Demonetization of ₹ 500 and ₹ 1000 currency notes in India: Sentiment Analysis using Twitter

Kushal Kokje
School of Informatics and computing
Indiana University – Bloomington
Bloomington, IN, USA 47408
kkokje@iu.edu

Gowtham Ayna Raveendran
School of Informatics and Computing
Indiana University – Bloomington
Bloomington, IN, USA - 47408
gowtayna@iu.edu

Abstract

This paper presents sentiment analysis strategy and results on a recent event that took place in India where in the prime minister Shri Narendra Modi made an announcement in an unscheduled live television on 8th November 20:15 IST that the rupee 500 and 1000 currency notes would be invalid midnight of that day. The prime minister also announced the issuance of new 500 and 2000 bank notes in exchange for the old bank notes. The decision was met by mixed initial reaction with general audience tweeting their feelings with the hashtag *#demonetization* on Twitter. In this paper we will present the sentiment analysis of the twitter users as a result of the demonization policy enacted by the government of India. We'll classify the tweets as Neutral, Positive and Negative to get aggregate review of the policy and also do Gender Classification and Time – series analysis of the tweets.

1. Introduction

The demonetization of rupee 500 and 1000 currency notes was a policy enacted by the government of India on 8th November, ceasing usage of all 500 and 1000 bank notes of Mahatma Gandhi series as legal tender in India after 9th November 2016. The government claimed that the Demonetization was an effort to stop the counterfeiting of the current bank notes allegedly used for funding terrorism in the country and to also crack down on the Black money * in the country.

*Black money – Illegally earned money on which tax is not paid

The move was also described as an effort to reduce corruption, use of drugs and smuggling of money for terrorism by several media channels. Twitter users reacted immediately to this action with hash tag *#Demonetization*.

With the current Sentiment Analysis and Machine Learning methods, it is possible to computationally identify and categorize the opinions expressed by the Twitter users in order to determine the public orientation of the demonetization policy. This is the motivation for the research described here. The approach discussed can be applied to many other domains like getting to know how people or government react during a natural calamity or any such event.

The first step in the process is to acquire twitter data containing hashtag *#demonetization* using the Twitter API and cleaning of the text data. The second step is to create a word cloud of the most popular words used in the tweets to get an idea of the general opinion of the public. Words like "terrorist", "Black Money" and "Narendra Modi" are the most commonly used in the tweets and hence we created a word cloud with tweets containing those words. It will be interesting to find out the context in which these words are used in the tweet from the word cloud.

Next, in the third steps we'll aggregate the tweets by hour, minute and seconds to do time series analysis of the user's reaction on tweets

after the announcement was made about the Demonetization policy.

In the fourth step we'll explore the tweets with the type of the device or social media platform the tweet originated from. In the fifth step we'll use k-means clustering methods and PCA to better visualize the categorized tweets and find the most common words per cluster. In the sixth step we'll classify the tweets based on gender using NLTK library and plot a distribution of the tweet across "male" and "female" gender. This step largely depends on the username set by the twitter users and hence the results obtained cannot be accurate enough. Lastly we'll do Sentiment Analysis and classify the tweets as Neutral, Positive or Negative.

2. Exploratory Data analysis of the tweets using Tag Cloud

The data contains 6000 tweets extracted at the time of event with hashtag *#demonetization*. There are 6000 rows, one for each tweet, and 14 fields. The major fields that we'll be using in our research are Text, created, screenname, retweet count and StatusSource. We'll do data basic data cleaning of the Text column and remove stop words for better analysis of the tweets.

We have created word cloud of the raw tweets to get the initial notion of the Twitter user, what they thought about the policy and what words the users used to describe the demonetization policy. As we can see in Figure 1 words like "PM" (which stands for Prime minister), "nareddramodi", "terrorist", "kishatwar" (a region in Northern part of India) and "politics" are more widely used by people. The word "support" is also extensively used as seen in the figure.

Figure 1: Demonetization tweets

Next we created word cloud for tweets containing only "narendramodi" word. The word cloud depicts what users think about the policy with reference to Narendra Modi, The Prime Minister of India. We built similar word cloud for word "terrorist"

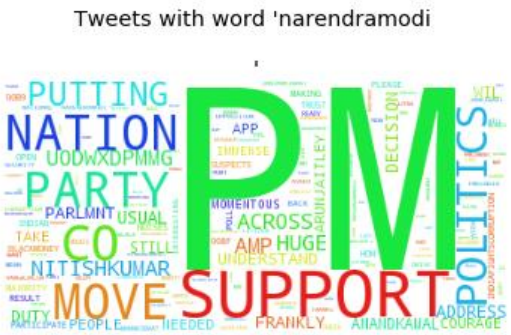

Figure 2: Word cloud "narendramodi"

