Aryabhatta Knowledge University

Patna, Bihar, 800001

2020 - 2021

A Project Report On

"Sentiment Analysis Using Machine Learning In R"
(on Demonetization in India based on opinions expressed via Twitter)

Submitted in partial fulfillment of the requirements for the award of degree.

Submitted By :-

ANSHU KUMAR SINGH	**17105108012**
SANU KUMAR	**17105108015**
MD AMAN	**17105108019**
KUMAR ANUBHAV	**17105108039**

B.Tech In Computer Science and Engineering

Bhagalpur College Of Engineering, Bhagalpur

Under the esteemed guidance of

Mr. GOVIND KUMAR JHA(Assist. Professor)

Department of CSE

Bhagalpur College Of Engineering, Bhagalpur

Aryabhatta Knowledge University

Patna, Bihar, 800001

Department Of Computer Science And Engineering,

Bhagalpur College Of Engineering, Bhagalpur

CERTIFICATE

This is to clarify that the project entitled **"Sentiment Analysis Using Machine Learning In R"** *has been successfully carried out by* **ANSHU KUMAR SINGH (17105108012), SANU KUMAR (17105108015), MD AMAN (17105108019) , KUMAR ANUBHAV (17105108039)** and *in partial fulfillment of the requirements for the award of degree of* ***Bachelor of Technology*** *in* ***Computer Science and Engineering*** *of* ***Aryabhatta Knowledge University, Patna,*** *during the academic year 2020-2021. It is certified that all corrections/suggestions indicated for Internal Assessment have been incorporated in a Report deposited in the departmental library. This project has been approved as it satisfies the academic requirements in respect of Project Work prescribed for the Bachelor of Technology Degree.*

Project Guide

Mr. GOVIND KR JHA
Prof.,Dept. Of CSE
Bhagalpur

Prof. Raj anwit

Head Dept. Of CSE Asst.
B.C.E Bhagalpur, B.C.E

Name of the Examiners **Signature with Date**

1._____ _____

2._____ _____

Aryabhatta Knowledge University

Patna, Bihar, 800001

Department Of Computer Science And Engineering,

Bhagalpur College Of Engineering, Bhagalpur

DECLARATION

We, ***ANSHU KUMAR SINGH (17105108012), SANU KUMAR (17105108015)MD AMAN (17105108019)*** and ***KUMAR ANUBHAV (17105108039)*** *and* Students of seventh semester Bachelor of Engineering, in the Department of Computer Science and Engineering, *Bhagalpur College of Engineering, Bhagalpur* declare that the project entitled ***"Sentiment Analysis Using Machine Learning In R"*** has been carried out by us and submitted in partial fulfillment of the course requirements for the award of degree in **Bachelor of Technology** in **Computer Science and Engineering** of **Aryabhatta Knowledge University, Patna** during the academic year **2020-2021.** The matter embodied in this report has not been submitted to any other university or institution for the award of any other degree or diploma.

ANSHU KUMAR SINGH (17105108012)

SANU KUMAR (17105108015)

MD AMAN (17105108019)

KUMAR ANUBHAV (17105108039)

Preface

This report documents the work done during the project in Bhagalpur College of Engineering, Bhagalpur under the supervision of Mr. Govind Kumar Jha.

The report first shall give an overview of Sentiment Analysis using machine learning in R taking data sets(containing 8000 tweets) on Demonetization in India based on opinions expressed via Twitter, as provided by kaggle.

Report shall also elaborate on the future works which can be persuaded as an advancement of the current work.

Acknowledgement

Our Project is the result of the encouragement of many people who helped us and provided the feedback, directions and valuable support. It is great pleasure for us to acknowledge their contributions for the successful completion of the project.

I would like to thank my internal guide **Mr. GOVIND KUMAR JHA,** Department of Computer Science and Engineering, BCE Bhagalpur, for the constant help and support extended towards us during the course of the project. It is with deep sense of gratitude and respect that I express my most cordial and humble thanks to our beloved HOD **Prof. Raj Anwit,** for his valuable and inspiring guidance and support. I express my gratitude to our Principal, **Dr. Pushpalata** who has always been a great source of inspiration.

I take this opportunity to express my sincere thanks to all the staff of BCE Bhagalpur for their cooperation and support. I would like to thank my **Family** and **Friends** who always supported and motivated me during the project. Special thanks to my **Parents** for their unending support, love and patience.

