TEXT ANALYTICS ASSIGNMENT

Topic : Sentiment Analysis on Demonetization twitter dataset

Dataset Chosen :The dataset (containing 14940 tweets) was downloaded from Kaggle.

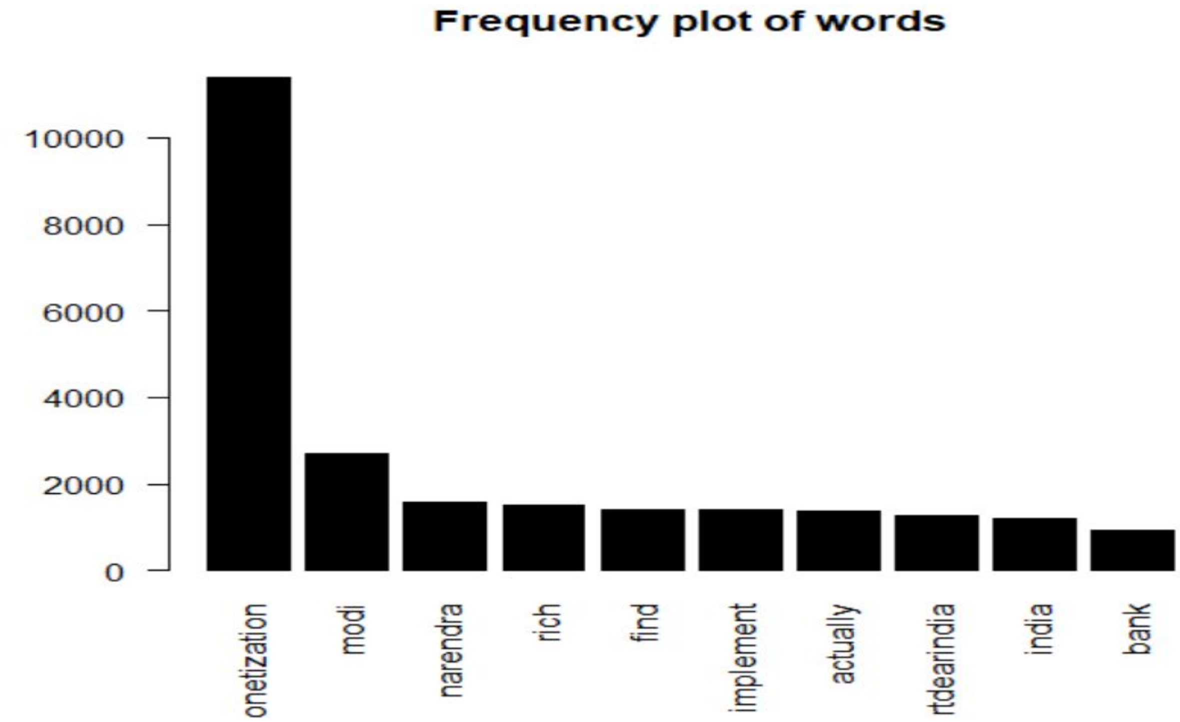
In this assignment, a detailed sentimental analysis is performed on demonetization, The demonetization of Rs.500 and Rs.1000 banknotes was a step taken by the Government of India on 8 November 2016, ceasing the usage of all Rs.500 and Rs.1000 banknotes of the Mahatma Gandhi Series as a form of legal tender in India from 9 November 2016. The data contains 14940 tweets on #demonetization.
The main motivation behind the analysis is to know the opinion/impact of demonetization.

Steps followed :

1. Import the dataset into R.
2. The dataset used here consists of different columns from which tweets alone should be considered
3. Followed by cleaning the data.Analysing the unique tweets by removing all the retweets
4. Punctuations, HTML links, people names(user name) are further removed to produce a perfect dataset for analysis.
5. Now,For every tweet (every row) a score is assigned based upon the text present in the specified tweet.
6. Visualize the result using ggplot2 package,A bar plot is seen with sentiment scores on X-axes, tweet counts on Y-axes.

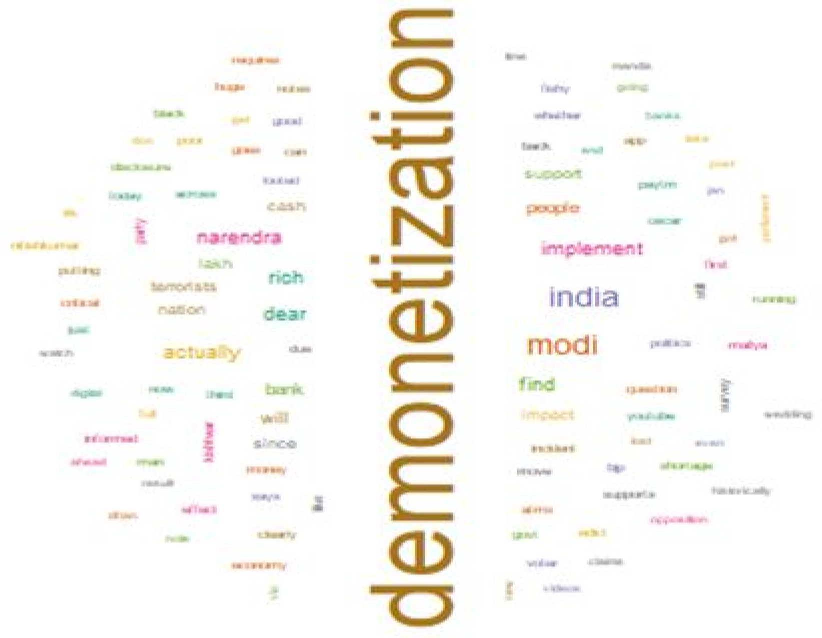
Results/Observations :

