Title
A Social Media Analysis of India Government that Ceased the Rupee Bank Notes

Author
Sun, Yifan

Publication Date
2018

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

A Social Media Analysis of India Government that Ceased the Rupee Bank Notes

A thesis submitted in partial satisfaction

of the requirements for the degree

Master of Science in Applied Statistics

by

Yifan Sun

2018

ABSTRACT OF THE THESIS

A Social Media Analysis of India Government that Ceased the Rupee Bank Notes

by

Yifan Sun
Master of Science in Applied Statistics
University of California, Los Angeles, 2018
Professor Mark Stephen Handcock, Chair

Nowadays online social media becomes the key factor for people to share and exchange information to the community, those real-time social media platforms such as Facebook, Twitter, and Instagram connect people to the others all over the world. With these platforms, people are able to share their thoughts and feelings to thousands of people, or even more.

In this thesis, I used data from Twitter to analysis people's opinion about the India government's announcement in 2016 for scraping the bank notes of 500 and 1,000 rupee. And try to figure out people's emotions with their words in their tweets. Thus, the people's thoughts could be collected in a more efficient way.

The thesis of Yifan Sun is approved.

Nicolas Christou

Yingnian Wu

Mark Stephen Handcock, Committee Chair

University of California, Los Angeles

2018

To my parents . . .

Thank you so much for your love and support.

I love you.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to thank all the people that discuss the topic on Twitter that provide the data for this thesis.

And I would like to thank my chair Professor Handcock, Professor Ying Nian Wu, and Professor Nicolas Christou. They were always willing to help when I asked them and steered me to the right direction. I enjoyed the class times with them.

Lastly I would like to thank Amandeep Rathee who collected the twitter data of this topic. It gave me the chance to analysis the steps of the India government

CHAPTER 1

Introduction

The Prime Minister of India has announced on 8th November 2016 that the 500 and 1,000 rupee notes " would be stopped from today" and it was unscheduled. The announcement was controversial in India, but it was difficult for people outside the India to know people's opinion about it. So, the tweets that discussed about the announcement were collected and I want to visualize it to make the analysis easier. This should be more precise than the questionnaire since the amount of people in the sample is much more than the usual questionnaire.

As of J. A. Barnes coined the word social network in 1954, many social network platforms came out and have some unbelievable growths. Twitter, the famous social network platform now has more than 336 million global active users. With Twitter, people could spread their latest opinions and information to the worldwide. Thus, in this thesis, I would be able to analysis the data from Indian's perspective with their own words. By analyzing people's tweets word by word, their thoughts would be clear about the supportive or the opposite side.

The main object of the thesis is to reorganize and visualize the data so that people's ideas could be explained and viewed in the simpler way. The data was collected from the tweet tag "demonetization" and over 10,000 tweets were collected after the unexpected live announcement.

After the introductory chapter, in Chapter 2 I reorganize the data to be friendly for the upcoming analysis. In Chapter 3 the data is visualized to be easier to catch the trends of the tweeting. The sentiment analysis is done in Chapter 4, focusing on the text of the tweets. The conclusion chapter discusses the results and summarizes the work briefly.

CHAPTER 2

About the Data

The data has over 14,000 observations, each tweet counts as one observation. It includes data such as the number of times that the tweet was retweeted or not, times the tweet was retweeted, and the device information. After importing the data some entries had to be transfer to numerical format for further analysis.

First we exchange the date and time format to be the general format we use for regular analysis, by taking the Year-Month-Day-Hour-Minute-Second mode. Add the new variable Hour to see the trend of the time that people tweeted. Then substitute the numbers in variables "Retweeted" and "isRetweet" to be 0 or 1 for the true-false analysis. Then in order to analysis the time and devices that people used, we need to reorganized the variables "hour" and "statusSource". Re-size the "hour" variable to be in a data frame and used the "gsub" function to take the data of device source from the "statusSource" variable. The number of the total observations is 14,940.

```
 [1] "X.1"          "X"            "text"         "favorited"    "favoriteCount" "replyToSN"    "created"      "truncated"    "replyToSID"
[10] "id"           "replyToUID"   "statusSource" "screenName"   "retweetCount"  "isRetweet"    "retweeted"    "created_date" "hour"
[19] "isRetweetNum" "retweetedNum" "tweet"
```

