**BUSINESS ANALYTICS AND BUSINESS INTELLIGENCE**

**CAPSTONE PROJECT | GROUP 1**

**INTERIM REPORT**

**20 - Jan - 2020**

**PROJECT ON RATING/SCORING OF DIFFERENT PRODUCTS BASED ON REVIEWS**

**Under Guidance of Ms Anjana Agrawal**

**Under Taken By**

Abhilash R

Angela Susan Mathews

Ashwin. S

Deepthi H

Sajin Madhavan S

Sneha Devadass

**TABLE OF CONTENT**

[**1. INTRODUCTION :**](#_axrg4cn45l14) **4**

[**2.DATA COLLECTION TECHNIQUES**](#_lgjoce67qcy7) **4**

[2.1 Purpose :](#_nft6z6gojy19) 4

[2.2 Sources Of Data :](#_u7cv2aorye1y) 4

[2.3Techniques Used to scrape the data:](#_vqkzxq1g5fyb) 4

[2.3.1.Web Scraper browser extension (mouthshut.com, official websites,gadgets.ndtv.com) :](#_7itm3791b9q9) 4

[2.3.2 API (For Twitter) :](#_b7xg7km4vhik) 5

[2.3.3 Inspect elements —> curl command —> python script (to scrape facebook data):](#_87n3gv1qdxyt) 6

[Limitations :](#_q5o333iexaw5) 6

[Methods used to overcome limitations :](#_rcpezcv9qbq6) 6

[**3.DATA CLEANING :**](#_mllg0gptk0de) **6**

[3.1 Cleaning Data using Excel :](#_9z53hvz53rkl) 6

[Step 1 : Removing null rows](#_99rktcnjynxd) 6

[Step 2 : Formatting “dateofreview” column to one single format](#_fens22oo18pf) 7

[Step 3 : Clustering the products into 3 common “categories”](#_z30ypar8f73m) 7

[3.2 Data Conditioning using Text Mining Package:](#_phfhtsl7o4yv) 7

[3.2.1 Observation on reading the dataset in R”](#_7tmdaxx2s5nb) 7

[Inference :](#_96xegb2ej0nv) 8

[3.2.2 Convert dataset to Corpus using tm text mining package:](#_offdvhgnf2qt) 8

[3.2.3 Convert document to lowercase :](#_jam7ibtg8aw6) 8

[3.2.4 Remove Punctuations (comma, hyphen, period):](#_nkrxq63zjrn8) 8

[3.2.5 Remove Stopwords :](#_mhqisck4bfp8) 9

[3.2.6 Remove Numbers:](#_rrv1psevbm0h) 9

[3.2.7 Remove Extra whitespaces:](#_qjor9ytexi33) 9

[3.2.8 Stem the document:](#_aovyr1iq0x4n) 9

[**4.EXPLORATORY DATA ANALYSIS**](#_bgi6b9jrpsgy) **10**

[4.1 Analyze the structure of dataset:](#_o8b0asupa41t) 10

[Inference:](#_fx3je4qrtlwh) 10

[Action:](#_cca6ww6k5y00) 10

[4.2 Summary of dataset:](#_6r9oilra4m3f) 10

[Inference:](#_i3v7w36otlml) 11

[Action:](#_20xo77yjko0w) 11

[4.3 Most Frequently Occurring Words :](#_hjs7zj9on2r3) 12

[Inference:](#_uynpk1tw2128) 12

[Action:](#_5wgytuua8nh3) 12

[4.4 Word Cloud :](#_wul6idanwg67) 13

[Inference:](#_k4ad2yaymcc8) 13

[4.5 Bag of n-Grams :](#_9d46ozp2x9k9) 13

[4.5.1 Creating Unigram:](#_f3i6vwge6d2g) 14

[4.5.2 Top 20 frequently occurring unigrams extracted :](#_e8hbdduogds4) 14

[Output:](#_vu6a666buk2m) 14

[4.5.3 Word Cloud with unigrams extracted:](#_lpwv14tx8ras) 15

[4.5.4 Interpret the word cloud with a bar graph for the unigrams extracted:](#_jybe5z3074i0) 15

[4.5.5 Creating Bigram:](#_wo3rkmf41ous) 16

[4.5.6 Top 20 frequently occurring sequence of bigrams :](#_57b3y1l6a7e9) 16

[Output :](#_lp8tw7ljo08a) 17

[4.5.7 Word cloud for bigrams extracted :](#_wm6cbgpgc17v) 17

[4.5.8 Interpret the word cloud with a bar graph for the bigrams extracted:](#_x6daaab1m9ta) 18

[4.5.9 Creating Trigram :](#_mrmwf9ku4qfp) 19

[4.5.10 Top 20 frequently occurring sequence of Trigrams :](#_yxwaphxpftll) 20

[4.5.11 Bar graph for the bigrams extracted:](#_lp1ljkcn4mk) 20

[**5.MODELLING TECHNIQUE:**](#_g9a1z2i9be1y) **21**

[5.1 Sentiment Analysis](#_xw3nvigrtlcl) 21

[5.2 Sentiment scores for each review:](#_5nb133uymf8h) 21

[5.3 Summary of Sentiment Scores obtained:](#_63hwcxo09yzj) 22

[Inference:](#_1d49j4o3vqph) 22

[Action:](#_w638yrexemde) 22

[5.4 Histogram :](#_mz1zg46rzlp2) 22

[Inference:](#_o4yy9t4ez5nt) 23

[5.5 Histogram to Visualise the positive,negative,neutral spread of reviews:](#_5nxpwuh2s7ge) 24

[Inference:](#_dp7kvk6331ry) 25

[5.6 Split Dataset For Modelling:](#_59qpogok4kfw) 25

[5.7 Random Forest Model :](#_ps4xy2f1dlkf) 26

[Inference:](#_lxi5huy33s52) 27

[5.8 Further Modelling Plans to be carried out:](#_iqx8qwjpw1uf) 27

# **1. INTRODUCTION :**

The project aims to perform a behavioral analysis of web and social media-based reviews of products. A prediction algorithm for the dataset using Natural Language Processing is modelled. This report captures a summary statistic of the dataset and brings out insights from the Exploratory Data Analysis. The report concludes with a brief explanation of the modelling techniques that is being used for prediction.

The reviews for the various products (brand and category wise) were extracted from different social media platforms like Twitter, Youtube, Mouthshut, Croma, etc. Our objective was to perform a sentimental analysis on the reviews collected and thereby classify the products as “Good”, “Bad” and “Neutral”.

