**CHAPTER I**

**INTRODUCTION**

Sentiment Analysis is a technique widely used in text mining. Twitter Sentiment Analysis, therefore means, using advanced text mining techniques to analyze the sentiment of the text (here, tweet) in the form of positive, negative and neutral. It is also known as Opinion Mining, is primarily for analyzing conversations, opinions, and sharing of views (all in the form of tweets) for deciding business strategy, political analysis, and also for assessing public actions.

Ingenuity, Revealed Context, Steamcrab, MeaningCloud, and SocialMention are some of the well-known tools used for the analysis of Twitter sentiment. R and Python are widely used for sentiment analysis dataset twitter. Sentiment Analysis of Twitter data is now much more than a college project or a certification program. A good number of Tutorials related to Twitter sentiment are available for educating students on the Twitter sentiment analysis project report and its usage with R and Python. You may also enroll for a python tutorial for the same program to get a promising career in sentiment analysis dataset twitter.

Our discussion will include, Twitter Sentiment Analysis in R and also throw light on its techniques and teach you how to generate the Twitter Sentiment Analysis project report.

**Context**

* Numerous outlets available for individuals to express opinions and emotions...positive, negative, and neutral.
* Need to promote positive news, react to the negative, and move the needle favorably on neutral news....as near real-time as possible
* Mining high volume, high velocity data for meaningful insights is not easy!...too much, too fast
* Similar challenges exist across all industries/verticals

**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, bar plots, timelines, word clouds, histograms and pie charts can be used for visualization.

**Twitter 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 October 2016, Twitter has more than [​315 million monthly active users](https://en.wikipedia.org/wiki/List_of_virtual_communities_with_more_than_100_million_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.

**Twitter Sentimental 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.

For example, the Obama administration used sentiment analysis to gauge public opinion to policy announcements and campaign messages ahead of 2012 presidential election.

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 adverts.

**Business**: Companies use Twitter Sentiment Analysis to develop their business strategies, to assess customers’ feelings towards products or brand, how people respond to their campaigns or product launches and also why consumers are not buying certain products.

**Politics**: In politics Sentiment Analysis Dataset Twitter is used to keep track of political views, to detect consistency and inconsistency between statements and actions at the government level. Sentiment Analysis Dataset Twitter is also used for analyzing election results.

**Public Actions**: Twitter Sentiment Analysis also is used for monitoring and analyzing social phenomena, for predicting potentially dangerous situations and determining the general mood of the blogosphere.

**CHAPTER II**

**LITERATURE SURVEY**

Sentiment analysis is the most important research area in business fields. Previously research was carried out for sentiment analysis in various domains like company product, movie reviews, politics etc. Previous research like Pang et al. has provided with the baseline for carrying out research in various domains. It uses star ratings as polarity signals in their training data. Even many authors have used the same concept provided by Pang et al.

## Earthquake shakes twitter users: Real-time event detection by social sensors:

T.Sakaki et al. developed an event notification system which monitors the tweets and delivers notifications considering the time constraint. They detect real-time events in Twitter such as earthquakes. They have proposed an algorithm to monitor tweets detecting target event. Each Twitter user is considered as a sensor. Kalman filtering and particle filtering are used for estimation of location.

# 

# Data Set:

# For classification of tweets, we prepared 597 positive examples which report earthquake occurrence as a training set.

# Advantages:

1. Main task of earthquake detection is done using the system. Users are registered with it and email messages are sent to them.

2. The two filtering techniques detect and provide estimation for location.

# Disadvantages:

1. Multiple events cannot be detected at a time.

2. It cannot provide advanced algorithms to expand queries.

3. Limited to only one target event detection at a single time event.

4. It uses SVM as a classifier into positive and negative sentiments which is not applicable to small data sets.

## CHAPTER III

## OBJECTIVE

We will take a trending statement so we could get more details about and big dataset for analyzing purpose. In Twitter sentimental analysis tweets are imported using R and the data is cleaned by removing emoticons and URLs. Lexical Analysis is used to predict the sentimental of tweets and subsequently express the opinion graphically through histogram. Sentiment Analysis is a technique widely used in text mining. Twitter Sentiment Analysis, therefore means, using advanced text mining techniques to analyze the sentiment of the text (here, tweet) in the form of positive, negative and neutral. It is also known as Opinion Mining, is primarily for analyzing conversations, opinions, and sharing of views (all in the form of tweets) for deciding business strategy, political analysis, and also for assessing public actions.

