International Journal of Data Analytics (IJDA)
Volume 01, Issue 01, January-December 2016, pp. 1-6, Article ID: IJDA_01_01_001
Available online at https://iaeme.com/Home/issue/IJDA?Volume=1&Issue=1
Journal ID: 6202-8020, DOI: https://doi.org/10.17605/OSF.IO/HAQMY

© **IAEME** Publication SCOPE DATABASE INDEXED

OPEN ACCESS

ANALYSIS OF TWITTER DATA BY EXTRACTING THE TWEETS ON DEMONETIZATION

Lakshmi Namratha Vempaty
International Institute of Information Technology Bangalore,
Bangalore, India
https://orcid.org/0009-0005-8426-8577

ABSTRACT

This report is on how to perform analytics on tweets of particular topic. In this we are going to review the tweets on demonetization. This report also presents various findings and visualisations while analysing the tweets extracted/provided by twitter on the topic demonetization and also the context is confined to analysis of twitter data.

Keywords: Tweets, Demonetization, Analytics of data, R, Twitter

Cite this Article: Lakshmi Namratha Vempaty, Analysis of Twitter Data by Extracting the Tweets on Demonetization, International Journal of Data Analytics (IJDA), 01 (01), 2016, pp. 1–6.
https://iaeme.com/Home/issue/IJDA?Volume=1&Issue=1

I. INTRODUCTION

On 8th November, 2016 Prime minister of India has announced that 500Rs and 1000Rs notes are no longer a legal tender. As a fight against black money and to cut the funding to terrorists, the central government of India took a controversial decision by banning the high denomination currency notes. Public got split on their opinions/views/experiences one supporting the currency ban, one against the ban and the other with neither of those opinions.

Social media platforms have become a place to quickly analyse the reactions, opinions of the people. The data produced by these platforms can be used to explore and analyse sentiment on various current issues/topics. Demonetization has become the topic of the hour all over the internet with people expressing their opinions on facebook, tweeting on twitter. Since then it is in huge interest to analyse the sentiment of the people based on data generated on the social media platforms like facebook, twitter etc. however twitter has been the most reliable data source due to its affordances like hashtag search, twitter API for data.

II. TWITTER AS A SOCIAL TOOL

Twitter is considered as one of the most powerful social media platform/communication tool. In this section well be looking at why we considered twitter as a tool for data extraction? and why it is the ideal platform for our purpose? Every social media platform has its own individual purpose for example facebook is used for maintaining friend network, catch up with friends, LinkedIn is to maintain professional relationship with co-workers etc, similarly twitter has been the place where people present opinions/views/experience publicly to friends and strangers. The 140 character limit makes the message crisp and clear to understand ones intent. As the tweets are available for not only followers but everyone else and also the reach (no. of users viewed the post) of the post is high and fast. These affordances are not provided by any other social media platforms which makes twitter stand out of the others.

Also one can search about a specific hashtag to get relevant tweets about the topic. Twitter provides API for collection of tweets of one account or tweets from various accounts with certain hashtag. In the context of data collection on a specific topic or burning issue it portrays the diversity in audience which help to understand the topic from different views. With all these affordances provided by twitter it has been the ideal platform for analysing and understanding publics reaction.

III. LEITERATURE REVIEW ON HOW PEOPLE FROM DIFFERENT INDUSTRIES REACTED TO DEMONETIZATION

The visualisations presented below represent the Industry wise segmentation of effects of demonetization. Clearly healthcare industry is the most effected with a negative sentiment of 66%, real estate, wedding, travel industries are the next in the list of effected industries due to currency ban.