In the dataset, the various variables extracted were, the Brand, Category of the products, Dates on which the reviews were written and the Reviews for the product.

# **2.DATA COLLECTION TECHNIQUES**

## **2.1 Purpose :**

To collect the review comments of products from social media and official websites to perform analysis on the products based on the collected reviews.

## **2.2 Sources Of Data :**

1. Twitter
2. Facebook
3. Mouthshut.com
4. Official websites of the respective product and brand
5. [gadgets.ndtv.com](https://gadgets.ndtv.com/)

## **2.3Techniques Used to scrape the data:**

### 2.3.1.Web Scraper browser extension (mouthshut.com, official websites,gadgets.ndtv.com) :

Downloaded and installed the google’s web scraper browser extension.

**Create sitemaps:**

1. Here we specify the url of a product for which we will be scraping the data.

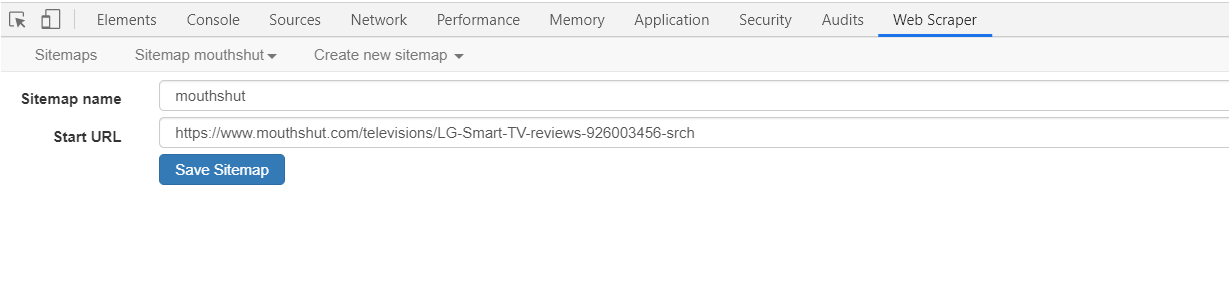


FIG 2.3.a

1. We created the sitemap where we have specified the details of variables like review comments and date. Also we have taken care of pagination.

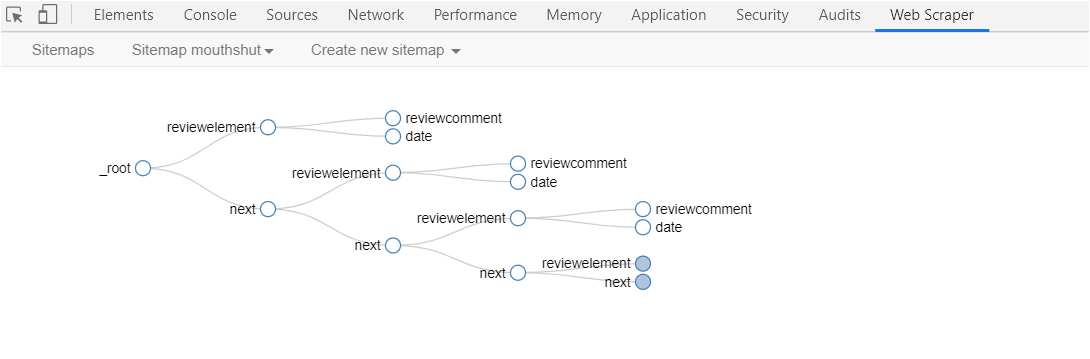


FIG 2.3.b

1. Output :

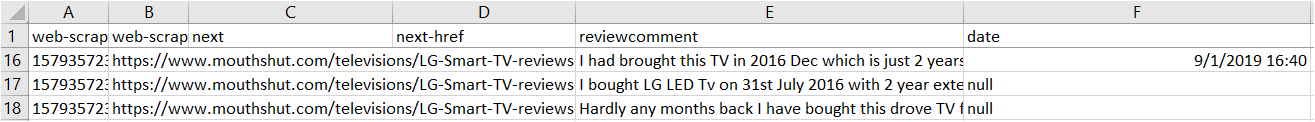


FIG 2.3.c

### 2.3.2 API (For Twitter) :

We used R language and twitter API to scrape data from twitter on particular products. Here shared the screenshot of one among them.

To scrape reviews or comments on oneplus mobiles

****

FIG 2.3.2

### 2.3.3 Inspect elements —> curl command —> python script (to scrape facebook data):

#### **Limitations :**

* Multiple requests from the same IP.
* Web scraping bots fetches data very fast which made website become unresponsive.

#### **Methods used to overcome limitations :**

* Rotating IPs using VPN.
* Added sleep calls and some delays after crawling

# 

# **3.DATA CLEANING :**

The data collected from various sources were collated in Excel.

Dataset contains 4 variables.

* Brand
* Category
* Review
* dateofreview

## **3.1 Cleaning Data using Excel :**

## **Step 1 : Removing null rows**

Skimming through the dataset, though filter,a lot of null values(in reviews column) were found and removed that particular row.

## **Step 2 : Formatting “dateofreview” column to one single format**

The dates extracted from different websites were of different formats , so the format has been made unique to all the rows

Due to lots of null values,**dateofreview** column has been dropped from the dataset.

## **Step 3 : Clustering the products into 3 common “categories”**

For each brand we have decided to collect different products and its reviews.

For example : For the brand “samsung”, we have collected the reviews for mobiles, televisions,washing machines. Here we have grouped mobile phones into “mobiles”, “televisions,washing machines” into “Home Appliances” and “fitband, watches” into “electronic gadgets” product categories.

So we have clustered various products into 3 categories which are “Mobiles”,”Home Appliances”,”Electronic Gadgets”.

Here is the overview of dataset.

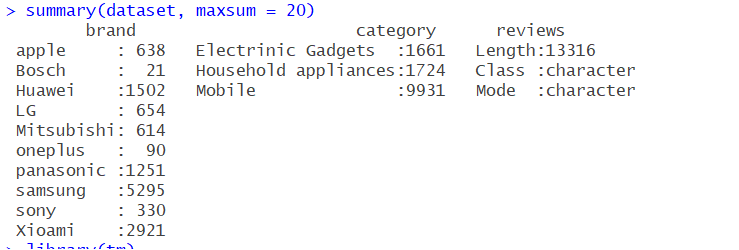


FIG 3.1

After performing these basic cleaning of data in Excel, the dataset was loaded into R to perform further cleaning of the data and analysis

## **3.2 Data Conditioning using Text Mining Package:**

### 3.2.1 Observation on reading the dataset in R”



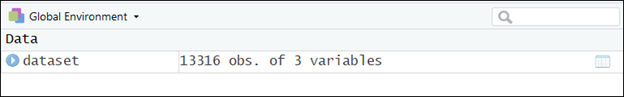


FIG 3.2.1

### Inference :

There are 3 columns and 13316 observations.

