RESEARCH ARTICLE OPEN ACCESS

Text mining and Sentiment Analysis of tweets on Demonetization in India

Varsha KR
Department of Computer Science and Engineering
RV College of EngineeringBangalore, India
varsha.rajendrakg@gmail.com

Abstract:

This paper focuses on analysis of Twitter data during demonetization in India and the sentiments and emotions exhibited by the people in the form of twitter tweets in the year 2016-2017. This analysis is carried out using R tool which includes collection, pre-processing, classification and visualization of twitter data/ text of tweets to derive meaningful insights of the data and the general common response for the decision of demonetization.

Keywords **— Text mining, sentiment analysis, term-document frequency matrix, wordcloud**

I. INTRODUCTION

The monetary stability of a country always depends on its government policy. One has to follow a far stricter code of austerity all-round. Government has its duty to curb conspicuous consumption by way of unearthing black-money. In this direction, the recent demonetization of higher denomination by the Government of India was expected to bring lot of positive effects in the strengthening of economy. This step was appreciated by the citizens in a large scale. In order to evaluate the reaction of common public in regard to this decision or any other decision that can be taken in the future, analysis of social media texts can be effective.

In this paper, we discuss the text-mining of data obtained by twitter tweets and estimate the emotion and sentiment of people through these tweets. This analysis is done in R with the aid of various libraries associated with it.

II. Techniques used

A. Text Mining

It is a procedure of obtaining useful information from text. In this paper, text mining is used to evaluate the reactions of common public about demonetization through mining the text in twitter data or 'tweets'.

B. Sentiment Analysis

It is a post-procedure of text mining. The metadata obtained after text mining is classified into different emotions and sentiments [1] using a trained classifier. The classes considered here are – 8 emotions: anger, fear, disgust, anticipation, trust, surprise, sadness, joy and 3 sentiments: positive, negative, neutral.

III. PROCEDURE

The procure employed here for the study of texts in the tweets involves many steps starting from data collection till the analysis of obtained results [2]. The entire procedure is carried out on R because there are multiple libraries available for all the steps

involved and the processing is faster when compared to other tools.

A. Data Collection

In order to analyse tweets with good accuracy, a large number of tweets along with details about user, time is required[3]. To extract tweets from R, 'ROuth Authentication' is used as a channel and the data can be saved as 'R.Data' format. There are also multiple datasets available in the web for the same period which can be modified according to the requirement. Here, we have used a dataset consisiting about 14,000 tweets from world-wide during the period 2016-2017.

B. Data PreProcessing

The twitter data collected cannot be used directly. It needs to be modified according to the requirements [4]. The twitter text consists of date along with the message. This needs to be sepearated and date needs to be converted into date format. Then, 'hour' is extracted to plot data on hourly basis.

In the twitter message data, text is composed of '#' tags, 'emoji' and sop words which needs to be removed as follows:

a) First blank space, tabs, punctuation, special characters and numbers using 'R regex' method.

b) All words are converted to lowercase text

c) Then, stopwords and fillers are removed. For this the 'tm' – Text mining package in R is used.

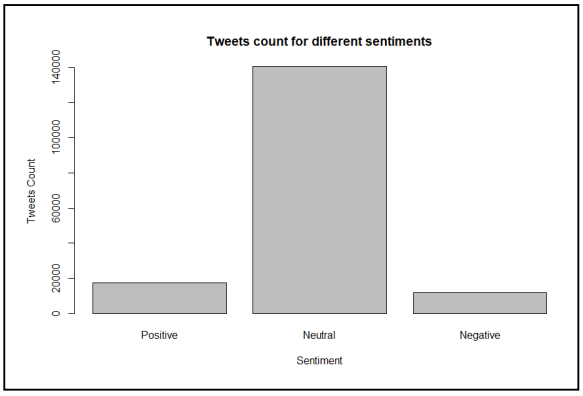
d) Then a unique set is obtained by removing duplicates.

e) Username, links and # tags are removed to get a clean raw text.

