**Sentiment Analysis of Tweets**

**#Demonetization**

Group 1

Jigyasa Kohli, Keka Nandi, John Cox, Saurabh Jadhav.

New Jersey Institute of Technology

IS 688

**Abstract**

*This paper**exemplifies**the use of sentiment analysis as it relates to the recent Demonetization that took place in India. By utilizing common analysis packages of R programming language and algorithms, we will demonstrate how sentiment analysis can be applied to Tweets and how it can ultimately be used to gain meaningful insights through visualizations and plots i.e. Opinion Mining. Methods will be discussed describing how the data set was extracted, collected, cleaned, and finally interpreted. Throughout this paper, we will describe the process used. Finally, we will determine whether the Demonetization was a good or bad decision.*

**INTRODUCTION**

Demonetization means the withdrawal of a certain form of currency (such as gold coins, currency notes) from circulation in an economy. The Government of India recently took a bold step to demonetize 500 & 1000 INR currency notes and the use of same were declared invalid from a specified date.

Most people hailed the Prime Minister Narendra Modi’s strong decision to fight black money, tax payment issues and other illegal activities but at the same time many poor and middle-class Indians criticized the same who had lost their jobs or did not receive daily wages. Our aim is to analyze these mixed emotions that flooded the social media and understand how people perceived the sudden decision of the Government.

**APPROACH**

We follow the approach in the following order:

* Data Loading
* Data Cleaning
* Sentiment Analysis
* Sentiment Analysis Visualization

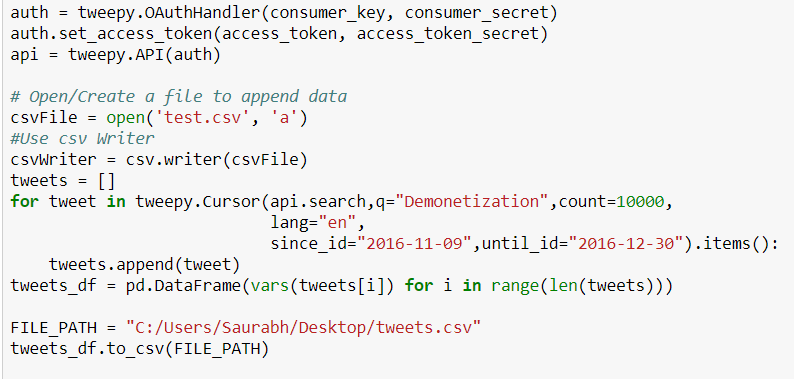
Before diving deep into the approach, let’s understand basics about the coding environment, data set, it’s attributes and the process of creation of the data set.

* **Business Understanding:**

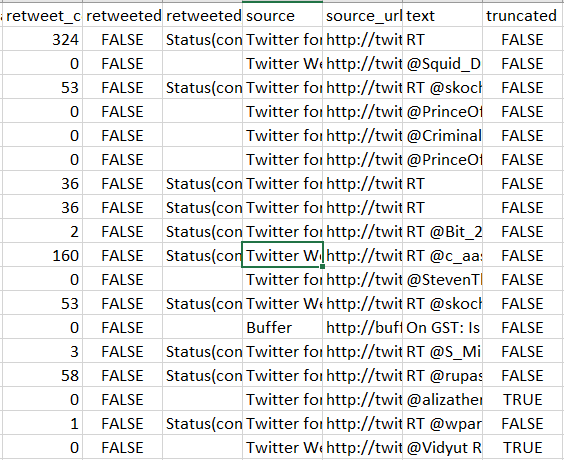
We are using RStudio environment with in-built packages like twitterR, sentiment (external package), ggplot2.

* **Data Understanding:**

The Demonetization data set consist of 10,000 columns and 14 actual interpretable rows with data. The data was extracted from the Twitter by using Python Script. After detailed discussion with group members and immense R&D, we found out that it’s more efficient to use Python instead of R while extracting the tweets. Below is the code used to gather the data.



To determine how to perform sentiment analysis, let’s first look at the few of the attributes and their values. Due to time constraints, no preprocessing was performed on the data set, but were directly done after loading the data set into RStudio environment which we will later in the report.