ANSHU KUMAR SINGH (17105108012)

SANU KUMAR (17105108015)

MD AMAN (17105108019)

KUMAR ANUBHAV (17105108039)

CONTENTS

INTRODUCTION

CONTEXT

Sentiment analysis (also called as opinion mining) is the task of identifying whether the opinion expressed in a text is positive or negative in general about a particular topic or context. Prime Minister of India, Narendra Modi announced the demonetization in an unscheduled live televised address at 20:00 Indian Standard Time (IST) on 8 November. In the announcement, Modi declared that use of all ₹500 and ₹1000 banknotes of the Mahatma Gandhi Series would be invalid past midnight, and announced the issuance of new ₹500 and ₹2000 banknotes of the Mahatma Gandhi New Series in exchange for the old banknotes. This was received by the masses with mixed feelings. This project aimed at structuring and analyzing those feelings, especially the ones that were expressed on Twitter. Due to this large amount of user base we hope to achieve a reflection of public sentiment by analyzing the sentiments expressed in the tweets.

WHY ANALYTICS?

1. What is trending positively/negatively over a period of time and why?
2. Who is being talked about, where, and why?
3. What college is being talked about?
4. What topics are being discussed the most?
5. Who is being talked about most positively?
6. What are the best sources for positive exposure?
7. What is the geographic location of the comments?

WHY VISUALIZATION?

Data visualization is the presentation of data in a pictorial or graphical format. It enables decision makers to see analytics presented visually, so they can grasp difficult concepts or identify new patterns. With interactive visualization,
you can take the concept a step further by using technology to drill down into charts and graphs for more detail, interactively changing what data you see and how it’s processed. Tables, barplots, timelines, word clouds, histograms and pie charts can be used for visualization. End Result: Informed Strategies, Improved Performance.

WHAT IS TWITTER SENTIMENT ANALYSIS?

Twitter is an online news and social networking service that enables users to send and read short 140-character messages called "tweets". Registered users can read and post tweets, but those who are unregistered can only read them. Hence Twitter is a public platform with a mine of public opinion of people all over the world and of all age categories. As of the third quarter of 2020, Twitter averaged 330 million monthly active users. Twitter Sentiment Analysis is the process of determining the emotional tone behind a series of words, used to gain an understanding of the attitudes, opinions and emotions expressed within an online mention.

WHY TWITTER SENTIMENT ANALYSIS?

The applications for sentiment analysis are endless. It is extremely useful in social media monitoring as it allows us to gain an overview of the wider public opinion behind certain topics However, it is also practical for use in business analytics and situations in which text needs to be analyzed. Sentiment analysis is in demand because of its efficiency. Thousands of text documents can be processed for sentiment in seconds, compared to the hours it would take a team of people to manually complete. Because it is so efficient (and accurate – Semantria has 80% accuracy for English content) many businesses are adopting text and sentiment analysis and incorporating it into their processes.

Applications:

The applications of sentiment analysis are broad and powerful. Shifts in sentiment on social media have been shown to correlate with shifts in the stock market.

The ability to quickly understand consumer attitudes and react accordingly is something that Expedia Canada took advantage of when they noticed that there was a steady increase in negative feedback to the music used in one of their television advertisements.

For example, Less than 24 hours earlier, Elon Musk appeared to prompt shares in CD Projekt, which makes the Cyberpunk 2077 computer game, to surge more than 12% after he said via Twitter that a new model of Tesla's Model S Plaid car would allow passengers to play the game.

OVERVIEW

The sentiment (score/mood) of each tweet was analyzed using ‘bag of words’ method. In this method, the sentences/text is broken down into words which are treated as a unit/token. The text field of each tweet is pulled out, cleaned and preprocessed. Lexicon based Sentiment Analysis was used for calculating the sentiment score of the tweets. Lexicon is a list of words tagged positive and negative. Each tweet was parsed and scanned for the presence of words matching the words from the list of positive and negative words from Lexicon. Based on the count of positive and negative words in the text, scores of +1 and -1 were allotted to each positive and negative word respectively. The score of a tweet was the net score of all its words. This process is repeated for all 8000 tweets. Based on the score, the tweet was identified as positive (>0), negative (<0) or neutral (=0). Example: "Demonetization is brilliant. But people are suffering a bit." Score = 0 (+brilliant, -suffering), Sentiment = Neutral

SYSTEM REQUIREMENTS

- Installation of R
- A Dataset Collection

Data Collection and Analysis

The dataset (containing 8000 tweets) was obtained from Kaggle and R was used for the analysis.

