Web and Social Media Analytics

Contents

[1 Problem Statement 1](#_Toc52132548)

[2 Introduction 2](#_Toc52132549)

[3 Brand Selection 2](#_Toc52132550)

[3.1 A brief Introduction on the application Zoom 2](#_Toc52132551)

[4 Scrapping of tweets and EDA 3](#_Toc52132552)

[5 Data Cleansing 4](#_Toc52132553)

[6 Analyzing Text Frequency 6](#_Toc52132554)

[7 Word Cloud 7](#_Toc52132555)

[8 Sentiment Analysis 11](#_Toc52132556)

[8.1 Constraints of Sentiment Analysis 14](#_Toc52132557)

[9 Correlation – Word Clustering 15](#_Toc52132558)

[10 Word Association 16](#_Toc52132559)

[11 Summary 19](#_Toc52132560)

# Problem Statement

Objective: Brand perception analysis (Social media analytics - Text mining)

Instructions for each group:

* Identify a brand - any global or Indian product, celebrity, company, etc
* For the identified brand download minimum 1000 twitter messages for the most recent period
* Perform EDA and Data Cleaning
* Conduct text mining on the data – Correlation, Frequency, Topic Modelling using Word Association, Sentiment Analysis
* Can leverage any tool for data visualization, text mining, ML, etc
* Can expand the source of data to Facebook or any other source as well, but Twitter is a must

# Introduction

The process of extraction of data can be collectively termed as text mining. The information extracted can be of various formats like information, data, or knowledge-based discovery. In text mining we often try to understand the meaning behind text / data which is available. Text analytics is usually done to understand customer behavior, sentiment of subjects, poll / review of subjects and to understand the meaning of many words. While collecting a database for the text mining, there can be a lot of stop words, punctuations, URL links, white spaces which can try to deviate from the analytics. Hence it is important to remove them from the dataset. Depending on what data is being used for text mining we create a word cloud of the words showing sentiments regarding the analysis. It is important that we know the meaning of words used since it plays an important role in identifying what the subject is telling and the nature of his tweet.

# Brand Selection

We have chosen to do our text mining analysis on “[zoom\_us](https://twitter.com/zoom_us)”. For this we downloaded tweets from twitter on topics in which had @zoom\_us was mentioned.

## A brief Introduction on the application Zoom

Zoom is a US based communications application which allows users to create a virtual meeting with attendees who can be friends, colleagues, students or business associates. It was originally developed in China. The app also allows one to share, receive and comment on files, folders, images, videos and has a wide range of tools which allows one to shift between modes. It also is widely used for the teaching since the meeting can be recorded. It allows the use of Sideboard and slate, which can be very useful for explaining concepts. In the recent situation because of the novel coronavirus (covid 19) there has been large dependency by companies, school, techno companies to use Zoom\_us to run their business. But there is also limelight on the fact that the application is trying to access data from personal files and storing these files in servers in remote locations of China. This has led to a large number of users to dislike Zoom and have a strong negative sentiment for the app. Also with the recent launch of Jio Meet app, which is said to be very similar to Zoom, it has started Zoom vs JioMeet discussions among the users. We have chosen this brand for our analysis since it will help to understand the overall feedback and the chatter on tweet.

The R codes in this document are represented using brown highlights.

# Scrapping of tweets and EDA

We started our analysis with understanding the tweets which were downloaded using the ‘rtweet’ package in R. Authentication was set-up and the following is the code to download.

We first tried downloading tweets and found a large number of retweets and resharing which first needs to be addressed. For this purpose, tweets which were shared or repeated were first removed so that the dataset has 3000 unique tweets.

#search for tweets in which mentions the brand "Zoom"

zoom\_tweets <- search\_tweets(q="@zoom\_us",n=3000,lang="en",include\_rts = FALSE)

The tweets were converted into data frames for further analysis.