### 3.2.2 Convert dataset to Corpus using tm text mining package:

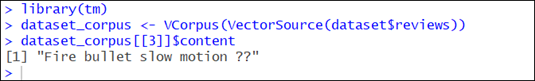


FIG 3.2.2

To perform text mining, the tm\_map() function is used.

### 3.2.3 Convert document to lowercase :

First, we convert the entire document to lowercase to avoid situations where for example, the model might treat a word which is in the beginning of a sentence with a capital letter different from the same word which appears later in the sentence but without any capital letter.



FIG 3.2.3

### 3.2.4 Remove Punctuations (comma, hyphen, period):



FIG 3.2.4

### 3.2.5 Remove Stopwords :

These are extremely common words like “and”, “or”, “not”, “in”,etc are removed. They are removed from the corpus as they provide no meaning or context for analysis.

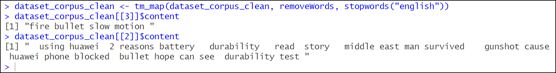


FIG 3.2.5

### 3.2.6 Remove Numbers:



FIG 3.2.6

### 3.2.7 Remove Extra whitespaces:



FIG 3.2.7

### 3.2.8 Stem the document:

As the final step of text preprocessing, stemming of the document is done. A word is replaced with its most basic conjugate form in stemming. For example, the stem of the word “typing” is “type”. We do this as we don’t want the words “type” and “typing” to convey different meanings to the algorithm.



FIG 3.2.8

# **4.EXPLORATORY DATA ANALYSIS**

## **4.1 Analyze the structure of dataset:**

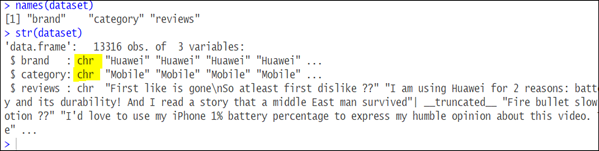


FIG 4.1a

### Inference:

“Brand” and “Category” variables are in character data type which has to be factor variable.

### Action:

Converted “brand” and “category” to factor variable.

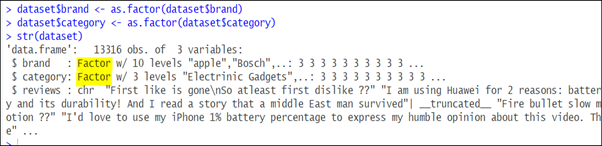


FIG 4.1b

## **4.2 Summary of dataset:**

**Summary functions** produce a **summary** of all records in the found dataset, or subsummary values for records in different groups.

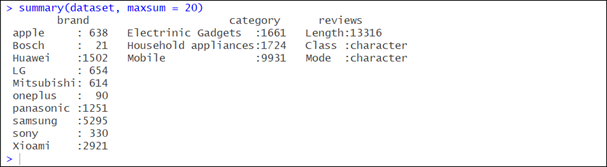


FIG 4.2

### Inference:

**Brand wise inference -** There is no sufficient data for “Bosch” and “oneplus” to analyse, and build a model.

**Category wise inference -** Data is biased as there are more data for mobiles but not to other categories.

### Action:

We need to remove those brands from the dataset and start performing sentiment analysis and apply modelling techniques on the updated dataset.

When coming to categories, collecting data for other categories would be the best solution but as there is no dependencies between the categories, having a good number of data in each category would make the model better is what the call taken.

## **4.3 Most Frequently Occurring Words :**

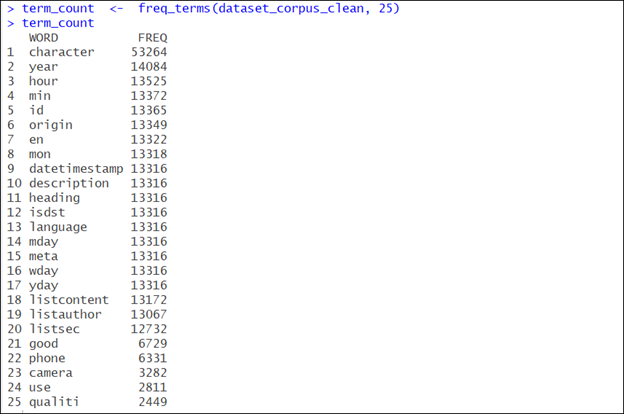
Top 25 most frequently occurring words and their frequencies are listed below.

FIG 4.3a

### Inference:

Words like “ year”,”hour”,”min”,”en”,”yday”,”wday”, etc.. seems to be unrelated or not much useful words.

### Action:

Using custom stop words, unwanted words has to be removed from the corpus.



FIG 4.3b

## **4.4 Word Cloud :**

A word cloud is constructed to visualise the frequencies of all the words that occur in the dataset. The larger the size, the higher the higher the frequency.

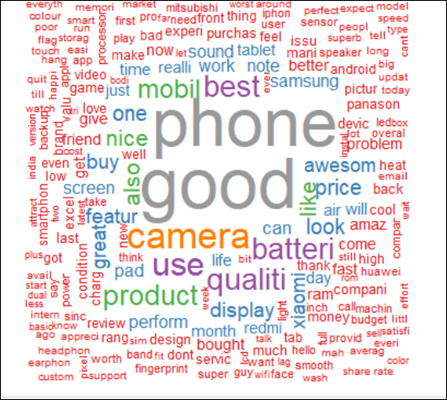


FIG 4.4

### Inference:

From the above cloud it is clear that the words “phone”, “good”,”camera”,”best”,”battery”,”sound”,”display”,etc has been used widely in the reviews shared by customers.

This infers that the dataset may give much **insights on the specifications of the products.**

## **4.5 Bag of n-Grams :**

To understand the data even better a sequence of N-words are extracted . This helps in deciding which sequence can be merged together as one word.

### 4.5.1 Creating Unigram:

Function is defined to call unigram words and then created a document term matrix from the unigram.