* **Data Preparation:**

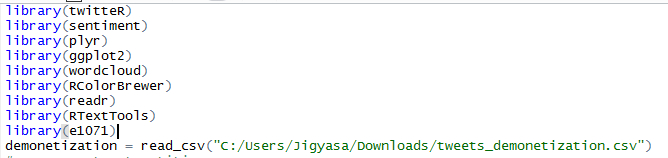
Before applying any function or algorithm, the dataset is checked for NaN values. The Data Preprocessing step involves rigorous methods which can be applied to validate the data i.e. remove the missing values. Bar chart and word-clouds are plotted to visualize the interactions between selected variables to find hidden patterns. The data looks like the table below.

**Approach:**

**Step 1: Importing the Data**

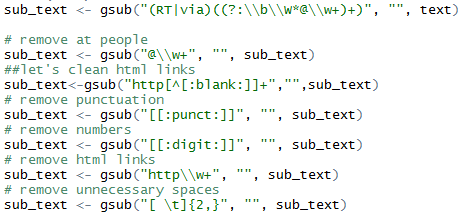
The first step is to import the data along with the inbuilt functions and packages. To import the sentiment package, we had to download the sentiment package and explicitly provide the link to it which was:

install.packages("C:/sentiment\_0.2.tar.gz", repos = NULL, type="source")

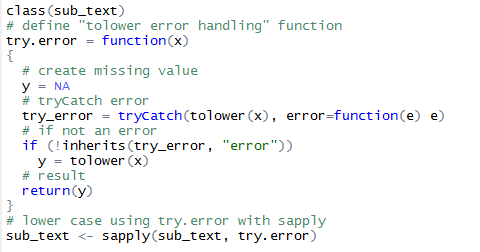
****

**Step 2: Data Preprocessing/Cleaning**

Since, the data set was extracted, it was raw and needed preprocessing. We followed the universal approach of cleaning data in R. Regular Expressions, html links, punctuations, numbers and unnecessary spaces removal constituted part of cleaning process. The code below provides a knowledgeable insight into this process.



The final step was to convert all the text to lower case and removing the NA values.

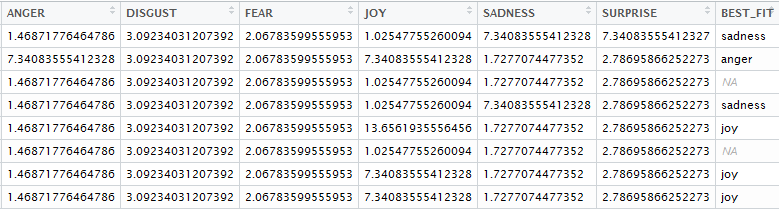
****

**A screenshot of a cell phone

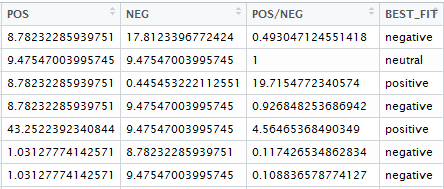
Description generated with very high confidence**

**Step 3: Sentiment Analysis**

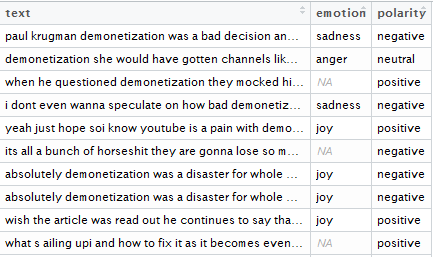
The data is now ready to be analyzed. In order to do sentiment analysis, we must know the emotions associated with the texts. To do so, we used the classify\_emotion function which was imported with Naïve Bayes as an external package. Naïve Bayes algorithm is used here to classify the polarity of the texts. Applying this algorithm resulted into creation of a very large matrix with different emotions.



The one thing we did here which we like to mention is replacing the NA emotions by Unknown emotion. Later, the same algorithm was used to check the polarity (i.e. positive or negative) and decide the best fit by dividing the positive value by negative one.

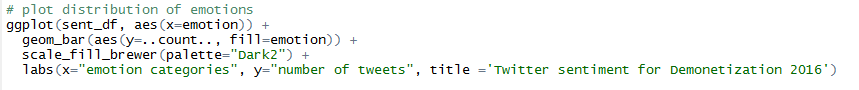


Finally, for better readability, each tweet was assigned its emotion and polarity as per above calculations and then inserted into a data-frame.