#converting to data frames

zoom\_tweets\_df <- zoom\_tweets

dim(zoom\_tweets\_df)

We carried our EDA on the data set and find out that for 3000 tweets we had 91 variables carrying various information such as text of tweet, User ID, screen name, hashtags, retweets attributes.

There were some interesting facts which was observed in the EDA;

1. The 3000 tweets were downloaded on 8 July 2020, so the context and topics are around that date.
2. There was one such tweet which has been favorited 14268 times and the same retweeted 2964 times.
3. 2301 tweets had hashtag in them.
4. 93% of the tweet are from Twitter for Android, iPhone and web-based applications.

The Tweets were converted into R object for carrying out further analysis, we retain only the text column for analysis

#extracting the tweet text for data cleaning and further analysis

tweets <- zoom\_tweets\_df$text

#converting text into document format

zoom\_source <- VectorSource(tweets)

#converting it into a R object

zoom\_corpus <- VCorpus(zoom\_source)

Let us view some tweets to understand

zoom\_corpus[[25]][1]

$content

[1] "@fs0c131y @zoom\_us What would you use indian or Chinese??"

zoom\_corpus[[130]][1]

$content

[1] "@zoom\_us launches #HardwareasaService - businesses today need exceptional resilience and agility for cost-effective and flexible technology investments https://t.co/QkgAFhojIb"

# Data Cleansing

After completing the EDA, it is important that all the data is cleansed of stop words, word clutter, punctuation marks and URL. It is the process of ensuring the data is correct, consistent and usable by identifying any errors or corruptions found in the data by correcting or deleting them. This helps to simplify the data and make good use in the analysis.

The following libraries were used

library(tm)

library(NLP)

#creating function for removing URL and cleaning corpus

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

clean\_corpus <- function(corpus){

corpus <- tm\_map(corpus,stripWhitespace)

corpus <- tm\_map(corpus, removePunctuation)

corpus <- tm\_map(corpus, content\_transformer(tolower))

corpus <- tm\_map(corpus, removeNumbers)

corpus <- tm\_map(corpus, removeWords,c(stopwords("en"),"zoom", "zoomus"))

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

return(corpus)

}

#creating a clean corpus

clean\_corp <- clean\_corpus(zoom\_corpus)

The above code was used to remove whitespaces, punctuations, remove all causing related issues, removing Numbers, removing stop words and removing URLs. We had introduced zoom and zoomus as additional stop words as many tweets may include them and these words won’t have a significant meaning in the analysis.

This helped to clean the corpus so that we can analyze the text data much better.

The next step was to create a document term matrix/ term document matrix, which is a mathematical expression that describes the frequency a term which occurs in a collection of documents. In a document term matrix, the number of rows correspond to the documents in the collection and columns corresponds to terms.

There are various schemes for determining the value that each entry in the matrix should take. Document term matrix is also very useful in the field of natural language processing.

#termdocument and documentterm matrix format

zoom\_tdm <- TermDocumentMatrix(clean\_corp)

zoom\_dtm <- DocumentTermMatrix(clean\_corp)

#create as matrix

zoom\_m <- as.matrix(zoom\_tdm)

zoom\_mc <- as.matrix(zoom\_dtm)

# Analyzing Text Frequency

After creating a term document matrix, it is important to identify words that are most frequently used so that we can analyze the sentiments in the dataset. For this purpose, we need to find out which are the frequently used words and what sentiment does it carry with it. This helps us understand the customer behavior pattern and also helps to understand the situation of the customer. There can be a situation where the words might have a tone of sarcasm, it is important in text analysis to carefully classify the sentiment in the correct basket.

#calcuate frequency of terms

term\_frequency <-rowSums(zoom\_m)

#sort frequcny of terms

term\_frequency <- sort(term\_frequency, decreasing = TRUE)

#view top 10 freqeuncy words

term\_frequency[1:10]

#plot

barplot(term\_frequency[1:10], col="blue", last =2)

We found the most frequently used words with the hashtag Zoom\_us were

* Chinese,
* amp,
* can,
* will,
* reliancejio,
* just,
* like,
* now,
* use,
* narendarmodi.