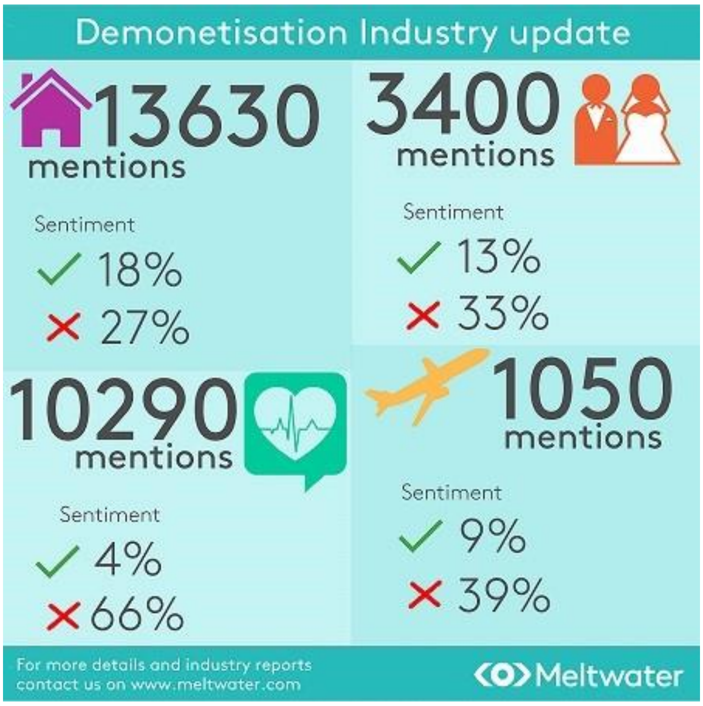

Fig. 1. Demonetization Industry Update

Figure 2 demonstrates the industry wise segmentation of twitter mentions with real state on top of the list with 42.8% followed by health care with 32.39% and wedding industry with 10.56%. These figures dont represent the actual effect of demonetization on respective industries but how people reacted relevant to the particular industry on twitter.

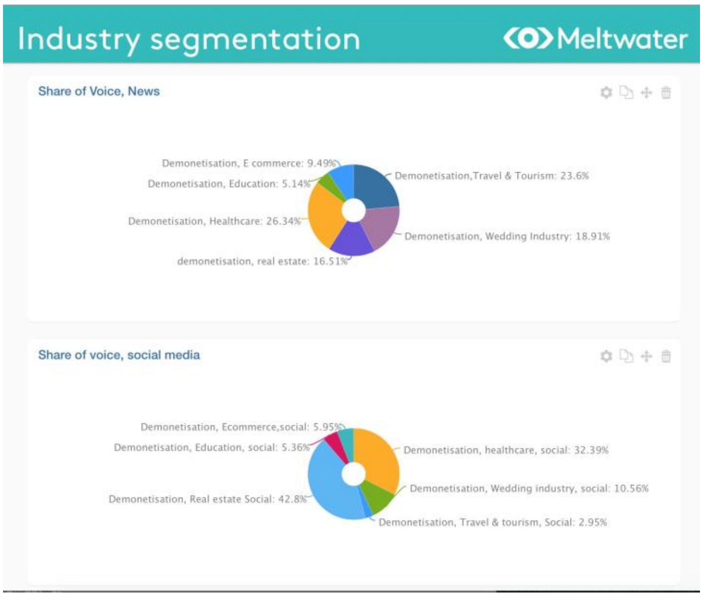

Fig. 2. Pie diagram Representation industry-wise

IV. METHODOLOGY FOLLOWED FOR THE EXTRACTION OF TWEETS ON DEMONETIZATION FROM TWITTER

To obtain the tweets, we need to extract the tweets related to the topic from twitter. So to do that the following process is followed:

A. Data Extraction:

The following are the three features which are a part of data extraction.

1.Volume: Almost 8,000 tweets (includes retweets, favourites) have been collected and provided as a (dataset) csv file which is hosted on the www.kaggle.com/datasets website. Along with these 1,000 most recent tweets are collected using twitter API in given data frame.

2.Timeframe: The tweets which include #demonetization are collected from Nov9th- 23rd, 2016.

3.Tools: RStudio has been used as an IDE and R language has been used for coding purpose.

B. Data Preparation/Cleaning:

The data is obtained in the form csv file. It has 8,001 rows with 15 columns. The metadata (columns) consists of tweet, favorited, favoriteCount, replyToSN, createdtruncated, replyToSID, id, replyToUID, statusSource, screenName, retweetCount, isRetweet, retweeted. The data is cleaned such that punctuations, unnecessary, common stop words and retweets have been removed.

V. RESULTS AFTER PERFORMING ANALYTICS ON THE COLLECTED DATA

By using the data we can perform many types of analytics which are as follows:

A. Exploratory Analytics

After cleaning of data as mentioned in the data preparation, a word cloud has been created from the corpus of the data after preparation. Word cloud represents frequency of the words, the more frequent the word is the bigger the size in the word cloud. Some observed common words are modi, bank, people, terrorists, support.

From these frequent words it can be understood that twitter reacted positively to demonetization but this inference needs to be confirmed in the further analysis.