Steps:-

1. Importing library and loading required packages

```
library("readr")
library("tm")
library("wordcloud")
library("RColorBrewer")
library("qdap")
library("ggplot2")
```

2. Loading Word Database

A database, created by Hui Lui containing positive and negative words, is loaded into R. This is used for Lexical Analysis, where the words in the tweets are compared with the words in the database and the sentiment is predicted. AFINN is a list of English words rated for valence with an integer between minus five (negative) and plus five (positive). The words have been manually labeled by Finn Årup Nielsen in 2009-2011. The file is tab-separated. The version used is: AFINN-111: Newest version with 2477 words and phrases.

3. Cleaning Tweets: The tweets are cleaned in R by removing:

- Extra punctuation
- Stop words (Most commonly used words in a language like the, is, at, which, and on.)
- Redundant Blank spaces
- Emoticons
- URLS

4. Algorithms used

- ***Lexical Analysis***: By comparing uni-grams to the pre-loaded word database, the tweet is assigned sentiment score - positive, negative or neutral and overall score is calculated.

Lexical analysis and parsing tasks model the deeper properties of the words and their relationships to each other. The commonly used techniques involve word segmentation, part-of-speech tagging and parsing. A typical characteristic of such tasks is that the outputs are structured. Two types of methods are usually used to solve these structured prediction tasks: graph-based methods and transition-based methods. Graph-based methods differentiate output structures based on their characteristics directly, while transition-based methods transform output construction processes into state transition processes, differentiating sequences of transition actions. Neural network models have been successfully used for both graph-based and transition-based structured prediction. In this chapter, we give a review of applying deep learning in lexical analysis and parsing, and compare with traditional statistical methods. All of these tasks can be regarded as structured prediction problems which is a term for supervised machine learning, i.e., the outputs are structured and influence each other.

Results :-

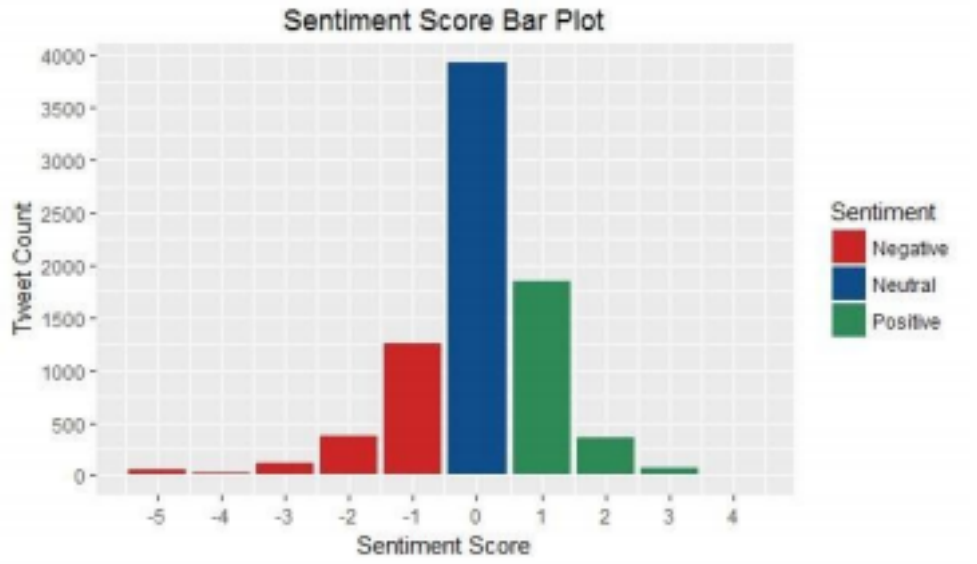
Sentiment Score

The distribution of scores obtained was as follows:

Score	-5	-4	-3	-2	-1	0	1	2	3	4
Tweets	47	26	112	371	1246	3922	1848	357	63	8

Mean score: 0.015

The mean score is slightly positive inferring that people are not having negative sentiment towards demonetization announcement.