This indicates that these words were most used when the tag Zoom\_us was used. This can also be understood that recently with the tension mounting between India and China the Indian Govt has decided the ban of large number of Chinese applications, there was a confusion as to whether Zoom\_us will be banned in India. Another observation is that Jiomeet is a communication application which is similar to zoom was launched by Reliance Industries.

Following is the plot of 10 frequently used words in the tweets.

A picture containing computer, drawing

Description automatically generated

# Word Cloud

Word cloud is an analytical tool to identify which are the most used words and how it plays a significant role in NLP for text-based analysis. The library “wordcloud” was used for plotting word cloud.

#overall word cloud

wordcloud(clean\_corp,colors=rainbow(7),max.words=50, min.freq=30)

zoom\_dtm\_as\_tidy <- tidy(zoom\_dtm)

These are the words which are the most frequently used in the dataset. This combines positive and negative attributes which have been used for conveying messages about Zoom\_us.

A screenshot of a cell phone

Description automatically generated

Before segregating the word cloud into positive and negative sentiments it is necessary to choose the correct library.

# assessing sentiments of the words, converting tidy model

zoom\_dtm\_as\_tidy <- tidy(zoom\_dtm)

We used the library bing Lexicon, there are another dictionary like NRC/ GFINN etc.

#This gives a list of positive and negative words form "bing" dictionary

bing <- get\_sentiments("bing")

#checking with words with the our dtm of tweets

as\_bing\_words <- inner\_join(zoom\_dtm\_as\_tidy,bing,by = c("term"="word"))

# check positive and negative words

as\_bing\_words

A screenshot of a cell phone

Description automatically generated

This indicates that the sentence having words like benefits, well, glorious, ample have a positive outcome. The word such as “problem” shows a negative outcome.

Positive Words:

Words which describe something in a positive manner are positive words. Words like good, best, pleased, brilliant a positive sentiment where the meaning will mostly be that the customer appreciates the Zoom\_us application. Here are some positive word clouds generated in R.

#selecting positive words

positive\_words <- as.matrix(as\_bing\_words[as\_bing\_words$sentiment=="positive",2])

#positive word cloud

wordcloud(positive\_words,colors=rainbow(7),max.words=50)

The words are highlighted in different colors. We can see that the below word cloud is made up of only positive words such as Like, great, work, better, good etc. already show a positive sentiment, higher the font size higher is the frequency.

A screenshot of a cell phone

Description automatically generated

Negative Word:

Words which describes something in a negative or derogatory manner are negative words. Words like problem, sorry, boycott show negative sentiments where the user must be having negative notation about the Zoom\_us app.

#selecting negative words

negatie\_words <- as.matrix(as\_bing\_words[as\_bing\_words$sentiment=="negative",2])

#negative word cloud

wordcloud(negative\_words,colors=rainbow(7),max.words=50)

Words like issue, sorry, problem, boycott show a negative sentiment since they can be words used to defame or reduce brand value of a particular brand. It is necessary to identify sentiment analysis where complex words can create a puzzle.

A screenshot of a cell phone

Description automatically generated

# Sentiment Analysis

Sentiment analysis is the interpretation and classification of emotions which can be positive, negative or neutral which is present in the text data using various text analysis techniques. Sentiment analysis tools allow businesses to identify customer sentiment towards products, brand or services in online feedback platforms.

