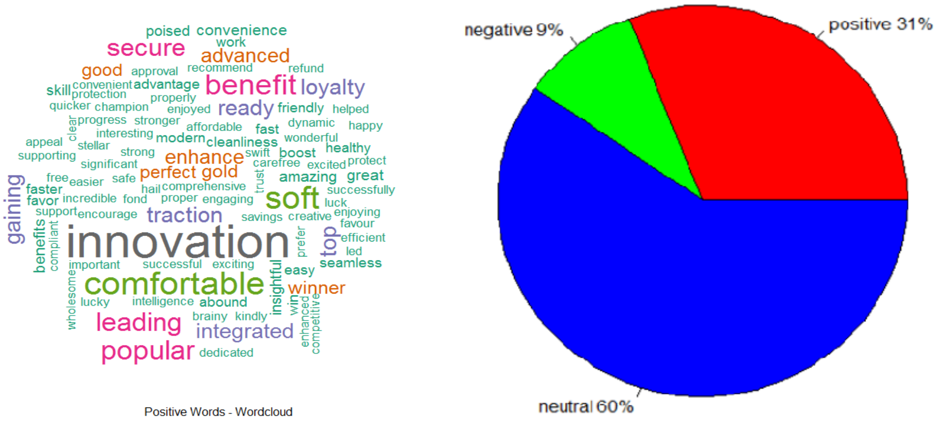

Figure 5: Tweet Visualization for Positive Word Cloud Sentiment Analysis for tweets

ITP: Income Tax Payments are drastically increased only with this effect, people also very positive for tax payments, in the word cloud there are positive words from the twitter data set, and the pie chart represents sentiment scores, which is 31% positive, only 20% negative, and finally 49% neutral words, here positive is more.

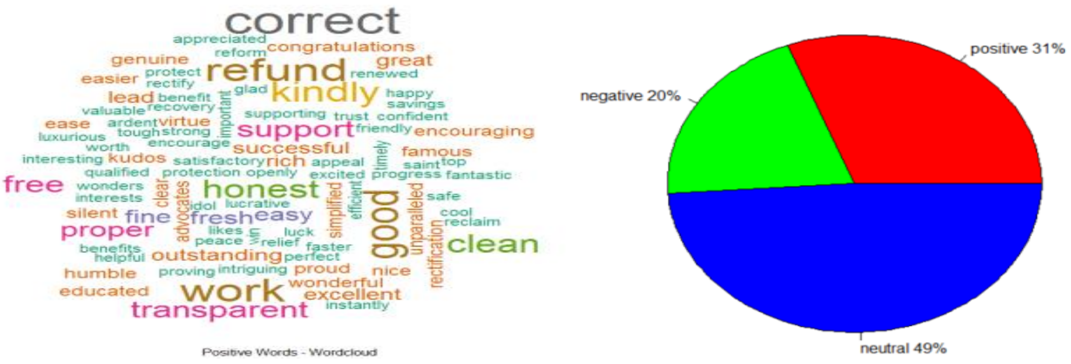

Figure 6: Tweet Visualization for Positive Word Cloud Sentiment Analysis

OCM: Operation on clean money is organized by Income Tax Departments of India, in this twitter data more direction from the Indian government about the clean money, actions on black money, and more government policies to clear the problems of the public.

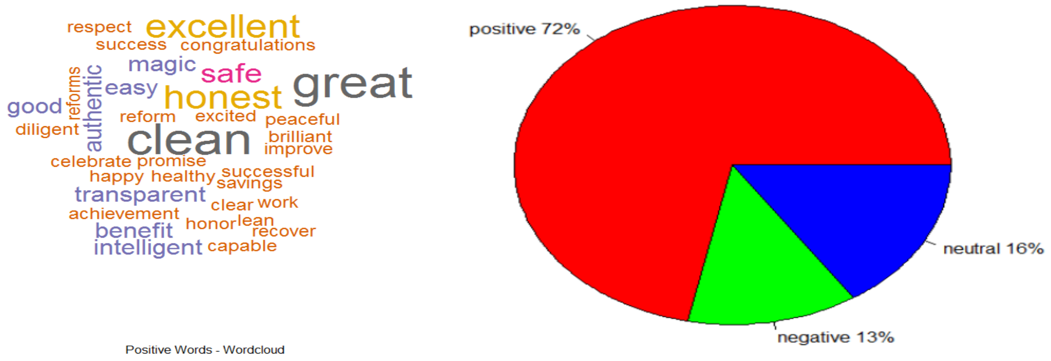

Figure 7: Tweet Visualization for Positive Word Cloud Sentiment Analysis

It is observed that in the analysis, it is also getting the positive 72% opinion more at an acceptable level. In all the above, we can observe that all positive sentiments are more compared to the negative. Hence the demonetization is in progress, most of the people are accepting is positive.

5. Conclusion

In this work, a new method is proposed to get the sentiment analysis on the demonetization from the twitter data sets, tweets are samples of the society opinions, with this sample tweets we can get the sample positive sentiments and negative sentiments of the demonetization decision of the Indian government, in this process data cleaning, bigrams, polarity, and sentiment scores and graphical methods are used for this Twitter sentiment analysis. In future work the Following aspects such as: to Apply the sentiment analysis on kidding words in the tweets, Handling huge twitter data sets in Big Data are proposed to be implemented.

References

1) Efthymios Kouloumpis. Theresa Wilson, Johanna Moore “Twitter Sentiment Analysis: The Good the Bad and the OMG!” In the Proceedings of Fifth International AAAI Conference on Weblogs and Social Media ,2011.

2) Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., & Kappas, “A Sentiment strength detection in short text”, Journal of the American Society for Information Science and Technology, 61(12), 2544–2558, 2010

3) https://apps.twitter.com/app/13647643, Date accessed:12/04/2017

4) https://about.twitter.com/company,Date accessed:12/04/2017

5) http://blog.revolutionanalytics.com/2017/01/cran-10000.html

6) Yu.H and Hatzivassiloglou.V “Towards answering opinion questions: separating facts from opinions and identifying the polarity of opinion sentences” In the Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-03),NOV 2003.

7) M De Choudhury, YR Lin, H Sundaram, KS Candan, L Xie, A Kelliher “Authoritative Sources in a Hyperlinked Environment” International AAAI Conference on Weblogs and Social Media 2010, ed. Conference Program Committee, AAAI Press, New York., MAY 2010, pp-34-41.

8) Hu, M., & Liu, B. (2004). Mining opinion features in customer reviews. National Conference on Artificial Intelligence.

9) https://lagunita.stanford.edu/c4x/Engineering/CS-224N/asset/slp4.pdf, Date accessed:21/04/2017

10) Jockers, M. L. (2017). Syuzhet: Extract sentiment and plot arcs from text. Retrieved from https://github.com/mjockers/syuzhet.

11) Sunil B. Mane, Kruti Assar, Priyanka Sawant, & Monika Shinde. (2017). Product Rating using Opinion Mining. International Journal of Computer Engineering in Research Trends, 4(5), 161-168. Retrieved from http://ijcert.org/ems/ijcert_papers/V4I503.pdf.