These days, the applications of such analysis can be easily observed during public elections, movie promotions, brand endorsements and many other fields. In this project, we exploited the fast and in memory computation framework 'Apache Spark' to extract live tweets and perform sentiment analysis. The primary aim is to provide a method for analyzing sentiment score in noisy twitter streams. This paper reports on the design of a sentiment analysis, extracting vast number of tweets. Results classify user's perception via tweets into positive and negative.

## 

**CHAPTER IV**

**METHODOLOGY**

**Overview**

Tweets are imported using R and the data is cleaned by removing emoticons and URLs. Lexical Analysis as well as Naive Bayes Classifier is used to predict the sentiment of tweets and subsequently express the opinion graphically through ggplots, histogram, pie chart, wordcloud and tables. The front end has been created using the Shiny App.

**System Requirements**

* Installation of R
* Twitter Authentication to access API

**Features**

**1. Extraction of Tweets**

* 1. Create twitter application
  2. twitteR - Provides an interface to the Twitter web API
  3. ROAuth - ​R Interface for OAuth

(iv)​Create twitter authenticated credential object (using key from step (ii) and cacert.pem certificate): It is done using consumer key, consumer secret, access token, access secret.

1. During authentication, we are redirected to a URL automatically where we click on Authorize app as shown in the image below and enter the unique 7-digit number to get linked to the account from which feeds are being taken.

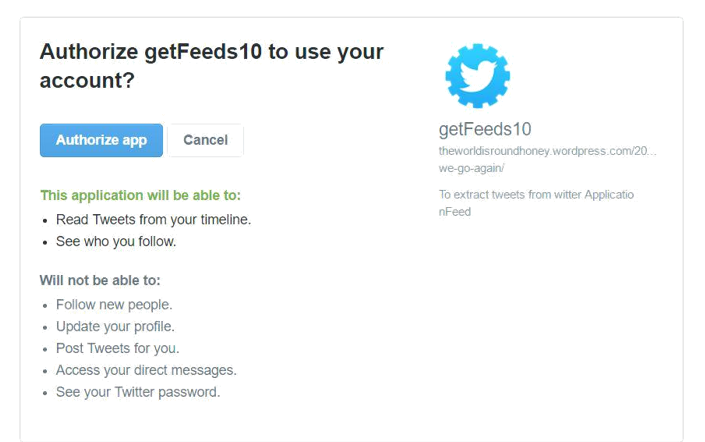


Fig 4.1: Extraction of Tweets

**2. 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

**3. 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.

For movie tweets, Naive Bayes Machine Learning Algorithm is used. AFINN is a list of English words rated for valence with an integer between minus five (negative) and plus five (positive).

**4. Algorithms used**

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

**Naive Bayes Machine Learning Algorithm: ​**Training data sets areused to teach the machine what kind of sentences are categorized as positive and what kind are categorized as negative. On arrival of a new tweet or sentence, the machine uses this algorithm to give the correct category to the new data and adds level to the emotion.

**5.Calculating percentage**

**​** In the table tab of our Shiny Web app as shown below, we have presented thescores, the tweets as well as the percentage of positive/negative emotion in the text. Th is calculated using simple arithmetic to understand the overall sentiment in a better manner.