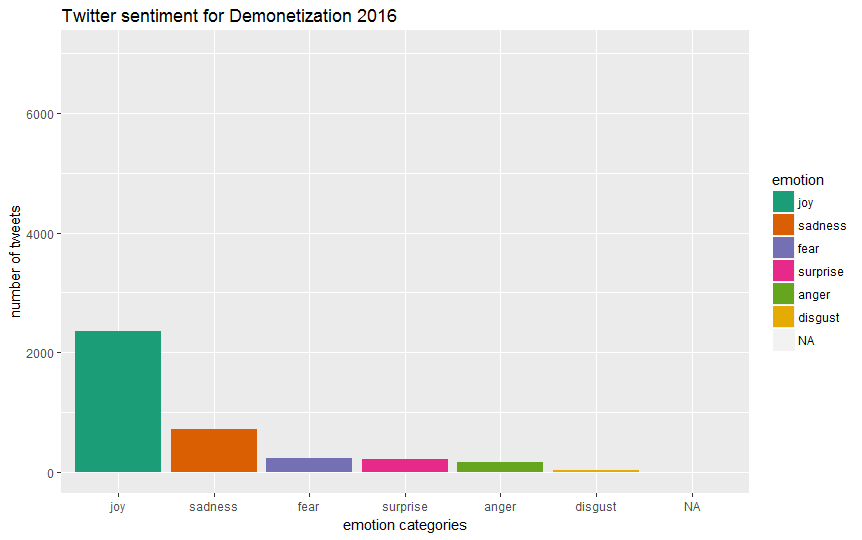


**Step 4: Visualizations:**

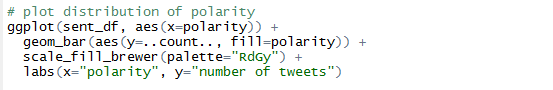
The above screenshots and calculations won’t make any sense to a normal person. So, we made sure that our analysis is understood even by a non-technical person and the only way to do so is visualize the results. Once our tweets were associated with their respective emotions, we plotted a bar chart depicting the total number of tweets belonging to a particular emotion category in descending order.

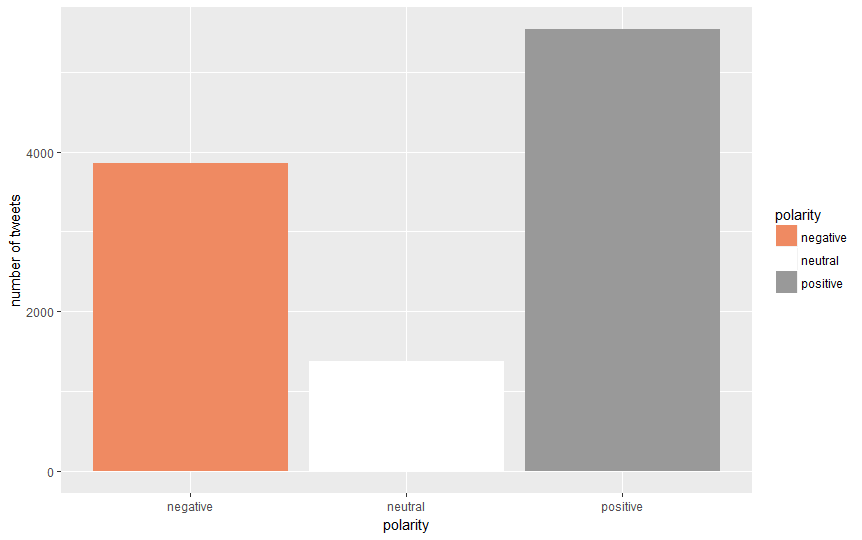


The above code resulted into following visualization:

****

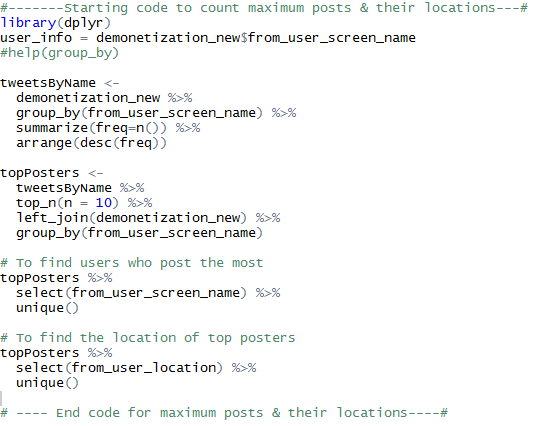
The graph explicitly describes that majority people were happy with the decision of Demonetization as compared to other emotions. To make this fact more solid and look convincible, we went further and plotted the bar chart of polarity of tweets and the results were quite assuring. The below screenshot is the code while the later one is the output.