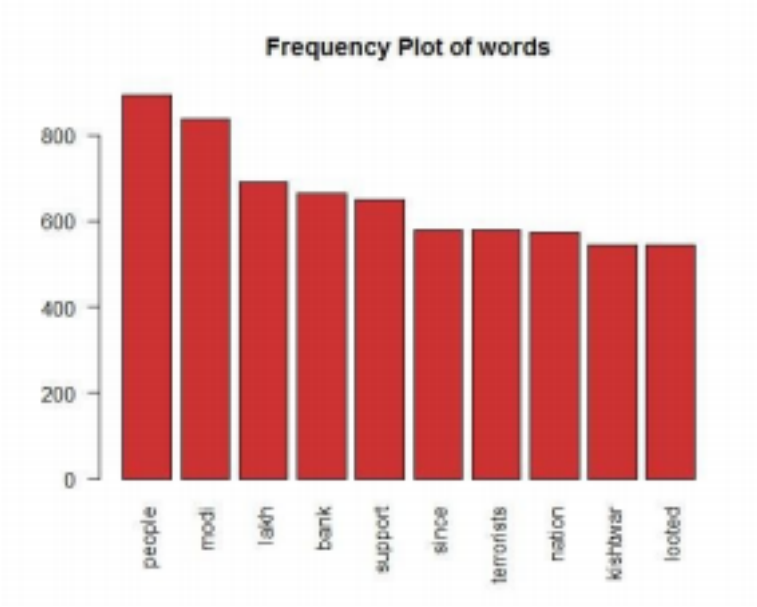


Sentiment	Tweets
Negative	1802
Neutral	3922
Positive	2276

Clearly, the sentiment of most of the tweets are neutral. Also, more number of positive tweets exist than negative ones which shows that the sentiments of people regarding demonetization are positive.

Wordcloud and Frequency plot of words

A word cloud was prepared based on the frequency of occurrence of words. The most frequently used keywords are towards the center of the word cloud. Size is related to the word frequency. As the frequency decreases, the size also decreases. Term Frequency plot gives the most frequent terms in the tweets

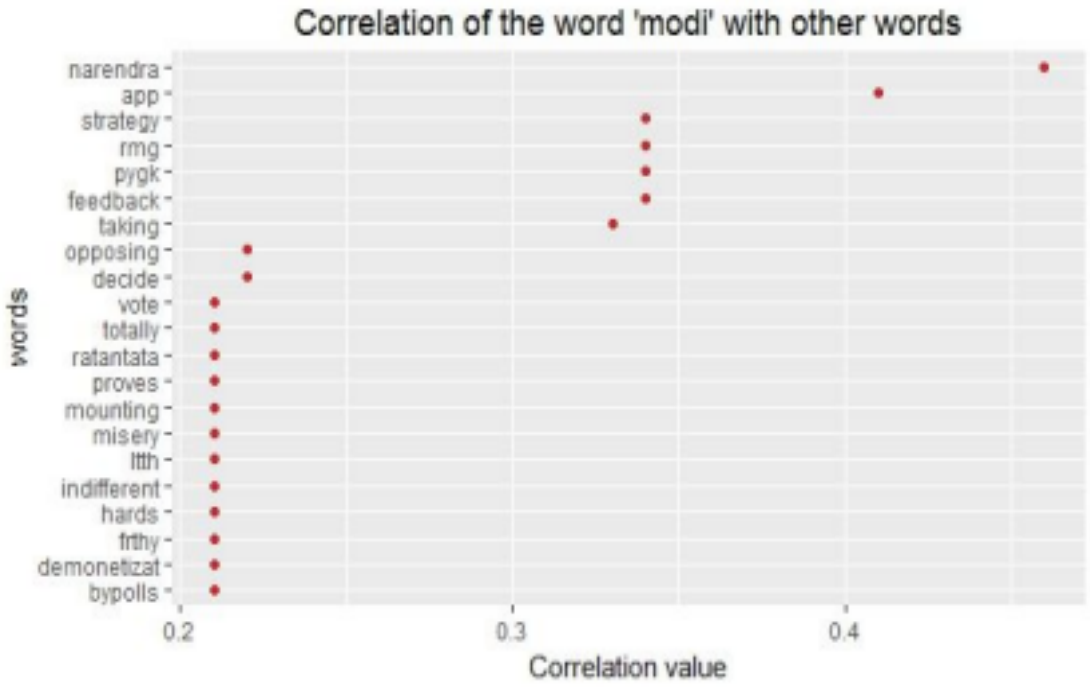


along with their frequencies as a bar plot.

From the word cloud obtained, it is clear that many people support demonetization. As expected, Modi, Bank, Nation, India, Loot, People – are prominently seen. The most frequent words are people, modi, lakh, bank and support.