Figure 2.1: Summary of the Twitter data — Variables

```
 Min.   :    1   Min.   :    1
 1st Qu.: 3736   1st Qu.: 3736
 Median : 7470   Median : 7470
 Mean   : 7470   Mean   : 7470
 3rd Qu.:11205   3rd Qu.:11205
 Max.   :14940   Max.   :14940


text
 RT @URautelaForever: Dear @evanspiegel \nIndia is so rich that PM Narendra Modi had to implement demonetization to find out who is actually…
: 1200
 RT @gauravcsawant: Rs 40 lakh looted from a bank in Kishtwar in J&amp;K. Third such incident since #demonetization. That's how terrorists have…
:  541
 RT @DrKumarVishwas: And the Oscar goes to "Mr.<U+092D><U+093E><U+0935><U+0941><U+0915>" <ed><U+00A0><U+00BD><ed><U+00B8><U+00A9><ed><U+00A0><U+00BD><ed>
<U+00B8><U+00A5><ed><U+00A0><U+00BD><ed><U+00B8><U+00A2><ed><U+00A0><U+00BD><ed><U+00B8><U+00AD>#demonetization https://t.co/ObQrhlNSL6:  350
 RT @rssurjewala: Critical question: Was PayTM informed about #Demonetization edict by PM? It's clearly fishy and requires full disclosure &amp;…
:  276
 RT @ModiBharosa: Putting Nation over Party Politics #nitishkumar supports PM @narendramodi on #Demonetization https://t.co/UodwXdPMmG
:  253
 RT @ModiBharosa: Huge support for PM @narendramodi 's #demonetization Move Across the Nation \n80-86% people back demonetization: C-voter su…
:  246
 (other)
:12074
 favorited       favoriteCount               replyToSN                   created      truncated         replyToSID              id
 Mode :logical   Min.   :   0.000   narendramodi   :   77   2017-04-18 15:51:06:   45   Mode :logical   Min.   :2.210e+10   Min.   :8.010e+17
 FALSE:14940     1st Qu.:   0.000   PMOIndia       :   21   2017-04-12 16:15:46:   32   FALSE:14243     1st Qu.:8.015e+17   1st Qu.:8.013e+17
                 Median :   0.000   sardesairajdeep:   17   2017-04-19 11:36:07:   22   TRUE :697       Median :8.529e+17   Median :8.015e+17
                 Mean   :   1.071   ArvindKejriwal :   16   2017-04-19 15:08:24:   22                   Mean   :8.384e+17   Mean   :8.256e+17
                 3rd Qu.:   0.000   centerofright  :   13   2016-11-23 06:03:53:   20                   3rd Qu.:8.540e+17   3rd Qu.:8.535e+17
                 Max.   :3166.000   (Other)        :  958   2016-11-23 09:57:50:   14                   Max.   :8.555e+17   Max.   :8.555e+17
                                    NA's           :13838   (Other)            :14785                   NA's   :14054
   replyToUID                                                                                  statusSource          screenName     retweetCount
 Min.   :1.918e+06   <a href="http://twitter.com/download/android" rel="nofollow">Twitter for Android</a>:7642   Stupidosaur  :   47   Min.   :   0.0
 1st Qu.:3.915e+07   <a href="http://twitter.com" rel="nofollow">Twitter Web Client</a>                    :2548   dhruvbhim    :   20   1st Qu.:   1.0
 Median :1.458e+08   <a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>  :2093   mituamin     :   20   Median :  40.0
 Mean   :4.303e+16   <a href="https://about.twitter.com/products/tweetdeck" rel="nofollow">TweetDeck</a> : 492   DipendraDipzo:   19   Mean   : 223.8
 3rd Qu.:1.480e+09   <a href="https://mobile.twitter.com" rel="nofollow">Twitter Lite</a>                  : 263   gauravcam    :   19   3rd Qu.: 197.0
 Max.   :8.543e+17   <a href="https://mobile.twitter.com" rel="nofollow">Mobile Web (M5)</a>              : 178   AmiKanu      :   18   Max.   :5170.0
 NA's   :13838       (Other)                                                                          :1724   (Other)      :14797
 isRetweet       retweeted        created_date            hour            isRetweetNum     retweetedNum     tweet
 Mode :logical   Mode :logical   Min.   :2016-11-22   Length:14940       Min.   :0.0000   Min.   :0     Min.   :1
 FALSE:3948      FALSE:14940     1st Qu.:2016-11-23   Class :character   1st Qu.:0.0000   1st Qu.:0     1st Qu.:1
 TRUE :10992                     Median :2016-11-23   Mode  :character   Median :1.0000   Median :0     Median :1
                                 Mean   :2017-01-28                      Mean   :0.7357   Mean   :0     Mean   :1
                                 3rd Qu.:2017-04-16                      3rd Qu.:1.0000   3rd Qu.:0     3rd Qu.:1
                                 Max.   :2017-04-21                      Max.   :1.0000   Max.   :0     Max.   :1
```

Figure 2.2: Summary of the Twitter data

CHAPTER 3

Data Visualization

After re-format the data in previous chapter, the visualization of the data could start. In this thesis I start by looking at the trend of the tweets in hours. Let the time variable "hour" to be the x-value and the number of total tweets to be the y-value. Would people prefer to retweet more in the first 10 hours? Or prefer to publish original tweets?