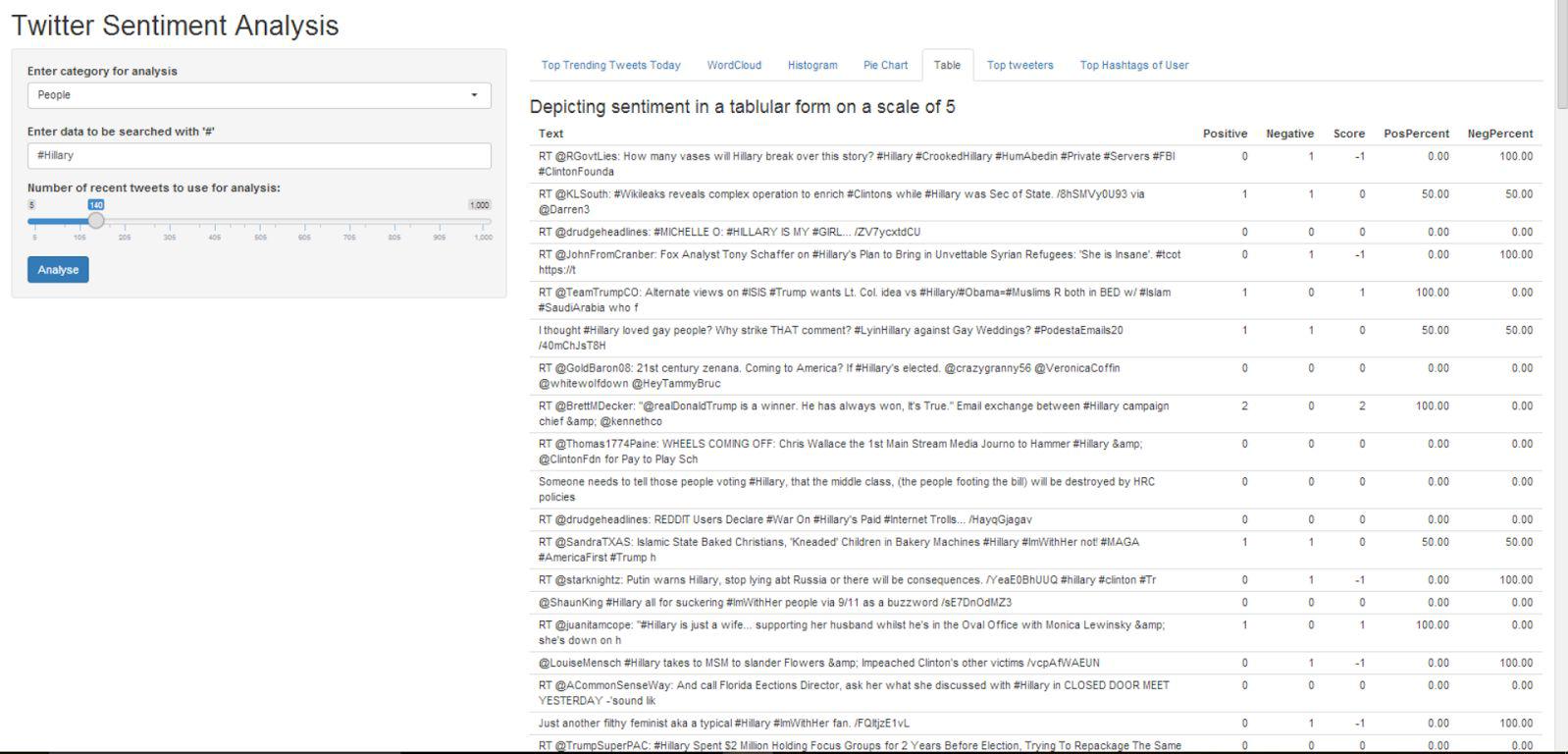


Fig 4.2: Calculating Percentage

**6.Top Trending Tweets Today tab: Table**

The table is shown which displays the top trending hashtags on Twitter of the location that has been selected.

A ​**WOEID (Where On Earth IDentifier) ​**is a unique 32-bit reference identifier, which is generated, and R uses the WOEID of the selected place to obtain the trending hashtags from that location.