Word Association Plot

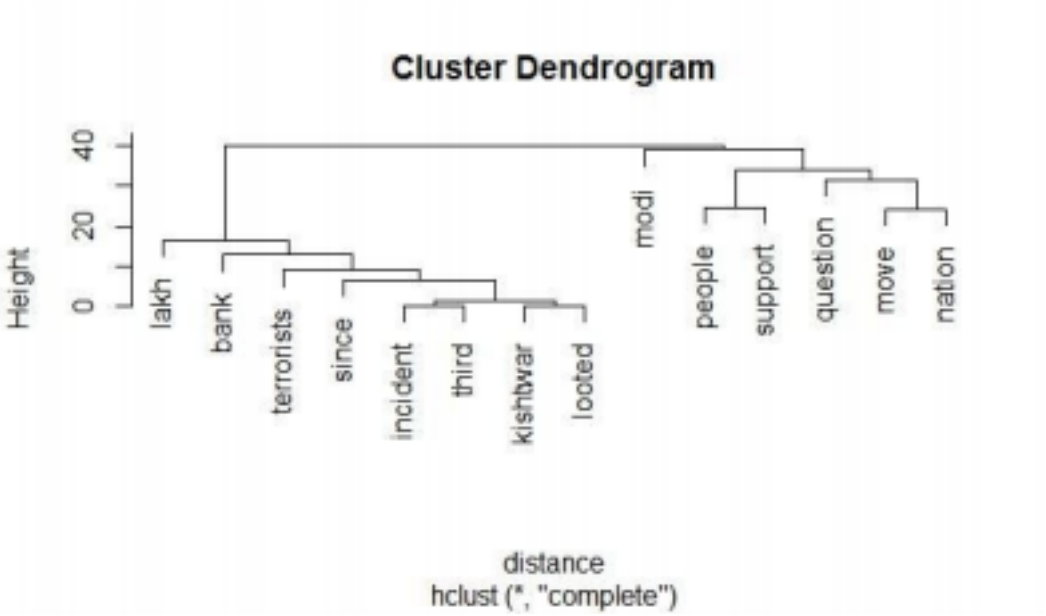
The word association was inspected for ‘Modi’ i.e. which words are closely associated with modi. Its correlation with every other word in the tweet was calculated and a dot plot was made along with their association values.



As seen from the plot, narendra and app have the highest association with modi and this makes sense. The Narendra Modi App was used to get feedback from people on the demonetization move and the results obtained in the app were overwhelmingly positive.

Dendrogram

Dendrogram is used to visualize hierarchical clustering of word frequency distances. A distance matrix is created for the words in the tweets and clustering is carried out to see how words fall into various clusters.



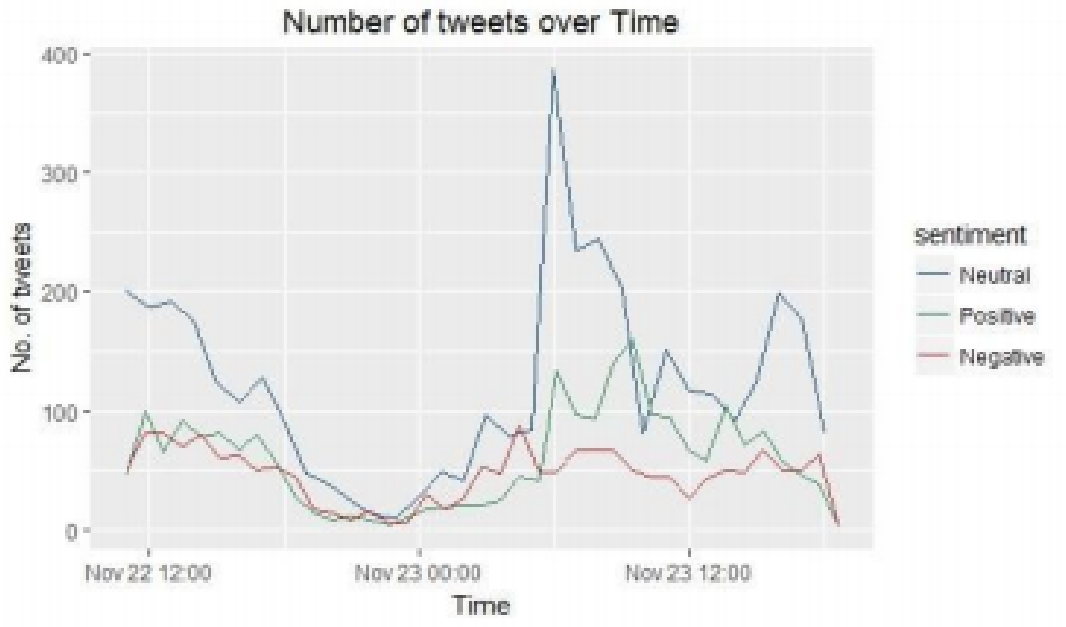
As seen from the plot, two well defined clusters are generated and the various terms in each cluster are depicted. The words people and support, move and nation are clustered together and fall under the same cluster indicating that there was support for modi from the people.