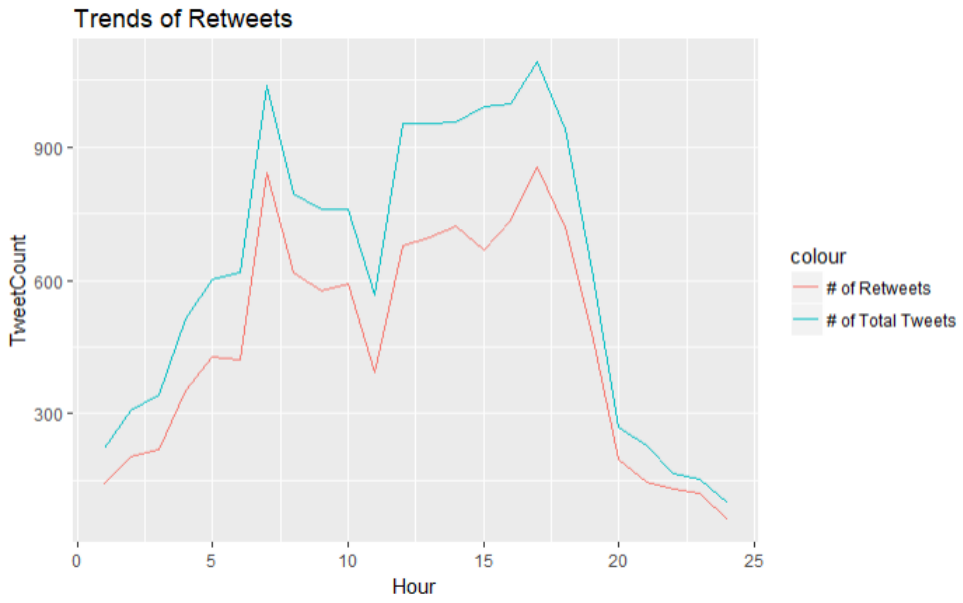


Figure 3.1: Trends of Retweets vs. Total Tweets

The above plot shows that people prefer to retweet for all the time! Those two lines on the plot do not have a lot difference which means that not a lot people send original tweets. People likes to retweet the original tweets that have similar opinions with them, and they would spend less time with tweeting since retweets only requires a few seconds. This is the same as the trend of the social network platform. In the early 21st century, people always

wrote blogs which always contains hundreds of words — unlike Twitter (founded in 2006) which would limit the words for a single tweet. From 6 pm to the late night, people are more willing to retweet. This also applies to the 6 am to 7 am frame. Before we analysis the device sources, I would like to make a guess that this is because people are busier during the rush hours that they do not want to type a lot words — a simple retweet would be more convenient.

Now look at the device sources. To support my previous guess, my expected result would be people use smartphones or portable devices to interact with other Twitter users. Thus Android, iOS, and Windows devices should take the most percentage in the chart. Otherwise, the guess must be false.