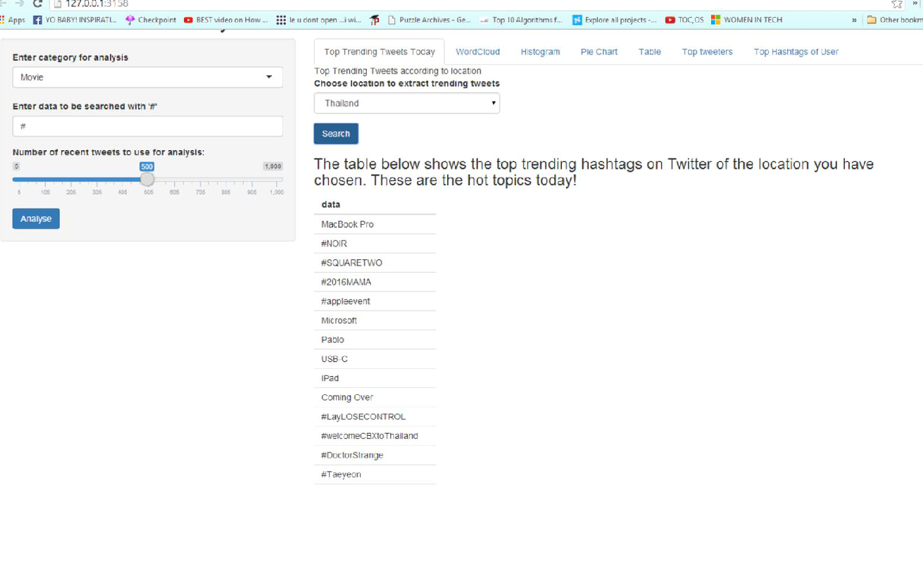


Fig 4.3: Top Trending Tweets

**7. Word Cloud tab: wordcloud**

A word cloud is a visual representation of text data, typically used to depict keyword metadata (tags) on websites, or to visualize free form text. This format is useful for quickly perceiving the most prominent terms and for locating a term alphabetically to determine its relative prominence. We have used tm and wordcloud package to depict the most used words associated with the hashtag in a pictorial representation under the Wordcloud tab.

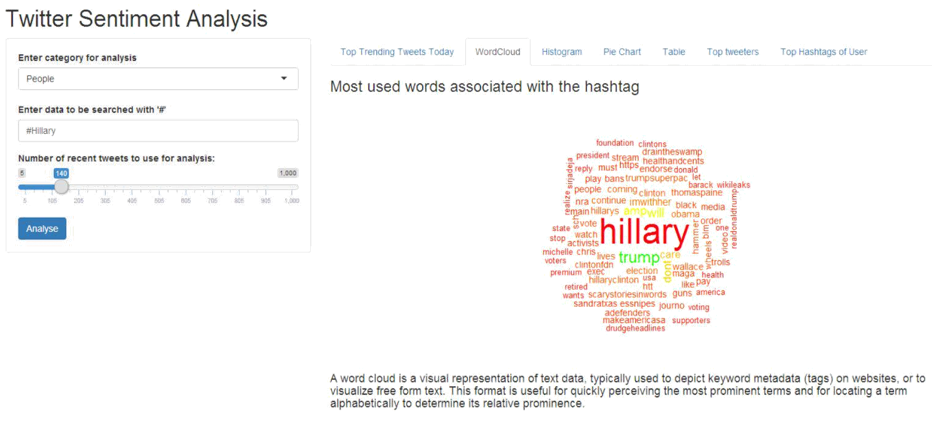


Fig 4.4: Word Cloud

**8. Graph tab: gg plot**

Histograms of positive, negative and overall score are found under the gg plot for graphically analyzing the intensity of emotion in the tweeters.