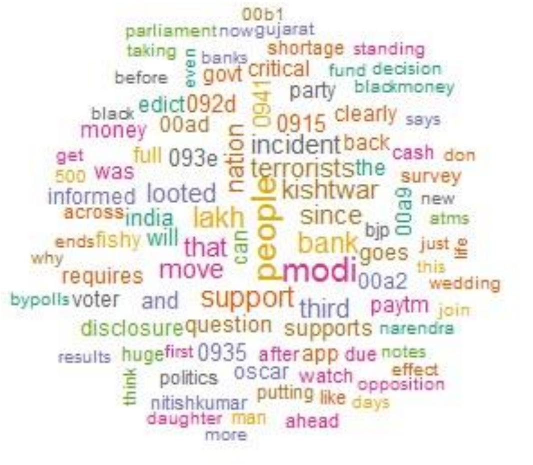
00b1
parliament now gujarat
taking banks shortage standing
before even govt critical fund decision
black edict 092d 0941 party blackmoney
money 00ad 0915 clearly says
get full 093e nation incident back cash don
500 was terrorists the survey
informed looted kishtwar new
across india lakh people since 00a9 atms
ends fishy will that can bank bjp just life
why goes this
requires move modi 00a2 wedding
bypolls voter and support third paytm join
disclosure question supports narendra
results huge first 0935 after app due notes
think politics oscar watch effect opposition
nitishkumar putting like days
daughter man ahead
more

Fig. 3. Word Cloud

B. Sentiment Analysis

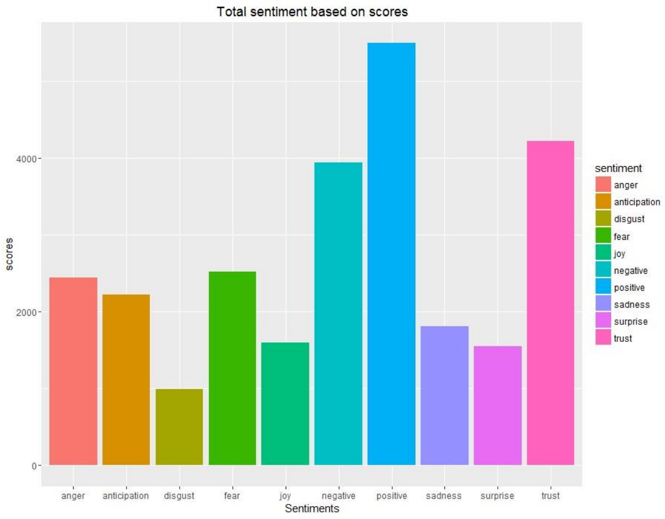

Fig. 4. Total sentiment based on scores

A sentiment analysis has been performed using the function get nrc sentiment available in R package syuzhet and the result has been plotted as shown in the above figure. A single can be categorized in to any number of sentiments out of 10 available. Positive, trust are highest observed sentiments from the data however anger and fear have also been prominent in the tweets. Even though it is evident from the above figure that positive is higher than negative, implying that internet, especially twitter has supported the currency ban decision, due to mixed emotions we cannot adjudge that our inference in exploratory analysis is correct.

The figure below represents sentiment analysis with polarity in which sentiments are categorized in to negative, neutral, positive. For each tweet the algorithm calculates the sentiment score and based on final score the tweet will be classified as negative or positive or neutral. From the above figure it can be observed that almost 3,400 tweets are classified as positive 3,200 tweets are classified as negative and 1,400 tweets are classified as neutral.

We can see the decrease in the difference between positive and negative from this figure to the previous one as some are classified as neutral.

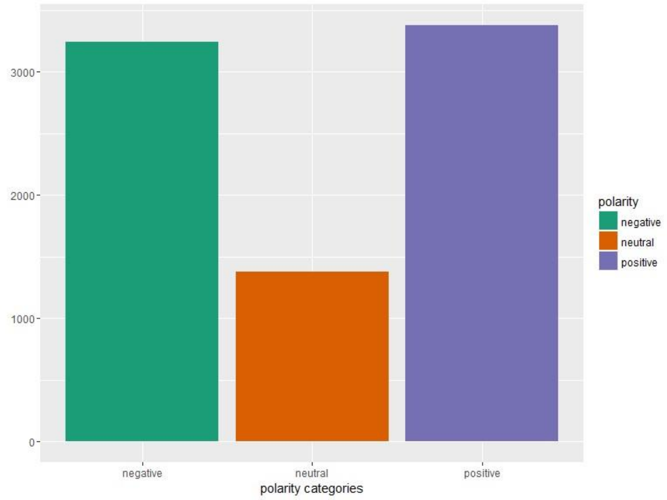

Fig. 5. Sentiment analysis with polarity categories

Hence we can confirm our inference/hypothesis that public has supported the decision of banning the 500Rs and 1000Rs currency notes.