1. FREQUENCY PLOT OF WORDS - After the data was loaded into R, we have performed text mining and obtained a cleaned data. Thereafter, we generated the corpus and counted the word frequency. According to the frequency of words we plotted the words.

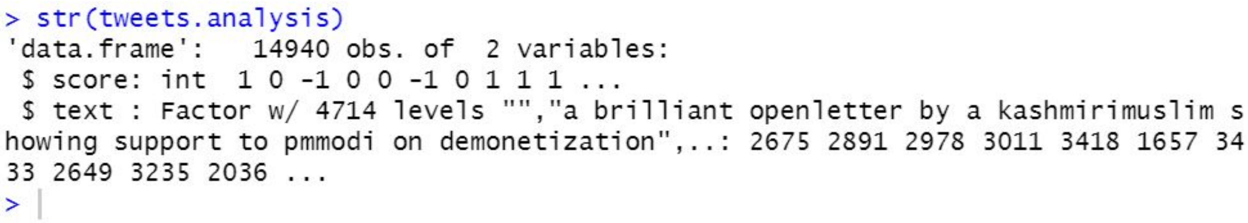


Here, “demonetization” has the frequency in the tweets.

2. WORD CLOUD - Another interesting plot of the frequency of words done is word cloud. The one with higher frequency appears to be in a large size and then in accordance to the frequency, it keeps on decreasing.

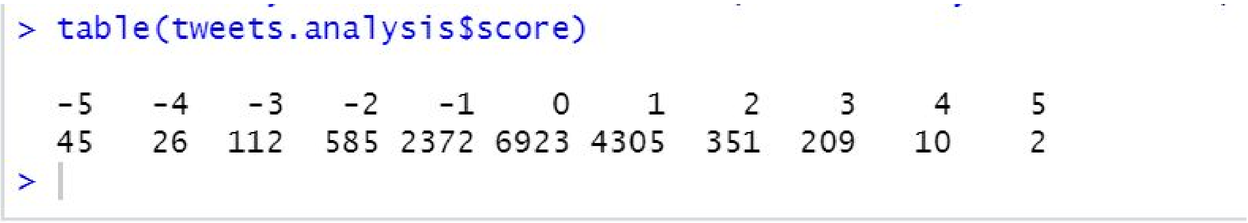


3. ANALYSIS -
    - Then we compared the words to the dictionaries of positive & negative terms. Here, match() returns the position of the matched term or NA, however we just want a TRUE/FALSE, and conveniently enough, TRUE/FALSE will be treated as 1/0 by sum().



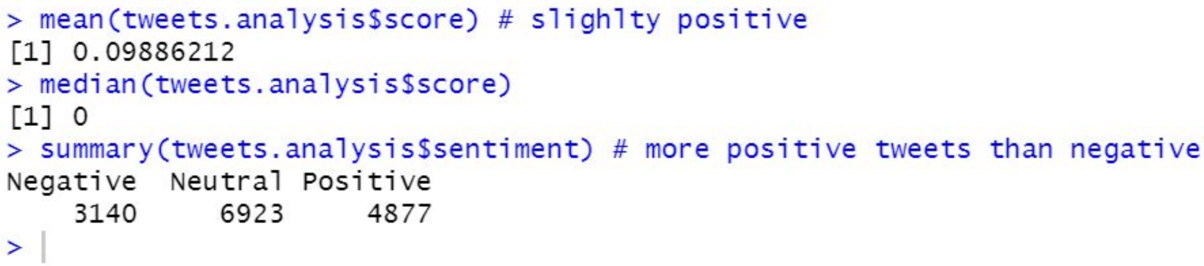
```
> str(tweets.analysis)
'data.frame':   14940 obs. of  2 variables:
 $ score: int  1 0 -1 0 0 -1 0 1 1 1 ...
 $ text : Factor w/ 4714 levels "","a brilliant openletter by a kashmirimuslim s
howing support to pmmodi on demonetization",..: 2675 2891 2978 3011 3418 1657 34
33 2649 3235 2036 ...
> 
```

    - Created the component that corresponds to the sentiments based on scores and made a table based on score and tweet counts