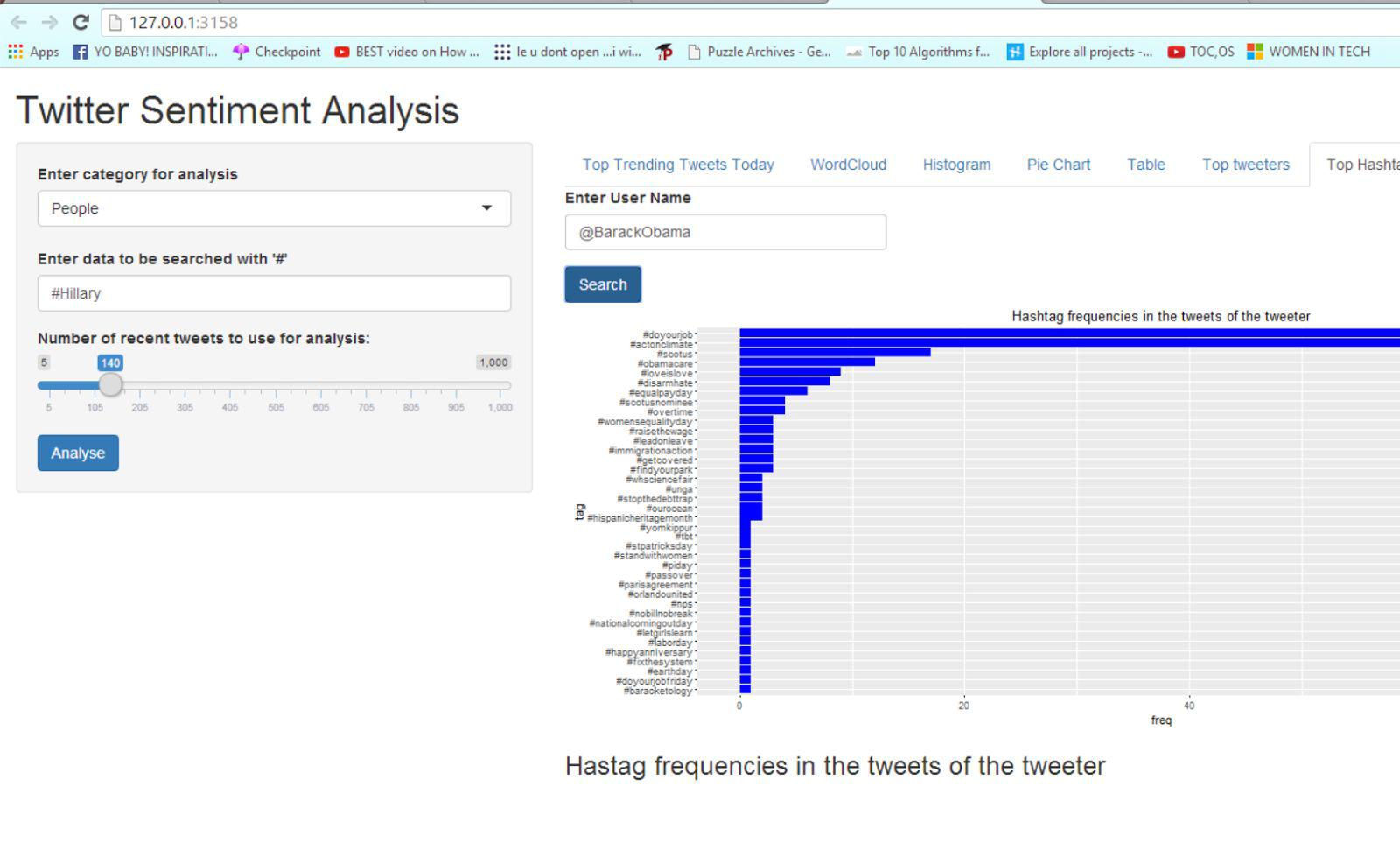


Fig 4.5: Graph Tab

**Project Implementation**

**Code**

#Connecting\_to\_Twitter\_api

api\_key = "GStibPIrmlNybfNajlIacNBDn"

api\_secret<-"jUaT7Bf09FP8HpmCdQicsUwPKpKcKdtOFzaV0b3Uc09IZEhiSh"

access\_token<-"212500992-Mrjm52E3BLc8gaoqGXLui8hgLkrezw2gpn58IocI"

access\_token\_secret<-"VfQhQXYIrCdrWhrFJrD0VZdBbk4DtfSfRc7TXZ1sj0rSb"

setup\_twitter\_oauth(api\_key,api\_secret,access\_token,access\_token\_secret)

#Cleanning\_Data

df1 <- twListToDF(bitcoin)

myCorpus <- Corpus(VectorSource(df1$text))

removeURL <- function(x) gsub("http[^[:space:]]\*", "", x)

myCorpus <- tm\_map(myCorpus, content\_transformer(removeURL))

removeNumPunct <- function(x) gsub("[^[:alpha:][:space:]]\*", "", x)

myCorpus <- tm\_map(myCorpus, content\_transformer(removeNumPunct))

myCorpus <- tm\_map(myCorpus, stripWhitespace)

myCorpusCopy <- myCorpus

myCorpus <- tm\_map(myCorpus, stemDocument)

myCorpus <- Corpus(VectorSource(myCorpus))

df <- twListToDF(bitcoin)

df <- df[, order(names(df))]

df$created <- strftime(df$created, '%Y-%m-%d')

if (file.exists(paste("bitcoin", '\_stack.csv'))==FALSE) write.csv(df, file=paste("bitcoin", '\_stack.csv'), row.names=F)

stack <- read.csv(file=paste("bitcoin", '\_stack.csv'))

stack <- rbind(stack, df)

stack <- subset(stack, !duplicated(stack$text))

write.csv(stack, file=paste("bitcoin", '\_stack.csv'), row.names=F)

score.sentiment <- function(sentences, pos.words, neg.words, .progress='none')

require(plyr)

require(stringr)

scores <- laply(sentences, function(sentence, pos.words, neg.words){

sentence <- gsub('[[:punct:]]', "", sentence)