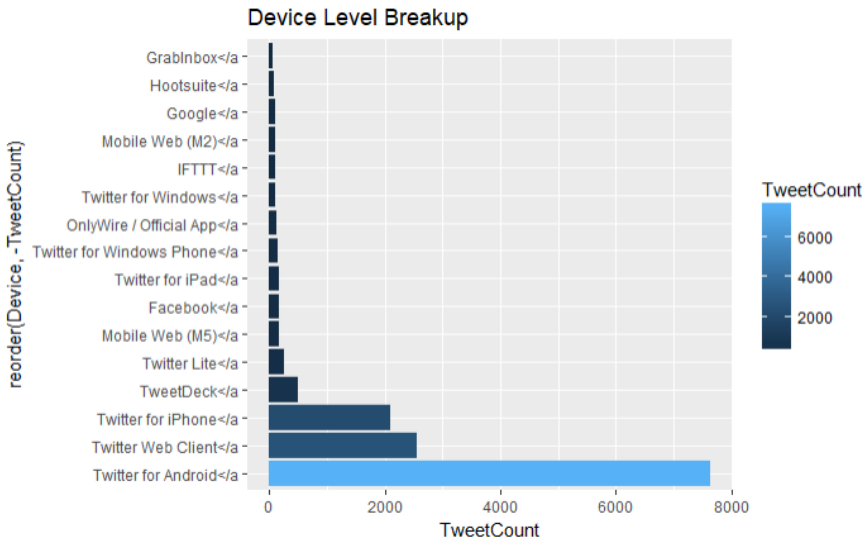


Figure 3.2: Summary of the Device Level Breakup

Table 3.1: Frequency of Device Breakup

Device	TweetCount
Twitter for Android	7642
Twitter Web Client	2548
Twitter for iPhone	2093
TweetDeck	492
Twitter Lite	263
Mobile Web (M5)	178
Facebook	167
Twitter for iPad	165
Twitter for Window Phone	139
OnlyWire / Official App	136
Twitter for Window	110
IFTTT	108
Mobile Web (M2)	105
Google	101
Hootsuite	75
GrabInbox	68

The chart shows that Android devices have the unreachable leading (see Table 3.1 for exact numbers) comparing to the other device sources. And the portable devices are much more than the desktop sources (Web Client and TweetDeck). Although it could not prove that my guess is correct, it somehow support the guess. It also shows that people in India uses Android phones more than the iPhones. These could be used to estimate people's budget for purchasing new phones, since for most cases iPhone would be much more expensive than Android phones. So when we want to advertising to India Twitter users, we should focus more on products that are close to the prices of Android phones, except those luxury level Android phones.

Next step I would group the users who received most retweeted. Are they general Indian famous stars, or politicians? Would people prefer to retweet famous person's tweets? If not, then we could guess that the Twitter users are more willing to retweet those who has opinions from their same sides.

	screenName <fctr>	retweetCount <int>	tweet <dbl>
1141	apoliceshanigm2	7677	2
5166	Krishna20977027	7677	2
7012	ParthPa07241800	5916	13
135	1SunnyElias	5170	1
8020	rayyat9tfoi	5170	1
9692	subhashjsr	5170	1
9942	sxP6DbxfufguCc0	5170	1
8509	sainath_kits	4280	11
8487	SahilBalu456	3772	2
9850	SurenderBalu1	3772	2

1-10 of 10 rows

Figure 3.3: Tweets that Got Most Retweets

Searching for these users who received top retweets, most of them are not famous persons that has the Twitter verified sign. That could be use to support the ideas that people on the social media platform would be more often to express their real ideas, while in the real-life they might be influenced by the environment they live in. This is an interesting result that when people are on the internet, since their information would not be provided or they do not have to talk to other people face to face, people are more courage to speak out the thoughts from their own perspectives.

Then check the users who received most replies. Would that be the same person who received the most retweets? The answer is no. Which means that those original tweets that received a lot retweets are not controversial. When people agree with the idea in the tweet, they would prefer to retweet rather than leave the reply.

```
##                User RepliesReceived
## 369    narendramodi              77
## 421         PMOIndia              21
## 518 sardesairajdeep              17
## 66   ArvindKejriwal              16
## 114   centerofright              13
## 377            ndtv              12
## 49         ANI_news              10
## 608    timesofindia              10
## 125       CNNnews18               9
## 180     evanspiegel               9
## 558     Stupidosaur               9
## 264        jamewils               7
## 41        AmmU_MaanU               6
## 99         BJP4India               6
## 276 John_Miller_GLR               6
## 313       madmanweb               6
## 370  NarendraMod198               6
## 63      arunjaitley               5
## 153    digvijaya_28               5
## 161             dna               5
```

Figure 3.4: Tweets that Got Most Replies

CHAPTER 4

Sentiment Analysis

After the work in Chapter 3, the data provides more and more information comparing to the raw Twitter data. Now it is the time to run the sentiment analysis. First, I took out the words that are not wanted from the text, such as numbers, punctuation, and white space. Then run the Rsentiment package in R to detect if the word is positive, negative, or neutral (The package would rate sentiment of the sentences). After creates new variables to store the words and summarize the frequencies of each words, use the wordcloud package in R to draw the wordcloud for positive and negative words — to understand people's thought about the government's announcement from both sides.