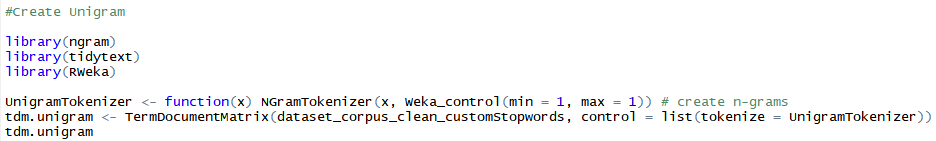


FIG 4.5.1

### 4.5.2 Top 20 frequently occurring unigrams extracted :



FIG 4.5.2a

#### **Output:**

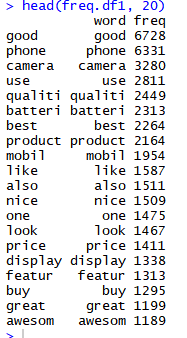


FIG 4.5.2b

### 4.5.3 Word Cloud with unigrams extracted:

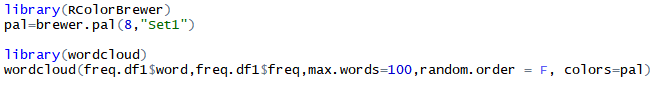


FIG 4.5.3a

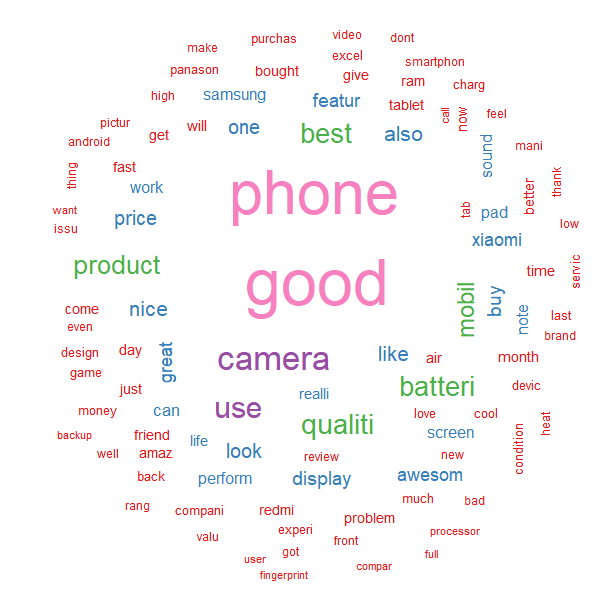


FIG 4.5.3b

### 4.5.4 Interpret the word cloud with a bar graph for the unigrams extracted:

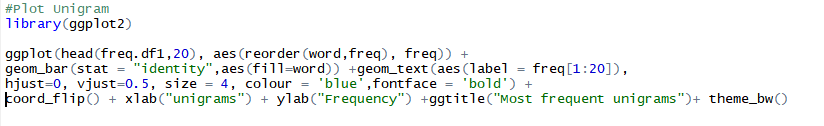


FIG 4.5.4a

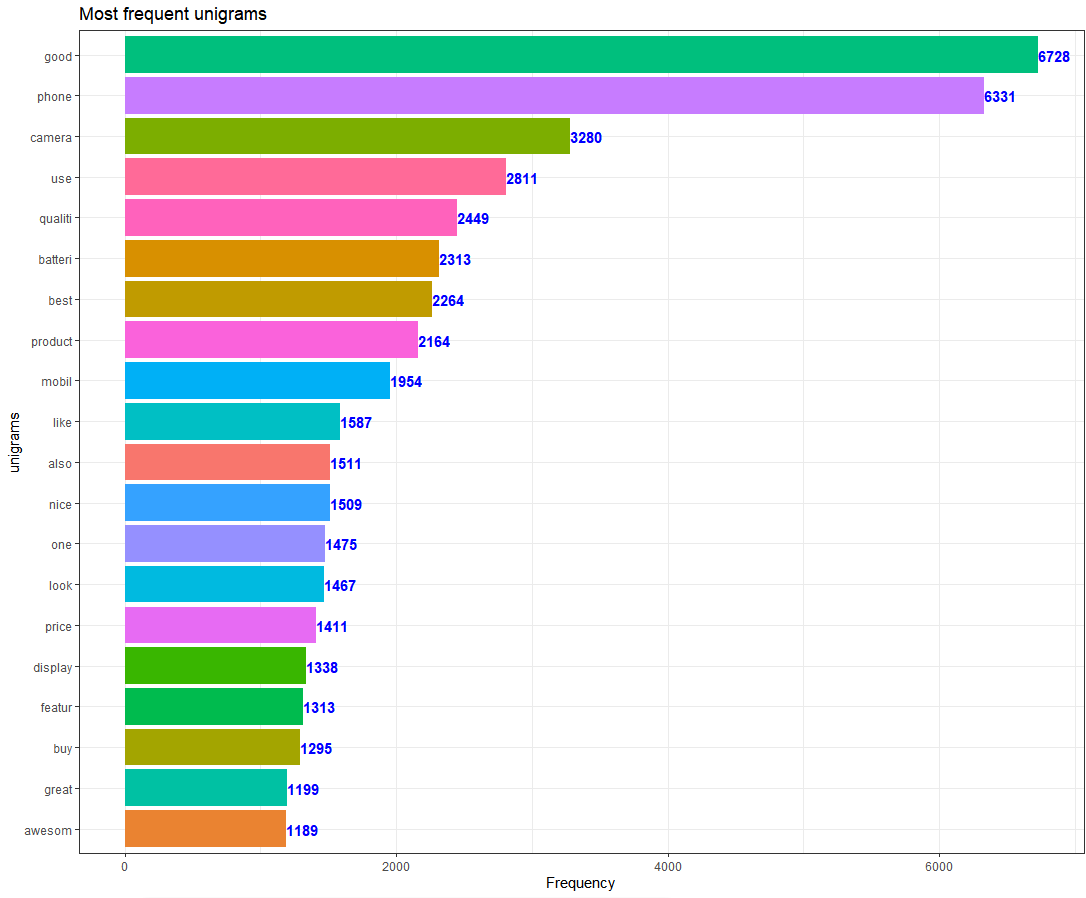


FIG 4.5.4b

### 4.5.5 Creating Bigram:

A 2-gram sequence of words are extracted to create a Bigram



### 4.5.6 Top 20 frequently occurring sequence of bigrams :



FIG 4.5.6a

#### **Output :**

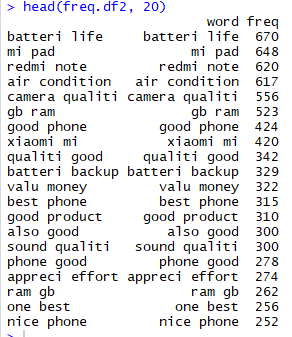


FIG 4.5.6b

### 4.5.7 Word cloud for bigrams extracted :



FIG 4.5.7a

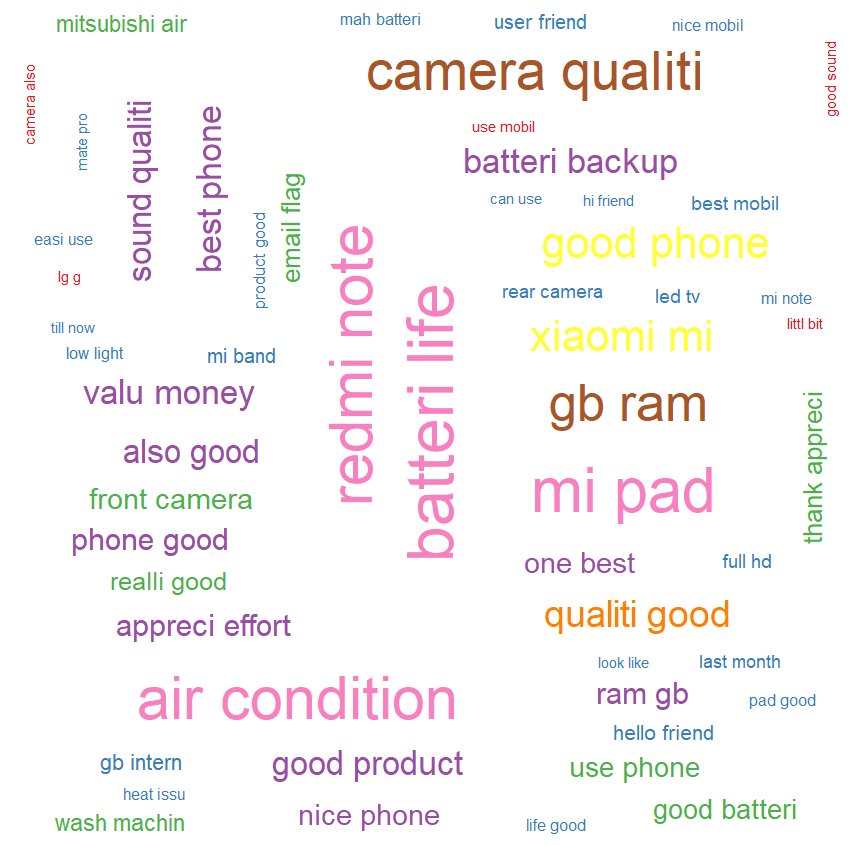


FIG 4.5.7b

### 4.5.8 Interpret the word cloud with a bar graph for the bigrams extracted:

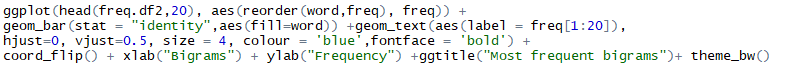


FIG 4.5.8a