#Plot of the sentiment analysis

ggplot(by.tweet, aes(created, number)) + geom\_line(aes(group=tweet, color=tweet), size=2) +

geom\_point(aes(group=tweet, color=tweet), size=4) +

theme(text = element\_text(size=18), axis.text.x = element\_text(angle=90, vjust=1)) +

ggtitle(bitcoin)

by.tweet <- group\_by(stat, tweet, created)

by.tweet <- summarise(by.tweet, number=n())

write.csv(by.tweet, file=paste("bitcoin", '\_opin.csv'), row.names=TRUE)

pos <- scan('C:\\Program Files\\R\\positivewords.txt', what='character', comment.char=';')

neg <- scan('C:\\Program Files\\R\\negativewords.txt', what='character', comment.char=';')

**CHAPTER V**

**RESULT AND DISCUSSION**

**Creating a Word cloud**

We will create a world cloud for the most frequent words.



Fig 5.1: Representing the word cloud

**Plot Most Frequent Words**

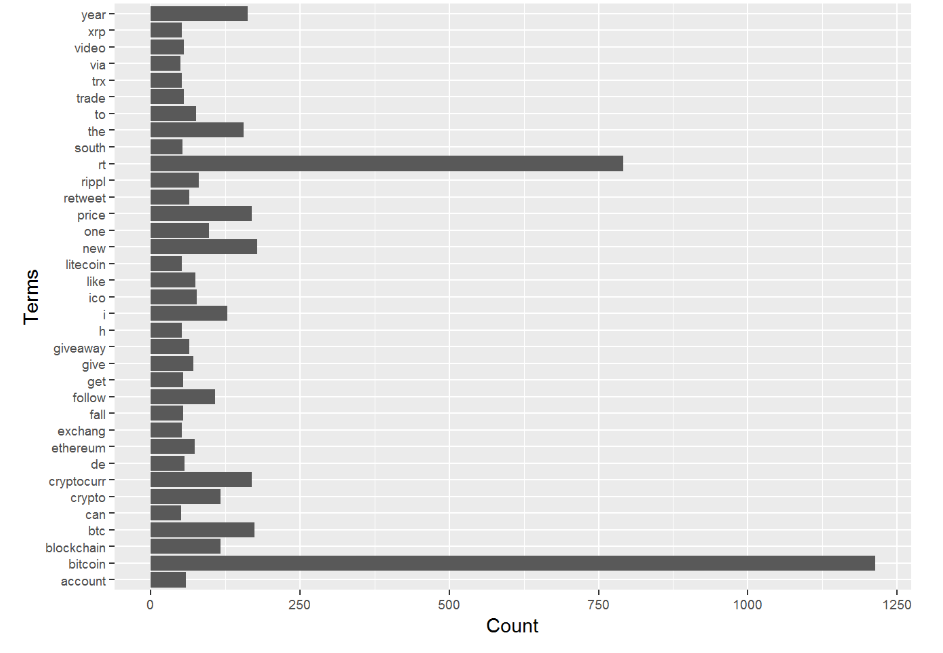


Fig 5.2: Representing the Most Frequent Words

**Plot of Sentimental Analysis**

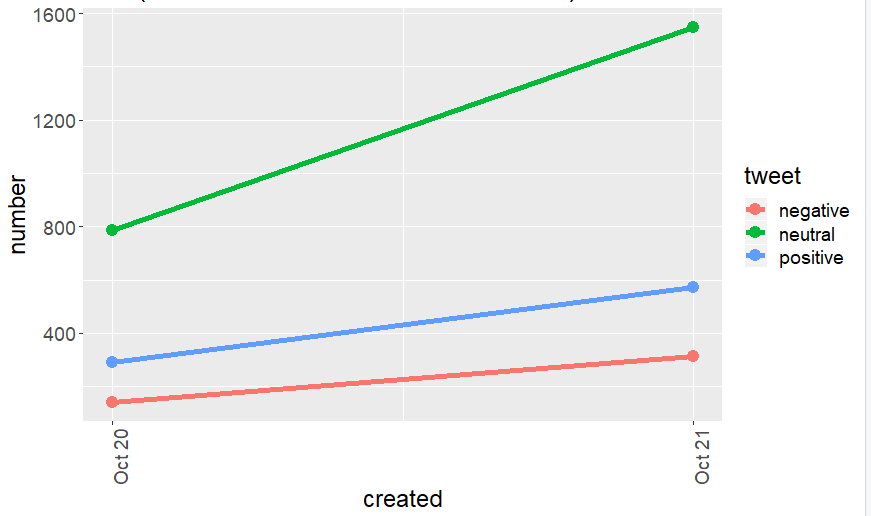


Fig 5.3: Representing Sentimental Analysis

Hence it can infer that the public sentiment towards bitcoin was neutral in oct 20 2019.There is noticeable change in the sentiment attached to each category. Especially in **Bitcoin** related Tweets, where the number of positive comments has dropped from 46% to 29%.

This gives us a glimpse of how CSS can generate in-depth insights from digital media. A brand can thus analyze such Tweets and build upon the positive points from them or get feedback from the negative ones.

**CHAPTER VI**

**CONCLUSION AND FUTURE WORK**

The rise of social media has fueled interest in sentiment classification. Promptly and correctly classifying sentiment from the text has become an important task for individuals and companies. In the development of prediction models to classify the reviews, more reliable approaches are expected to reduce the misclassifications. We can use this sentimental analysis for predicting the positive and negative outcome for a given topic. we can analyses it for better result in future and predict some of the upcoming threats and neglect it or tackle it

In future, we can use this analysis to predict upcoming threats and tackle it like suicide and many more related problems can be easily solved or at least used to get an idea of the project.

**Future Work**

* Detect sarcasm in tweets