Figure 4.1: Worldcloud of the Positive Sentiment

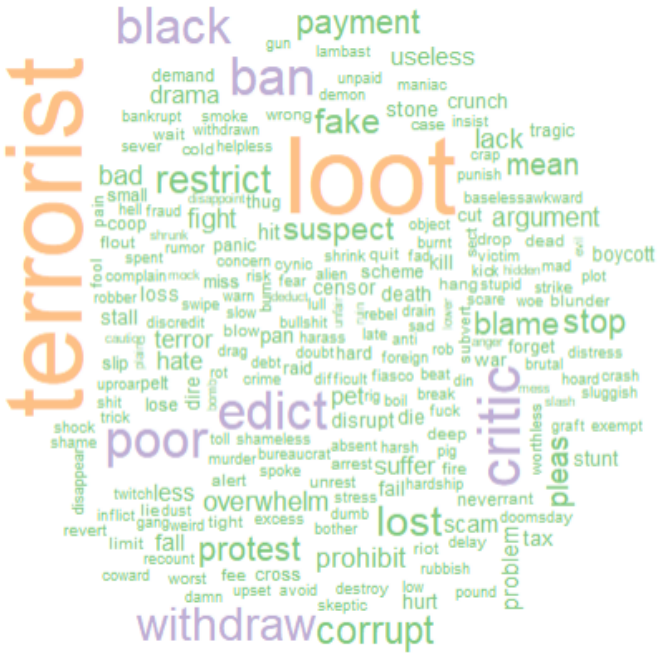


Figure 4.2: Worldcloud of the Negative Sentiment

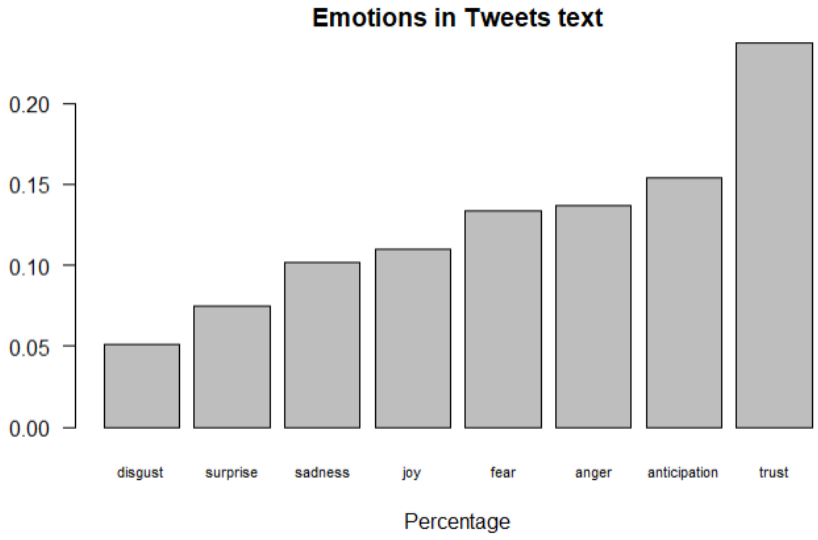
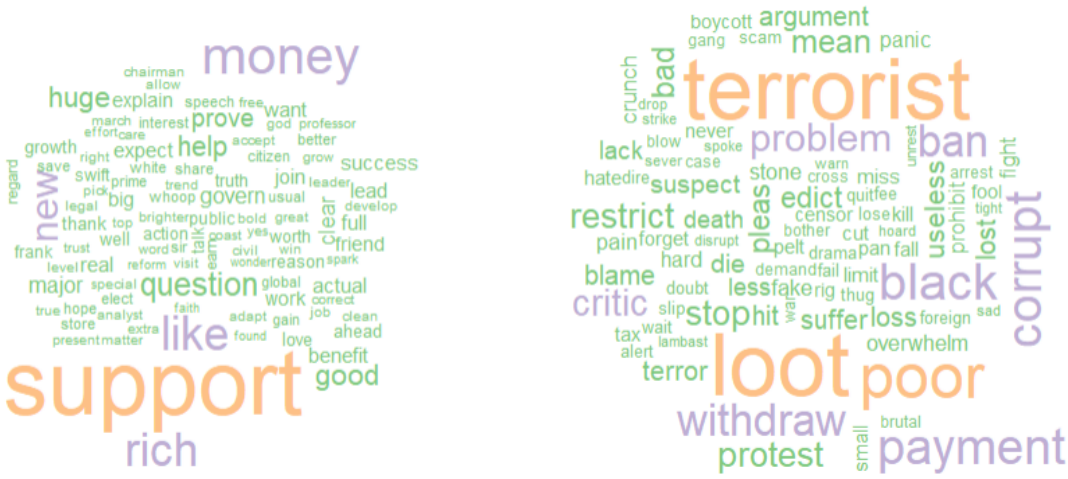


Figure 4.3: Emotions in the Twitter Texts

Besides the wordclouds, I used another sentiment package which is called syuzhet in R to focus on the eight words that express people's emotions (The eight words are selected by the NRC sentiment dictionary[MT10]). The most frequent word is "trust". And for the other

words, only "joy" could be considered as a positive emotion. Thus most people is reflecting some negative emotions with the announcement since the percentage of trust and joy is less than the others.