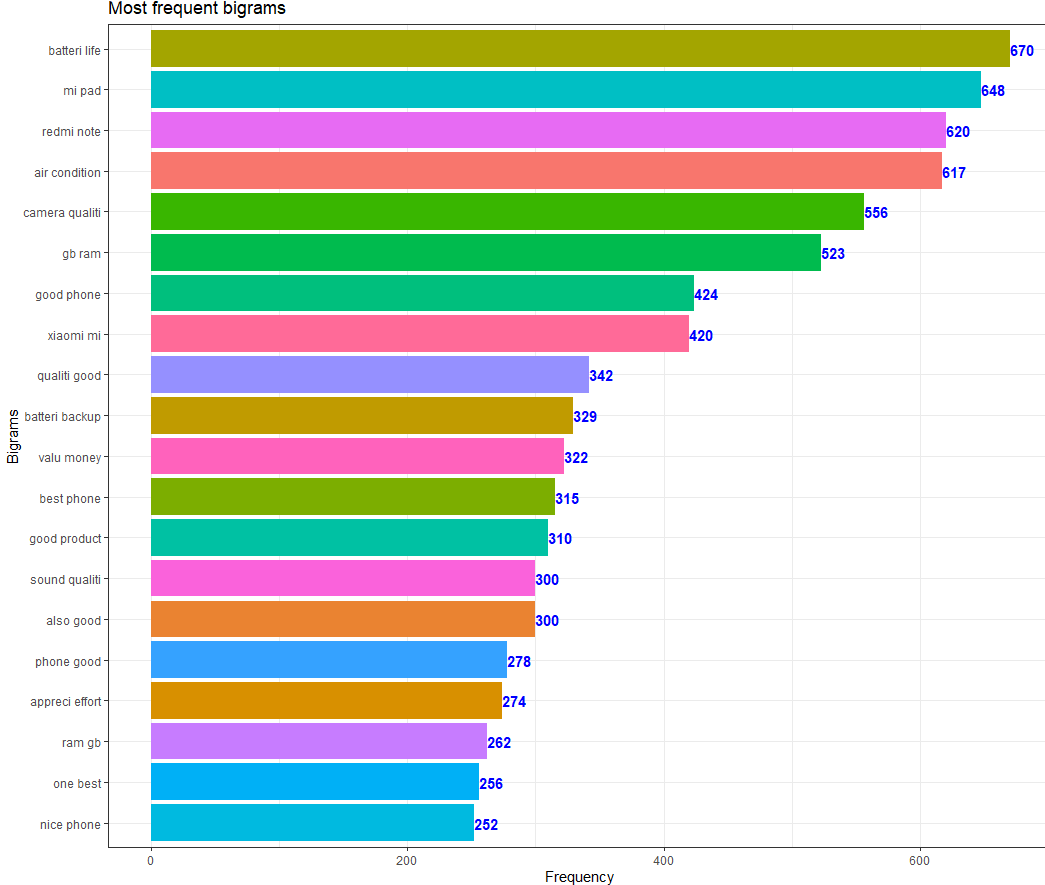


FIG 4.5.8b

### 4.5.9 Creating Trigram :

A 3-gram sequence of words are extracted to create a Bigram

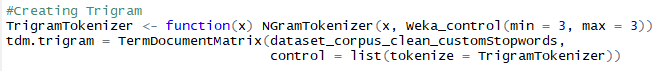


FIG 4.5.9

### 4.5.10 Top 20 frequently occurring sequence of Trigrams :





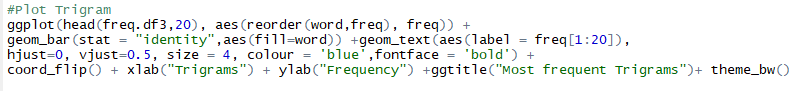


FIG 4.5.10

### 4.5.11 Bar graph for the bigrams extracted:

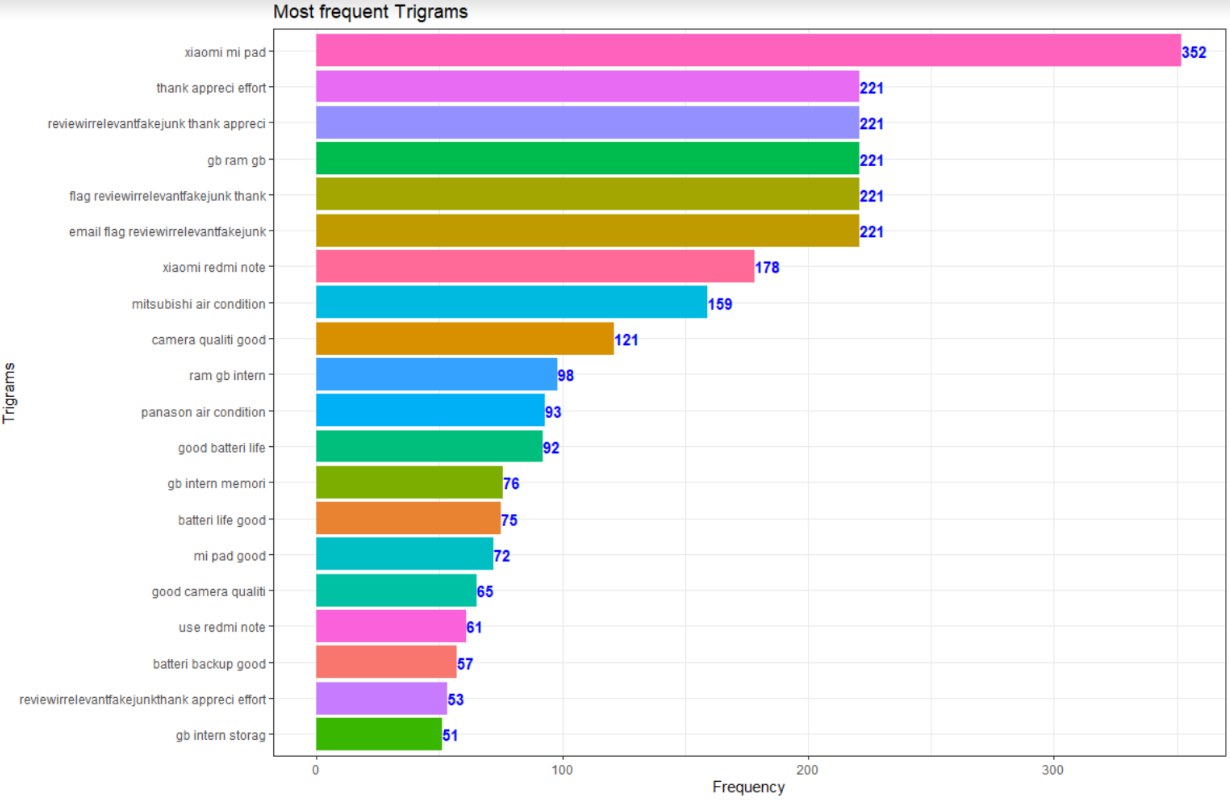


FIG 4.5.11

# **5.MODELLING TECHNIQUE:**

## **5.1 Sentiment Analysis**

The reviews are bought in to a clean format to perform sentiment analysis.

First, “**Review**” column is converted to a character data frame.

****

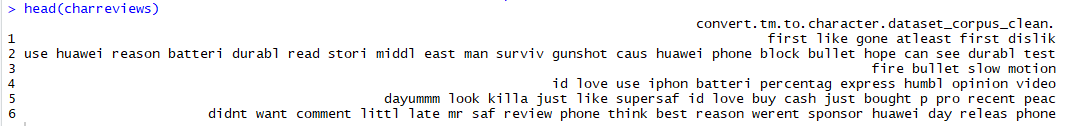
****

FIG 5.1a

Afterwhich, the column is merged to the final dataset on which analysis is done.

****

****

****

FIG 5.1b

## **5.2 Sentiment scores for each review:**

In order to perform sentiment analysis, and extract the sentiment scores for each reviews from the dataset, **“Sentimentr”** package is installed.