C. Analysing the Data After Clustering

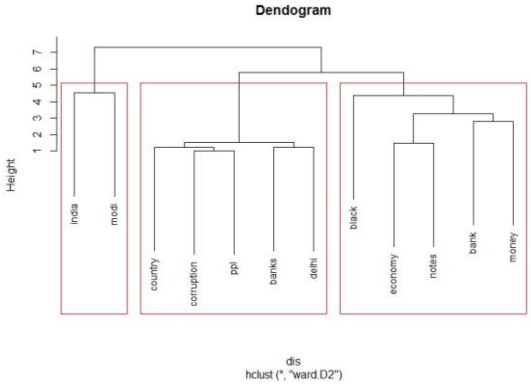

Fig. 6. Hierarchical Clustering on the words extracted from tweets

The above figure is the plot of hierarchical clustering on the words extracted from the tweets. Hierarchical clustering produces a set of nested clusters organized as a hierarchical tree. This can be visualized as a dendrogram, a tree like diagram that records the sequence of merges or splits. In the resulting dendrogram we obtained, cluster-1 specifies tweets about modi, cluster-2 represents people talking about corruption in the country, and cluster-3 represents people chattering about economy of India and banks. These clusters can be interpreted as co-occurrence of the leaf nodes together or frequency of the words. The height represents the dissimilarity between these clusters i.e. when two leaf nodes are at same height that implies both are in this context co- occurring together.

VI. CONCLUSIONS

Analysis of social media data has always been in the interest of the analysts because of its ease to collect data on topic of interest to evaluate sentiment of people, reach of advertisement etc. From the sentiment analysis on the tweets data it is evident that people have supported 500Rs, 1000Rs note ban however the margin between support and oppose is less. It is also observed that effect of demonetization varied for different industries with the effect on real estate and health care industry being most mentioned.

REFERENCES

[1] Rathee, A. Demonetization in India Twitter Data. Retrieved from Kaggle: https://www.kaggle.com/arathee2/demonetizationin-india-twitter-data.

[2] Social Samosa. Social Samosa. Retrieved from socialsamosa.com: https://www.socialsamosa.com/2016/11/data-social-mediareactions-demonetization/Appendix A - Techniques in R that are used to perform analytics on Twitter data

[3] To get the word cloud, use library(”wordcloud”) and to generate the word cloud use wordcloud(tweets.text.corpus,min. freq = 2, scale=c(7,0.5),colors=bre wer. Pal (8,” Dark2”), random.color= TRUE,random.order = FALSE, max.words =250).

[4] We need to remove links,punctuation marks,blank spaces ,tabs and the user name from the obtained tweets as part of data cleaning, because these are unnecessary for analytics. 3. To remove something from a text we need to use the funtion gsub. So in case we want to clean the html links we use new txt¡-gsub(” http[ˆ[:blank:]]+”,””,ini txt).

[5] To get the plot of hierarchical clustering on the words extracted from the tweets,clust¡-hclust(dis,method=”ward.D2”), plot(clust,cex=0.9,hang=1,main= ”Dendogram”), rect.hclust(clust,k=3).

[6] The algorithm used to classify emotion and polarity is bayesclassifier.In R we use class emo = classify emotion(some txt, algorithm=”bayes”, prior=1.0) and for polarity class pol = classify polarity(some txt, algorithm=”bayes”).

Citation: Lakshmi Namratha Vempaty, Analysis of Twitter Data by Extracting the Tweets on Demonetization, International Journal of Data Analytics (IJDA), 01 (01), 2016, pp. 1–6

https://doi.org/10.17605/OSF.IO/HAQMY

Article Link:
https://iaeme.com/MasterAdmin/Journal_uploads/IJDA/VOLUME_1_ISSUE_1/IJDA_01_01_001.pdf

Abstract:
https://iaeme.com/Home/article_id/IJDA_01_01_001

editor@iaeme.com