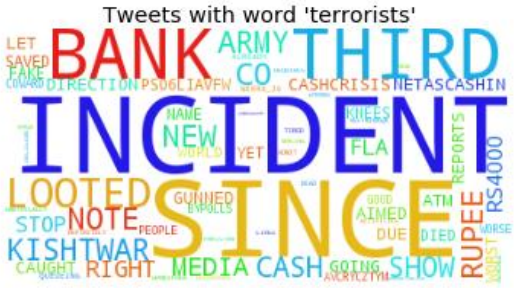

Figure 3: Word Cloud "terrorists"

One of the reasons for implementing the demonetization policy was to curb the black-money problem in India. Black Money is the income illegally obtained or not declared under the income tax by individuals or corporations. The word cloud in Figure 4 provides a view of the words used by Twitter users while tweeting about black money problem

Figure 4: Word Cloud "black money"

3. Time Series Analysis

Further we extracted the hour, minute and seconds information from the "created" field in the tweets data to analyze the Number of tweets per hour and Number of tweets per minute distributions. The statistical measures for the retweet count are as follows.

count	8000
mean	167.26325
std	272.506961
min	0
25%	4
50%	41
75%	221
max	1944

Table 1: Statistical figures for retweet count

Figures 5-8 on the right show time series analysis of the tweets.

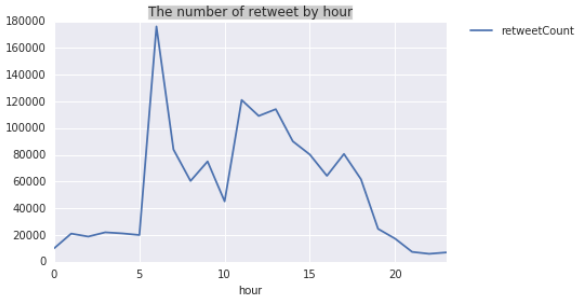

Figure 5: Number of retweet by Hour

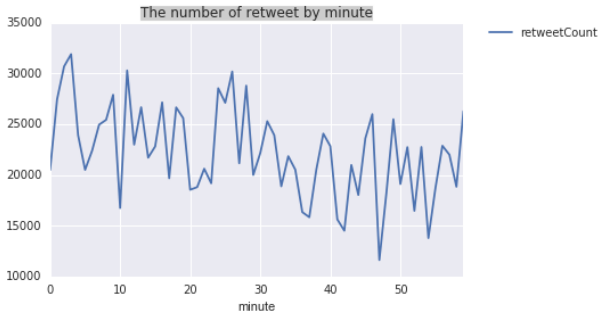

Figure 6: Number of retweet by minute

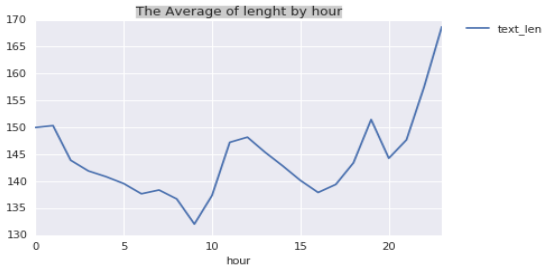

Figure 7: Avg. Tweet length by Hour

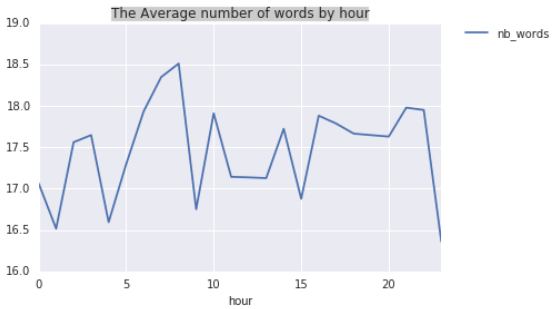

Figure 8: Avg. words in tweet per hour

4. Tweets by Source

In this part we'll classify the tweets based on the device or the platform it originated from. This feature is extracted from field "statusSource" from the tweets. The table shows sample data from the originated platform and the retweet count.

statusSource	Count
Trendinalia Singapore	0
TweetCaster for Android	3426
TweetCaster for iOS	105
TweetDeck	18508
Tweetbot for i<U+039F>S	21
TwitPane Android	138
Twitter Ads	1
Twitter Web Client	163302
Twitter for Android	860969
Twitter for BlackBerry	759
Twitter for BlackBerry®	2
Twitter for Mac	0
Twitter for Windows	9222
Twitter for Windows Phone	11339
Twitter for iPad	21299
Twitter for iPhone	211447
TwixxyBot	24
WordPress.com	0
bitcoinagile	1

Table 2: sample source values and tweet count

Figure 9 is a bar graph of the device/platform the tweet originated from to its count. We can see that a large number of tweets originated from android and IPhone devices and a few from twitter bots and other Twitter API platforms.