**Sentiment(**) function is applied on the character vector of reviews, which in turn returns the text polarity sentiment in a sentence level.

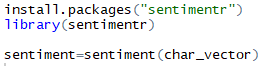
****

FIG 5.2

## **5.3 Summary of Sentiment Scores obtained:**

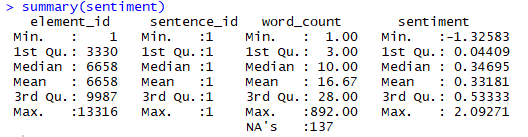
****

FIG 5.3a

### Inference:

Range of sentiment scores is between **-1.32583 to 2.09271**.

On an average, the scores are seen to fall around 0.33181.

### Action:

The sentiment scores are now grouped into 3 categories - **Positive**, **Negative** and **Neutral** comments for better analysis.

The categorization parameter is taken based on the mean of the sentiment scores (0.34).

Hence, Positive group includes all values greater than 0.34,

Negative group consists of all values less than 0, and

Neutral group consists of the values under the cap of 0 and 0.34.

The sentiment scored extracted are now merged to the final dataset data frame to proceed with further analysis.

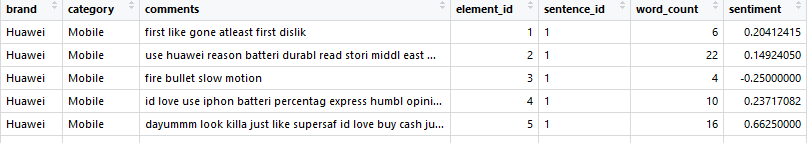


FIG 5.3.b

## **5.4 Histogram :**

To visualize thetext polarity sentiment scores.