This research could be expanded as we go further with the device sources. What would iPhone customer think comparing to Android phones users? The iPhone has a very specific price that is higher than most of the Android phones. So I splitted the people into different groups by their device sources, and see if the groups would have the same or different opinions based on different perspective. I took the Android group and iPhone group to make the comparison by creating the wordcloud graphs in groups, and the Web group to see the difference between portable device users and computer users.



(a) Positive Sentiment – Android

(b) Negative Sentiment – Android

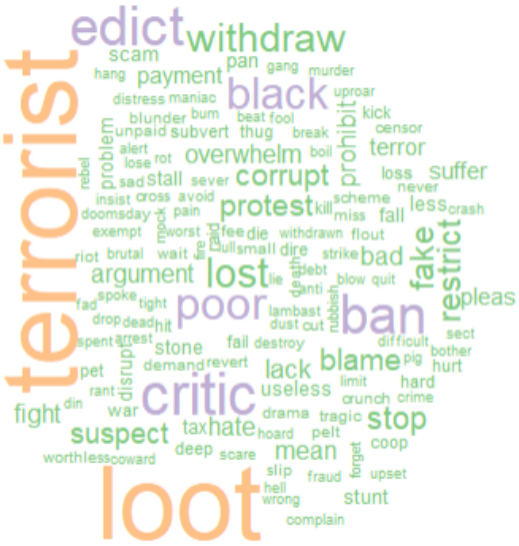
Figure 4.4: Wordclouds of Android user

(a) Positive Sentiment – iPhone

(b) Negative Sentiment – iPhone

Figure 4.5: Wordclouds of iPhone user

(a) Positive Sentiment – Web Client



(b) Negative Sentiment – Web Client

Figure 4.6: Wordclouds of Web Client

The wordcloud figures did not show a lot difference between Android users and iPhone users. People's consuming preference did not effect their opinions about the bank notes

suspension. The figure for Web Client also shows that when people were not using portable devices, they would still have the similar ideas.

CHAPTER 5

Conclusion

In this thesis, I run the data of 14,940 Twitter tweets that are collected under the tag of India's "demonetization". For many times I have felt that I was able to speak more when I was "online". Now the world's trend is to make all information "online" : the news, the music, the TV dramas, and so on. It is good that we could receive information from the other countries easier and faster (and cheaper).

The idea about this thesis is to analysis people's opinion about the country government's decision by matching their tweets. And I have also used the data to get some additional information such as the possible purchase power of Twitter users by grouping their devices. With the part visualizing the tweeting hours of the users, I found out that the peak occurs when people are mostly on their way to work or home. These users tend to retweet what they have the same opinions instead of retweeting famous people's tweets.

The 2016 India demonetization is a controversial topic. Most Indians were really shocked by the unscheduled announcement and they discussed a lot. In the result it shows that most people do not think it was a good news and they became afraid of the subsequent result. Coming to the 2018, we could see that people's fear is reasonable. The Government did not meet their expectation with the new 500 and 2,000 rupee bank notes. But it is a great example showing that social media platform is turning to become reliable data source for the analysis. It is an upgrade for the traditional questionnaire while people all talk about what they really want to say instead of simply answering specific questions.

5.0.1 Future Thought

And for the people who had positive or negative opinions, it would be interest to find their characteristics for further studies. However, the tweets did not provide information such as sex, age, or location although some users would present these on personal sites and these would not be able to collect when we were looking for tweets under the tag. This is the disadvantage when we are comparing with traditional questionnaire that it could focus on specific groups or ask people to provide some information to help the characteristics process. In order to have some further studies with this topic, I think that the data would need to have some more variables that have to be collected in other ways.

REFERENCES

[MT10] Saif M. Mohammad and Peter D. Turney. "Emotions Evoked by Common Words and Phrases: Using Mechanical Turk to Create an Emotion Lexicon." In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, CAAGET '10, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.