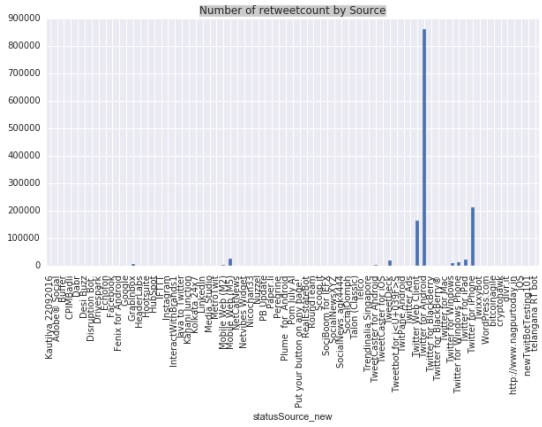

Figure 9: bar graph for number of retweet by source

Further we exclude sources like Twitter payed channels, Twitter bots that are used by paid media and instead concentrate on tweets originated from Android, twitter web client and IPhone devices. Rest of the tweets are tagged as others.

Others	102388
Twitter Web Client	163302
Twitter for Android	860969
Twitter for iPhone	21144

Table 3: Tweets from Android, Iphone and Twitter web client

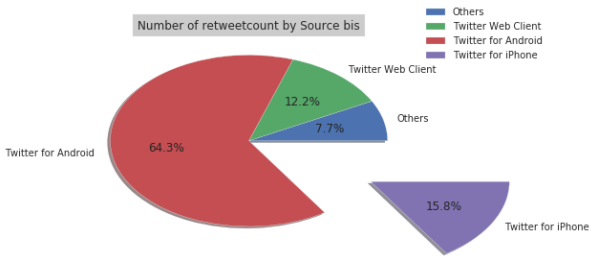

Figure 10: Pie chart for Tweet source

5. K – means clustering

The k-means algorithm creates clusters of data by trying to separate samples in a group of equal variance based on the Euclidean distance measure. The number of clusters needs to be specified by the user or can be derived by experimenting various values of K and choosing the K with the least average distance from the centroid. The feature set needs to be numeric in order to apply K-means as its uses distance to measure how similar objects are. Since we have text data, we will be using bag-of-words approach to create document-term matrix. The total rows in the matrix are the total number of tweets and the columns values are the unique words used in all the tweets text field. We also removed the "english" stop words using the NLTK library before creating the matrix. Next we did TF-IDF transformation on the matrix values which decreases the weights of the most common words that appear in the most of the tweets and hence are less informative.

We executed the K-means algorithm from the python's sklearn library with K = 5. The most commons words per cluster can be found in the below table. Table 4: Cluster key words

Cluster	Words
cluster 0	nationalists calling join benefits https government going gone good got
cluster 1	amp rs lakh bank terrorists looted kishtwar incident modi effect
cluster 2	pm support huge nation question narendramodi edict informed critical fishy
cluster 3	ed bd oscar ad obqrhlnsl mr goes https narendramodi bc
cluster 4	https narendramodi supports party pm politics nitishkumar putting uodwxdpmmg nation

As we can see certain fuzzy words are accumulated in the cluster. The results of the k-means can be further improved by removing such words.

Using PCA we reduced the high dimensional sparse matrix into a 2 dimension for better visualization with little loss of information. Distance measure used was Cosine similarity. Below is the 2-D projection of the five clusters created.

Figure 11: PCA

6. Gender Classification

In this part we classify the tweets based on the gender of the users. We used NLTK names corpora to train a Naïve Bayes classifier. The test set consisted of Screen-Name field from the input tweets data. Through this we got the male-female distribution of the tweets data.

The authenticity of this results cannot be verified as we don't have labeled test data and also people are likely to use a fancy username for a twitter account which can mislead our classifier.

Figure 12 shows Gender distribution for the tweets.