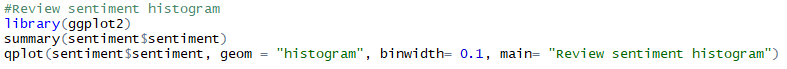
****

FIG 5.4a

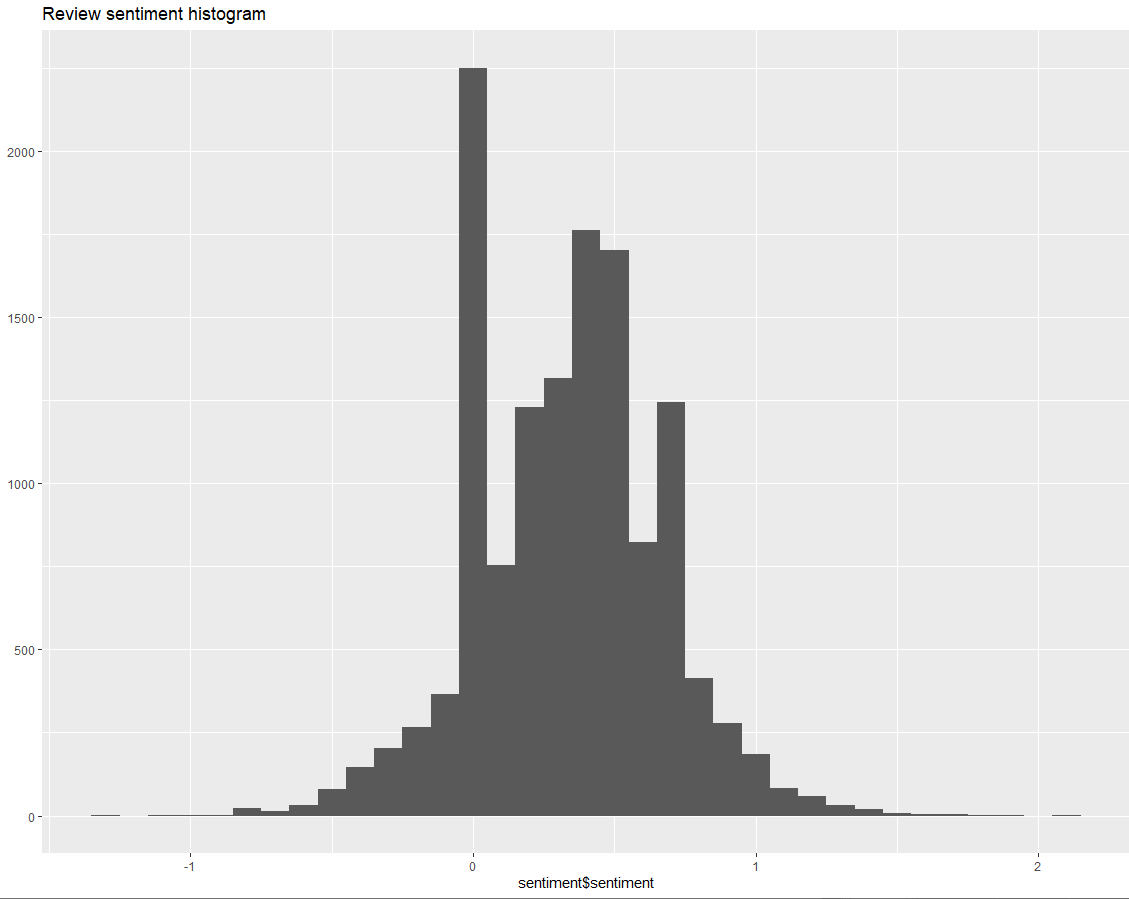
****

FIG 5.4b

### Inference:

Scores get overpopulated in the center of the histogram and fall under the 0-1 range of values which means neutral reviews are much in dataset.

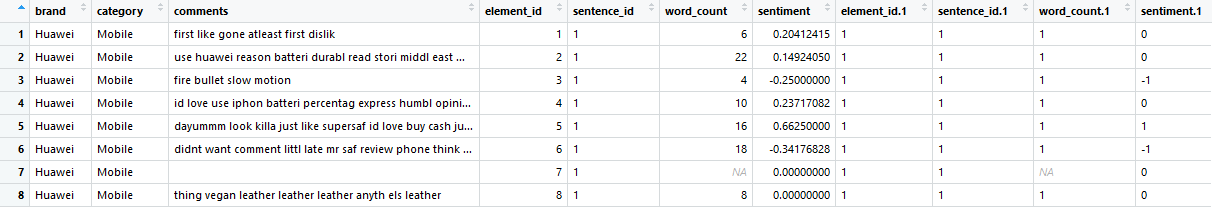
****

FIG 5.4c

As seen in the above compiled dataset, row 7 is null, and such empty values would be considered as meaningless values for our analysis.

Hence, all such NULL values are **omitted**.