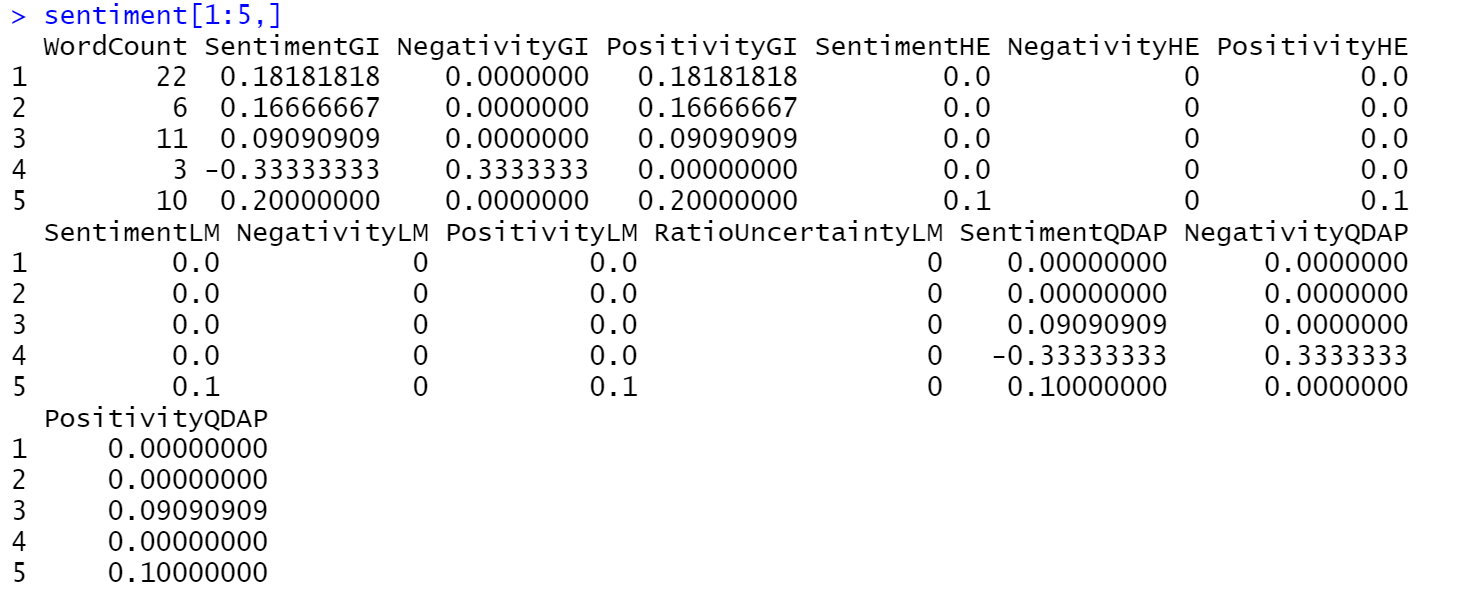
Now we know what words are positive and negative, now it becomes important that each tweet is given a number of positive or negative so that we can understand the analysis better.

Here we use the “SentimentAnalysis” library on the dtm matrix.

#analysisng each tweet to find out sentiment

sentiment <- analyzeSentiment(zoom\_dtm)

Now let us find the sentiment score of the first five tweets.



There are 4 dictionaries used, for each dictionary a positive, negative and sentiment (negative +positive) score which is given;

#GI - Harvard-IV dictionary

#HE - Henry's Financial dictionary (Henry 2008)

#LM -Loughran-McDonald Financial dictionary

#QDAP - QDAP dictionary

The first tweet/document has a sentiment score of 0.18181818 as per GI dictionary which means it is positive tweet. See the scores are different for each dictionary meaning they are computed in different ways

The sentiment scores are in decimals which may be difficult to interpret so

we need to convert them to sentiment direction (positive / negative)

#-1 mean negative sentiment + 1 positive sentiment, 0 means nuteral

#We will QDAP dictionary.

sentimentdirection <- convertToDirection(sentiment$SentimentQDAP)

#let understand count of positives

table(sentimentdirection)

A picture containing knife, table

Description automatically generated

There are in total 1209 neutral comments, 1549 positive and 240 negative sentiments as per the analysis by the function.

plot(sentimentdirection)

A screenshot of a cell phone

Description automatically generated

This indicates that almost 50% of the tweets have a positive sentiment this can also be visible by plotting sentiment score plot.

plot(sentiment$SentimentQDAP)

A screenshot of a cell phone

Description automatically generated

There are less sentiment scores values below the 0-line mark which means relatively fewer negative tweets.