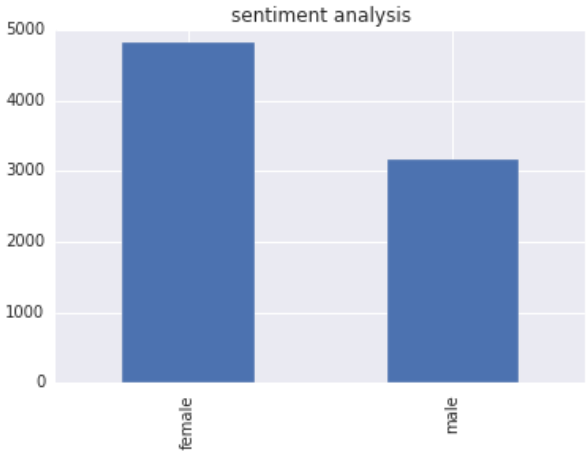

Figure 12: Tweets by Gender

7. Sentiment Analysis

Sentiment analysis was performed using the Vader sentiment analyzer package available in python sklearn library. The tweets were classified as Neutral, Positive or Negative. Through sentiment analysis we get the opinion of the general public, what people think about the demonetization policy. The texts used in the tweets indicate the Positive, Negative or Neutral alignment of an individual towards the policy.

Plotting the results of sentiment analysis, we get the following distribution. It is clear that more people are in support of the Prime Ministers Demonetization policy. 40.3 % of the people express positive sentiment as compared to 30.6 % neutral and 29.1 % as negative.

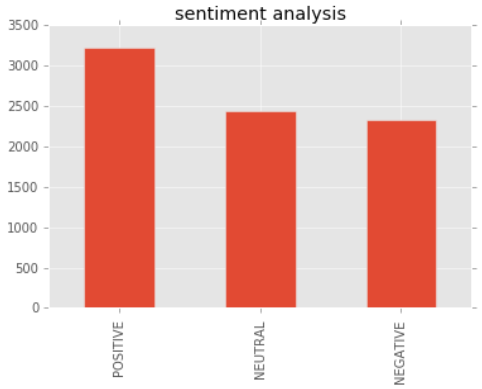

Figure 13: Bar graph sentiment by Tweet

Sentiment Type	Count
NEGATIVE	2331
NEUTRAL	2445
POSITIVE	3224

Table 5: Tweet Count by sentiment

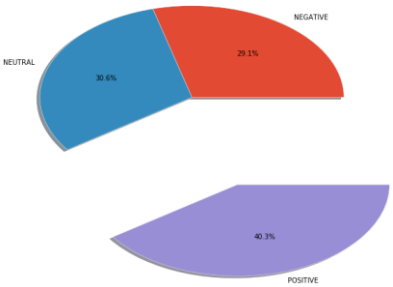

Figure 14: Tweets by sentiment percentage in pie chart.

8. Applications

There are variety of potential applications of using Sentiment Analysis for opinion mining using user's text data. It can be used by retail companies to get the sentiment of the user's product reviews, general sentiment of Forum threads and gauging how people on social media sites like Twitter, Facebook response to an emergency situation or a calamity. How do responses change over time during a crisis event, from initial emergency to support work later? We can also find how people self-organize using micro-blogging as in information network. What kind of information people tweet about depending about the proximity of the crisis? Twitter data can also be used to extract meaning out of the text and construct a model for each user. Modeling users, figuring out personality and ability/interest of a person are many of such techniques that marketing companies use to sell their products. Twitter data

was widely used in predicting the electoral state votes in the 2016 US presidential election.

9. Conclusion

This paper introduces techniques for doing sentiment analysis on recent event using exploratory Data Analysis on twitter data through word clouds. We started with a demonetization word cloud to get a feel of what most people tweeted about the policy. We then subsequently created word cloud only for tweets containing key words like “terrorist”, “narendramodi” and “black money” to know the twitter user’s opinion with reference to these words. We did clustering to separate out clusters and got the top words for each of the clusters. We ended our analysis by sentiment classification of the tweets which gave us an insight into the positive or negative feeling of the users with respect to the demonetization policy.

The limitation of this work includes the size of the data points, more the twitter data, more accurate the analysis would be. The results of the word clouds and K-means can be further improved by data cleaning techniques.

References

Indian 500 and 1000 rupee note demonetization
https://en.wikipedia.org/wiki/Indian_500_and_1000_rupee_note_demonetisation

Python sklearn
http://scikit-learn.org/stable/

Thumbs Up or Thumbs Down? Semantic Orientation Applied to
Unsupervised Classification of Reviews
Peter D. Turney
Institute for Information Technology
National Research Council of Canada
Ottawa, Ontario, Canada, K1A 0R6

NLTK.org
http://www.nltk.org/api/nltk.sentiment.html

Kaggle
https://www.kaggle.com

NLP Stanford
http://nlp.stanford.edu/sentiment/