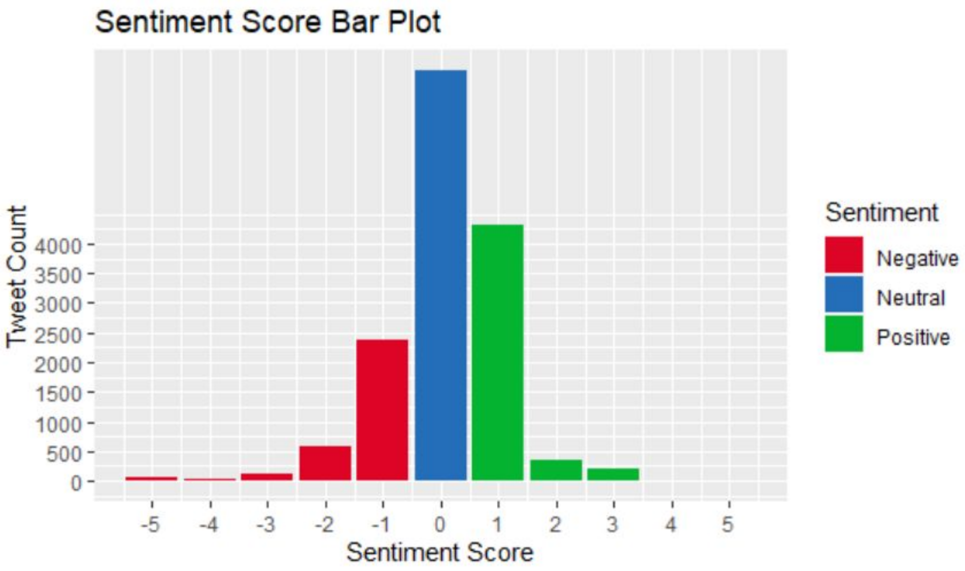
```
> table(tweets.analysis$score)

  -5   -4   -3   -2   -1    0    1    2    3    4    5 
  45   26  112  585 2372 6923 4305  351  209   10    2 
> 
```

    - Next, we performed the basic statistics



```
> mean(tweets.analysis$score) # slighlty positive
[1] 0.09886212
> median(tweets.analysis$score)
[1] 0
> summary(tweets.analysis$sentiment) # more positive tweets than negative
Negative  Neutral Positive 
    3140     6923     4877 
> 
```

    - Based on the sentiment scores calculated, it was clear that most tweets were positive in nature. Here is a comprehensive graph of it

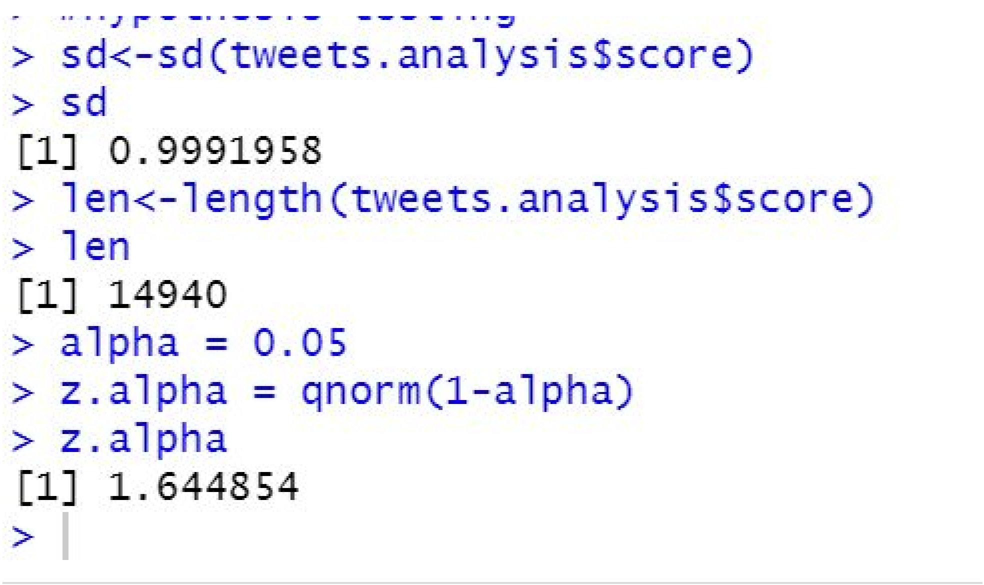


4. HYPOTHESIS TESTING -

Let,

H0: Demonetization had majority popular support.

H1 (Alternative Hypothesis):Demonetization does not have majority popular support.



```
> #hypothesis testing
> sd<-sd(tweets.analysis$score)
> sd
[1] 0.9991958
> len<-length(tweets.analysis$score)
> len
[1] 14940
> alpha = 0.05
> z.alpha = qnorm(1-alpha)
> z.alpha
[1] 1.644854
>
```

At alpha = 0.05 the calculated-value < tabulated-value i.e. 1.644 < 1.96.Hence null hypothesis is accepted.Hence, demonetization had majority popular support.

Conclusion : The conundrum of demonetisation’s effect on the general public was solved easily using Twitter as a sample space and with the use of R Studio. While a hot button issue like the one above may require more data for a conclusive result, it is proved that positive sentiment score is more compared to the other sentiment scores, concluding that support for demonetization is high.But there are also people who feels negative about it.