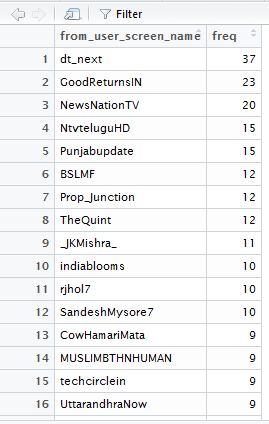
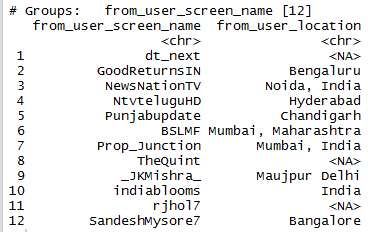


****

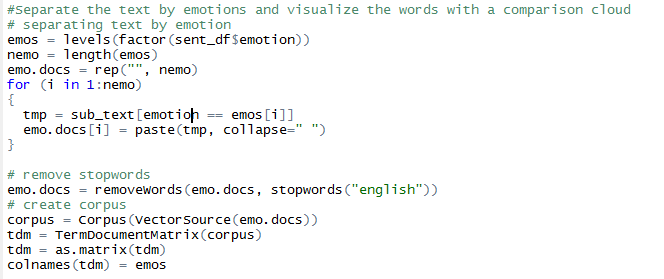
The graph depicted a crystal-clear conclusion that people were supporting Demonetization while a considerable amount of people showed negative response. Of course, the neutral polarity should also be considered. But, in this case it is less as compared to the positive and negative polarity.

To gain some more insights, we evaluated the dataset based on the author of the tweets and their geo locations. We found that some users posted more frequently than others and their locations which gave us an idea that users from the metropolitan cities are more aware and eloquent about their opinion regarding recent developments in the country.



Further, we thought of diving more into the emotions and separating it and visualizing the most commonly typed words by using word-cloud visualization. The initial step in implementing word cloud visual was to create TDM and to do so, we needed to remove all the stopwords from the tweets text.

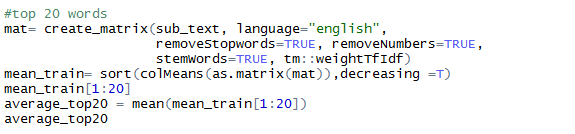


Our above efforts resulted into something like this below:

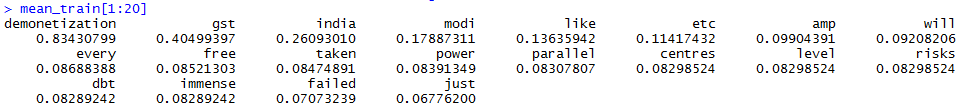
A close up of text on a white background

Description generated with high confidence

Since the above visual looks little cluttered we thought of displaying the top 20 words that were used.



The frequency of each word is displayed as follows:



**CONCLUSION:**

The main motive of the project was to understand the sentiment of Twitter population on the topic of Demonetization. In process of doing so, we had the opportunity to implement various concepts on a real data. As we can see from the above analysis, majority of twitter population bolstered the Demonetization resolution. However, these were the sentiments of only Twitter population within India. If another data source is considered with broader geographical aspect, the results may vary ranging from small to relatively large scale.

**CHALLENGES:**

The sentiment analysis is not a perfect science since the human language, on which we train our algorithms, is complex. Teaching a machine to understand the tone, grammatical nuances, slangs, misspellings, social media lingos is quite difficult due to its huge variations. Again, all expressions cannot be strictly classified as positive, negative or neutral. Thus, it is quite prone to errors and would need a human eye to watch over it.

However, with large datasets, like the one that is used above, the insights gained will overshadow the concerns at granular level and the focus will be to make the results more interpretable and actionable.

**FUTURE WORK:**

Implementation of classification algorithms and prediction is a part of the future scope. On the other hand, handling negation to improve the classification accuracy is also very important. Also, analysis of tweets may lead to loss of sentimental information. To minimize the loss, we must understand how the different polarities are being mixed to give the final result i.e. we would want to adopt algorithms different that Naïve Bayes to do so.

**References**

* [Www.kaggle.com](http://Www.kaggle.com)
* <https://en.wikipedia.org/wiki/Sentiment_analysis>
* <https://www.lexalytics.com/technology/sentiment>
* <https://www.brandwatch.com/blog/understanding-sentiment-analysis/>