Variation of Tweets with time

The plot below shows the variation of the number of tweets with time on an hourly basis starting from November 22 11 A.M.

PM Narendra Modi asked for feedback on demonetization through twitter by requesting people to rate the measure on the modi app on 22 November, 11.25 A.M. There has been a decrease in the number of tweets since then until midnight. 40 lakh was looted from Jammu and Kashmir Bank in Kishtwar district on November 23. That explains the peak in the curve since there were many tweets reporting the incident. This incident also explains the considerably visible presence of the word ‘kishtwar’ on the wordcloud.

The number of positive tweets are mostly greater than the number of negative tweets at any point of time. The number of neutral tweets are greater than both at any point of time.



CODE :-

The entire code and the details of each part in a modular version can be find in our Github Repository.

The link: https://github.com/anshukrsingh/sentiment-analysis-on-demonetisation-using-R

PACKAGES USED :-

1. **readr** : readr provides a fast and friendly way to read rectangular data (like csv, tsv, and fwf).
2. **tm** : A framework for text mining applications within R.
3. **wordcloud**: visual representation in the form of wordcloud where size of the word is proportional to the frequency of words used in the tweets
4. **RColorBrewer**:Provides color schemes for maps (and other graphics)
5. **Qdap**: package provides parsing tools for preparing transcript data containing discourse including frequency counts of sentence types, words, sentences, turns of talk, syllables and other assorted analysis tasks.
6. **ggplot2**: An implementation of the grammar of graphics in R. It combines the advantages of both base and lattice graphics: conditioning and shared axes are handled automatically, and you can still build up a plot step by step from multiple data sources.

INFERENCES :-

A sentiment analysis was carried out on demonetization with twitter data to understand the people's reaction towards the demonetization step by PM Narendra Modi. Bag of words approach with lexicon based analysis was employed to calculate the sentiments. The results obtained gave strong evidence that the people of India support demonetization and are having a positive sentiment towards the

Demonetization step. Social media gives a thumbs up to Demonetization!

LIMITATIONS :-

1. The effect of negation words and sarcasm was not considered while calculating the sentiment score so not effective in detecting sarcasm.
2. Since the approach used was lexicon based, the list of positive and negative words can never be the complete list and hence most tweets were scored as neutral.
3. Cannot get 100% efficiency in analysing sentiment of tweets.

FUTURE WORK:-

1. Detect sarcasm in tweets
2. Analyse images for emotions
3. Add hindi words to the dataset.
4. Star rating (Negative and Positive [According to percentage]) (BOX PLOT)
5. Find no of mentions of n particular organizations (And analyse sentiment) ● Timeline of 7 days for emotions: anger, anticipation, disgust, fear, joy, sadness, surprise, trust
6. Extract from newspapers(TOI)
7. Apply better Machine Learning Algorithms (Like Support Vector Machine)

REFERENCES:-

- http://www.rdatamining.com/docs/twitter-analysis-with-r
- https://jeffreybreen.wordpress.com/2011/07/04/twitter-text-mining-r-slides
- https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html
- http://www.rdatamining.com/docs/twitter-analysis-with-r
- https://github.com/datumbox/twitter-sentiment-analysis
- https://www.r-bloggers.com/emoticons-decoder-for-social-media-sentimentanalysis-in-r/
- https://www.r-bloggers.com/twitter-sentiment-analysis-with-r/
- https://www.r-bloggers.com/how-to-create-a-twitter-sentiment-analysis-using-r-and-shiny/
- https://link.springer.com/chapter/10.1007/978-981-10-5209-5_4/
- The Arc | Medium
- Demonetization a drive towards digital economy
- Demonetization is Narendra Modi-made disaster: Rahul Gandhi