* Analyze images for emotions
* Add tamil, hindi words to dataset.
* Star rating (Negative and Positive [According to percentage]) (BOX PLOT)
* Find no of mentions of n particular organizations (And analyze sentiment)
* Timeline of 7 days for emotions: anger, anticipation, disgust, fear, joy, sadness, surprise, trust
* Extract from newspapers (TOI)
* Parallelizing code
* Apply better Machine Learning Algorithms (Like Support Vector Machine)

**REFRENCES**

1. Stone, Philip J., Dexter C. Dunphy, and Marshall S. Smith. "The general inquirer: A computer approach to content analysis." MIT Press, Cambridge, MA (1966).
2. Gottschalk, Louis August, and Goldine C. Gleser. The measurement of psychological states through the content analysis of verbal behavior. Univ of California Press, 1969.
3. [USA Issued 7,136,877](http://patft.uspto.gov/netacgi/nph-Parser?Sect1=PTO2&Sect2=HITOFF&u=%2Fnetahtml%2FPTO%2Fsearch-adv.htm&r=2&p=1&f=G&l=50&d=PTXT&S1=(fogel.INNM.+AND+volcani.INNM.)&OS=in/fogel+and+in/volcani&RS=(IN/fogel+AND+IN/volcani)), Volcani, Yanon; & Fogel, David B., "System and method for determining and controlling the impact of text", published June 28, 2001.
4. Turney, Peter (2002). "Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews". Proceedings of the Association for Computational Linguistics. pp. 417–424
5. Pang, Bo; Lee, Lillian (2004). ["A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts"](http://www.cs.cornell.edu/home/llee/papers/cutsent.home.html). Proceedings of the Association for Computational Linguistics (ACL).
6. Zhai, Zhongwu; Liu, Bing; Xu, Hua; Jia, Peifa (2011-01-01). Huang, Joshua Zhexue; Cao, Longbing; Srivastava, Jaideep (eds.). Constrained LDA for Grouping Product Features in Opinion Mining. Lecture Notes in Computer Science. Springer Berlin Heidelberg. pp. 448–459.
7. Galitsky, Boris; Chen, Huanjin; Du, Shaobin (2009). "Inversion of Forum Content Based on Authors' Sentiments on Product Usability". AAAI Spring Symposium: Social Semantic Web: Where Web 2.0 Meets Web 3.0: 33–38.
8. [Karlgren, Jussi](https://en.wikipedia.org/wiki/Jussi_Karlgren" \o "Jussi Karlgren). "Affect, appeal, and sentiment as factors influencing interaction with multimedia information." In Proceedings of Theseus/ImageCLEF workshop on visual information retrieval evaluation, pp. 8-11. 2009.