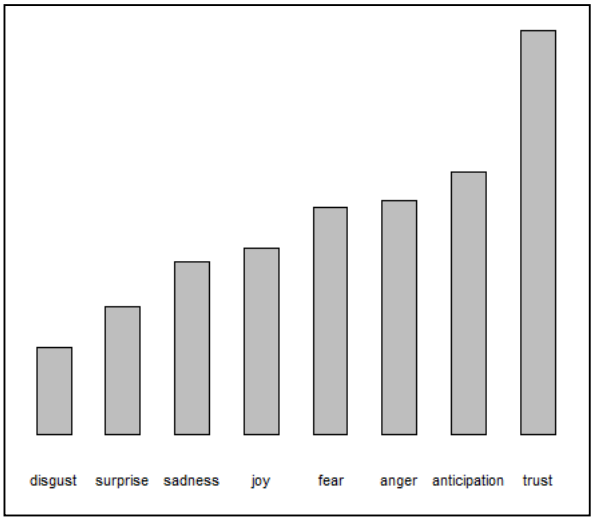
f)

C. Data Classification

The pre-processed data is imported into R and converted to a term-document frequency matrix. Each row is classified as positive, negative or neutral using RSentiment package. The polarity is saved as category name using the 'Bayes' algorithm[5]. For each record, the hour is extracted from 'created date'. The classification obtained from the matrix is concatenated with the data of tweets having hour data. The sum of records on each category is calculated and plotted against the hour. The same matrix is used for creating word-clouds.



The term-document matrix is also classified as 8 emotions[6] + 2 sentiments with 'classify_emotion'. The overall sentiment analysis on the tweets is done



using the 'syuzhet' package and represented as shown in the figure.

D. Data Visualization

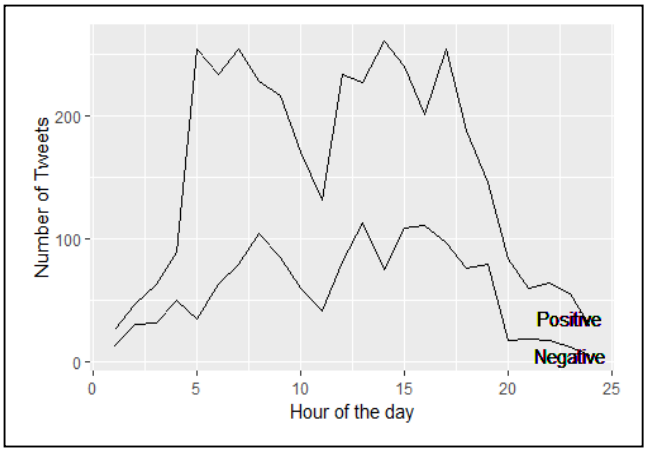
The modified data can be represented in various forms, so that better insights can be retrieved from the plots [7]. These plots are obtained from various libraries available in R.

Here, the data is represented with 3 types of plots – line-graph, bar-graph and word cloud. Line-plots are used to represent the number of positive and negative tweets distributed over 24 hours of a day. The difference between the positive and negative lines is the measure of difference in the number of reactions. The bar plot is used to show the classification of tweets into 8 emotions + 2 sentiments. The bars are sorted in the order of 'number of tweets' showing most seen emotion at the end of x'axis.

Word clouds are a unique way of representation, where the actual words used can be seen. Here, the more the frequency of the word as seen in the term-document frequency matrix, the bigger the size of the word in the word-cloud. Hence word clouds show words and frequency at the same time for multiple words giving a holistic view.

IV. ANALYSIS OF RESULTS

Analysis confirms the demonetization act as more utility based than a futile action. The dataset here is composed of tweets worldwide and the analysis of these texts show that the action has been appreciated world-wide. It can also be seen that there were a few negative responses too, but does not equal positive reactions.



A. Sentiment Analysis

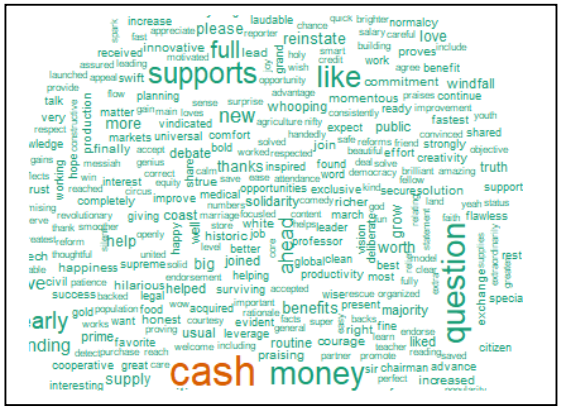
From the graphs obtained, we can see the distribution of text in the tweets over 3 categories – Positive, negative and neutral. The high number of neutral records obtained can be attributed to the inadequate text preprocessing and hence can be omitted for analysis.