## Constraints of Sentiment Analysis

One constrains to be taken into account while performing such basic sentiment analysis is the failure to detect sarcasm, context or perspective. For example, there are a few tweets in which users have said they will be joining “JioMeet” however these are classified as positive since the sentence itself is positive but may not be positive from Zoom’s perspective. Examples of such tweets are classified as positive/negative but mean the opposite from Zoom’s perspective are given in the below table.

|  |  |  |
| --- | --- | --- |
| **Tweet Extract** | **Classified Sentiment** | **Actual Sentiment knowing the context** |
| So don't be so quick to dump any other and opt for jio. I did the same. Left BSNL and chose jio and now I am suffering. So think wise. I have said what I faced. The worst customer care ever i have seen | negative | positive |
| I have already uninstalled zoom as now we have our own better version called #JioMeet. | positive | negative |
| I will prefer microsoft team or jio meet. | positive | negative |
| @reliancejio Its excellent. With new updates it's much better than Zoom.   Sharing audio seems a problem/not updated as of now. | positive | negative |
| @reliancejio with @jiomeet by providing 24hrs call feature and password protecting hosting. Good competitor of @zoom\_us | positive | negative |

# Correlation – Word Clustering

Let us find out which words come together – this can be achieved with doing clustering of terms across the tweets. Words that appear together will be grouped into one cluster.

The first step is to remove words which are sparse words, i.e. those very less frequent words and convert the TDM matrix into data frame.

tdm2 <- removeSparseTerms(zoom\_tdm, sparse = 0.975)

tdm\_m <- as.matrix(tdm2)

tdm\_df <- data.frame(tdm\_m)

The Euclidean distance is calculated and then hierarchical clustering on TDM data frame was performed.

#euclidean distance is calcuated

tweets\_dist <- dist(tdm\_df)

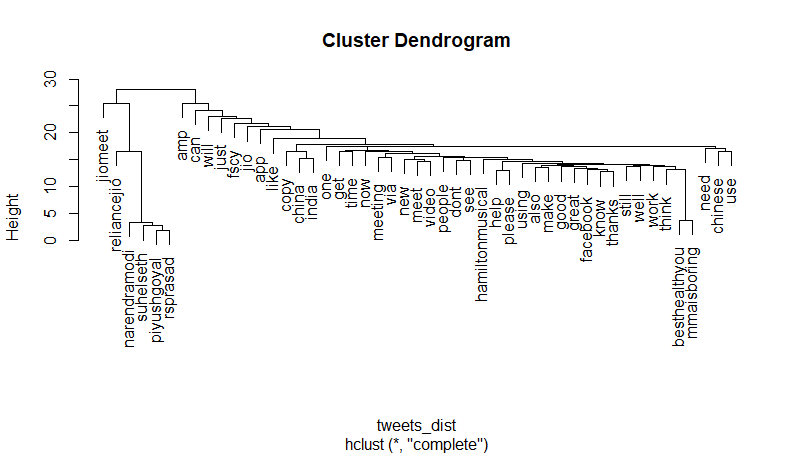
#hierarchical clustering

hc <- hclust(tweets\_dist)

The results of clustering can be viewed using dendrogram.

#plotting dendrogram

plot(hc)



Let’s understand the output

The output of the dendrogram depicts a logical association of words.

1. **JioMeet** and **Reliance** Jio are correlated since there were many tweets comparing Zoom\_us to Jiomeet which was launched recently
2. There is mention of 4 people in few of the tweets which the clustering algorithm has clubbed together, they are **Suhelseth, Narendarmodi, Piyushgoyal** and **RSPrasad**.
3. There is a large correlation of the recent **Indo**-**China** ripple which relates to tweets about banned applications.
4. There is a strong classification of people requesting for support with respect to usage of the application by using words like **Help** and **Please**.
5. **@MMAisBoring @besthealthyou** are clubbed together as these tweets have been reply to the same tweet.