****

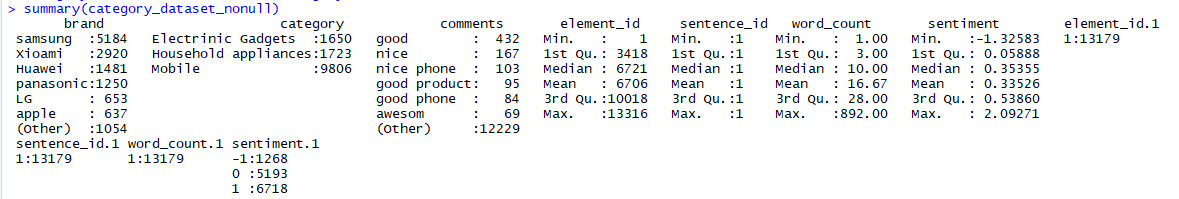
****

FIG 5.4d

## **5.5 Histogram to Visualise the positive,negative,neutral spread of reviews:**

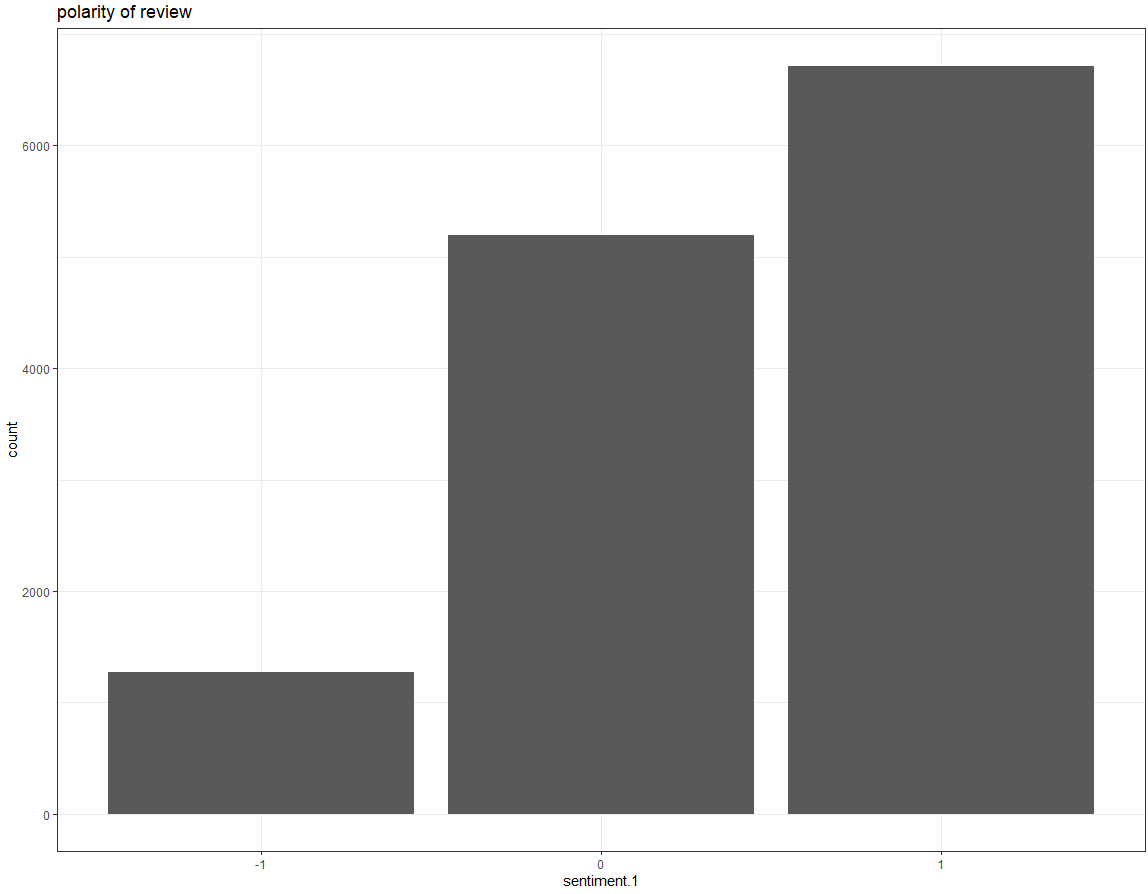
****

FIG 5.5

### Inference:

The majority of values fall under the **“Positive”** category, greater than 0.34.

This indicates that the dataset consists of more Positive Reviews followed by Neutral and Negative Comments.

## **5.6 Split Dataset For Modelling:**

Complete dataset is split into train and test set on 80-20 ratio to apply the modelling techniques.

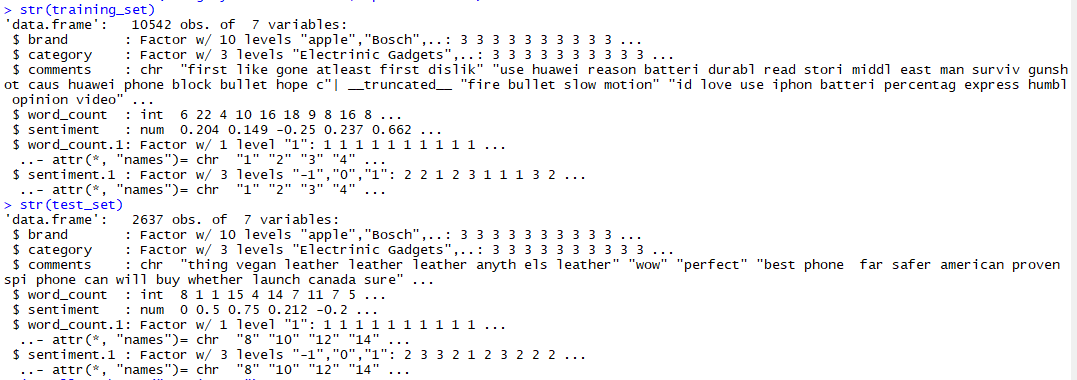
****

FIG 5.6

## **5.7 Random Forest Model :**

**Random Forest classification model** is first tried to predict the class of the recipient variable.

Parameter **mtry**, number of variables randomly sampled at each split is first considered as 4.

Parameter **ntree**, number trees to grow is first considered as 601.

Basic Random forest model is first applied to the train dataset before tuning the parameters and so forth.

And once the dataset is trained, the model is applied to the test data, which provides us with the below predictions and its respective confusion matrix.

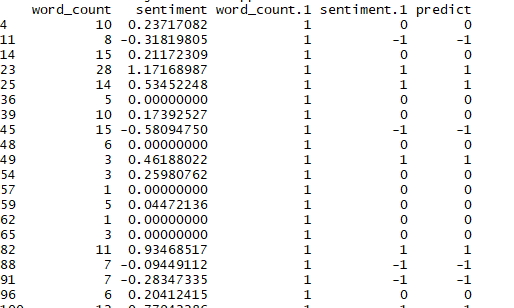


FIG 5.7a

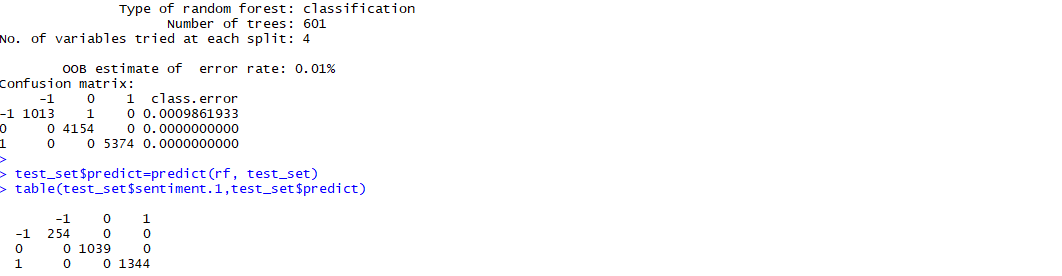


FIG 5.7b

### Inference:

Expected results are not derived from this basic model. Further tuning and following modelling in Random Forest is to be applied.

## **5.8 Further Modelling Plans to be carried out:**

1. Decided to work on below mentioned models.

* Naive Bayes Algorithm
* Decision Tree Algorithm
* Random Forest Algorithm

1. Comparing Accuracy, Precision, Recall, and F1 Scores, we will be arriving at the best model which gives us the scores/rating for each product category of a particular brand.

# **APPENDIX A**

1. <https://chrome.google.com/webstore/detail/webscraper/jnhgnonknehpejjnehehllkliplmbmhn?hl=en>
2. <https://rtweet.info/>
3. <https://www.pluralsight.com/guides/machine-learning-text-data-using-r>

# **APPENDIX B**

Attached the raw code