On comparing the positive and negative records, the positive ones are comparatively high. The plot over 24 hours also shows that the positive tweets are high in number all through the day.

B. Word Cloud Analysis

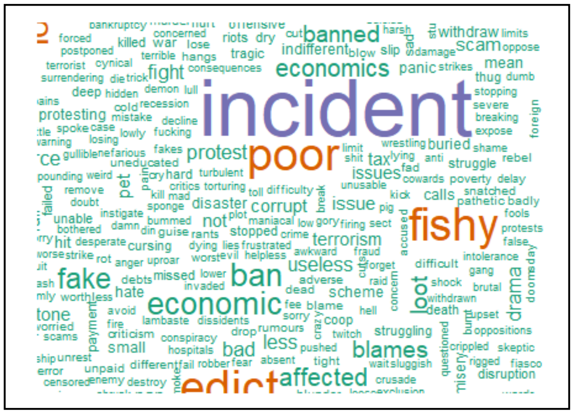
Word clouds are an amazing visualization tools. In an instant, most of the reactions can be captured. In this analysis, we have created 2 word-clouds – positive and negative, which show various 'most-used' words. The clouds also represent words according to the frequency. Most frequent words are displayed with bigger size.

The positive word cloud shows many words, but not many words in bigger size. This shows that positive responses are not focused on any one kind of topic and is distributed. Though words like 'dear','money','cash' are considered common words in this scenario, they are bigger and are classified as positive because of their association with other positive words in the text. Words like 'support', 'like', 'good', 'benefit', 'successful',



'love', 'standing', 'thanks', 'huge' – all of these show huge support and appreciation for the demonetization decision.

On the other hand, the negative clud shows many bigger words, which indicate few groups with concentrated negative reactions eg. ‘poor’. Words like ‘economic’, ’shortage’, ’incident’,’ban’ can be seen in the cloud that represent negative reactions. In comparision, the distribution of these words is more in the first half of the dataset.



C. Emotions Classification Analysis

Here, the tweets are classified into 8 emotions excluding 2 sentiments. From the graph, it is evident that maximum number of tweets belong to ‘trust’ category which shows the trust Indians had on the policy, followed by ‘anticipation’. Though there are a large number of tweets for ‘anger’ and ‘fear’, it can be attributed to the confusion and problems that the policy caused. Overall, the tweets belonging to ‘positive’ category is significantly higher.

V. REFERENCES

[1] Vishal A. Kharde, S.S. Sonawane, “Sentiment Analysis of Twitter Data: A Survey of Techniques,” International Journal of Computer Applications (0975 – 8887) Volume 139 – No.11, April 2016

[2] Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, Rebecca Passonneau “Sentiment Analysis of Twitter Data”, Department of Computer Science Columbia University New York, NY 10027 USA.

[3] Mitali Desai, Mayuri A. Mehta, "Techniques for sentiment analysis of Twitter data: A comprehensive survey", Computing Communication and Automation (ICCCA) 2016 International Conference on, pp. 149-154, 2016.

[4] Eman M.G. Younis Kaur, "Sentiment Analysis and Text Mining for Social Media Microblogs using Open Source Tools: An Empirical Study", International Journal of Computer Applications (0975 – 8887) Volume 112 – No. 5, February 2015.

[5] Harsha Sinha, Arashdeep Kaur, "A detailed survey and comparative study of sentiment analysis algorithms", *Communication Control and Intelligent Systems (CCIS) 2016 2nd International Conference on*, pp. 94-98, 2016.

[6] Kamath S Sowmya, Anusha Bagalkotkar, "Sentiment Analysis Based Approaches for Understanding User Context in Web Content", International Conference on Communication Systems and Network Technologies, 2013.

[7] M. Srinath Vijayaragavan, Abhishek Anand, Sundar Vignesh, R Arockia Xavier Annie, "Visualization of big data analysis on social media", Energy Communication Data Analytics and Soft Computing (ICECDS) 2017 International Conference on, 2017.