# Word Association

Now we tried to create a few words association where words having close correlation is used in similar context.

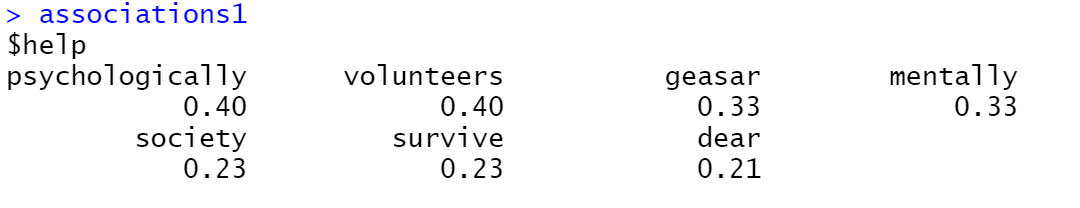
A correlation plot was plotted for the words associated with **Help** with a correlation limit of 0.2.

associations1 <-findAssocs(zoom\_tdm, "help", 0.2)

#view assocations

associations1

We see the following words are used along with “help”. The number indicates the correation – the higher the numeber more is the chance of these two words coming together.



Let’s plot the association

library(qdapTools)

library(ggplot2)

associations\_df1 <- list\_vect2df(associations1)[,2:3]

ggplot(associations\_df1, aes(y=associations\_df1[,1])) +geom\_point(aes(x=associations\_df1[,2]),data=associations\_df1, size=3)

A close up of a piece of paper

Description automatically generated

It can be seen the words volunteers and psychologically have a strong correlation to the word Help.

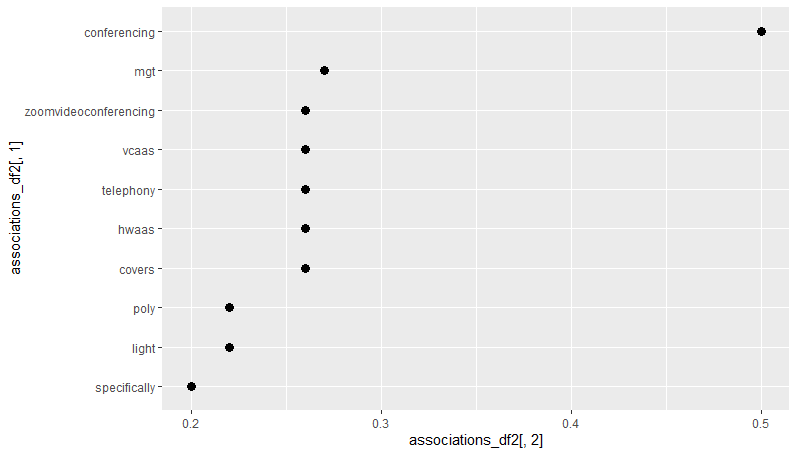
Another association graph was plotted for the word video.

#one more word association for "video"

associations2 <-findAssocs(zoom\_tdm, "video", 0.2)

associations\_df2 <- list\_vect2df(associations2)[,2:3]

ggplot(associations\_df2, aes(y=associations\_df2[,1])) +geom\_point(aes(x=associations\_df2[,2]),data=associations\_df2, size=3)



Video seems to be used most with conferencing, mgt, zoomvideoconferencing - which makes logical sense

# Summary

From the tweets downloaded following are observations

* The overall image of zoom seems to be not negative; a caution has to be applied while trusting the results of the sentiment analysis as it fails to understand the context as explained in the above section.
  + From the brand’s perspective advance analysis such filtering on specific comparison like “Jio” and should be done.
* The comparison of JIO meet and Zoom was pretty evident, in which Zoom is considered as inspiration for JIO meet
* We also had tweets associating it with the recent China apps ban in India, the image of Zoom not been seen as Chinese product is what the marketing team at Zoom may also be concentrating on.