BUSINESS INTELLIGENCE AND DATA MINING (BIDB17-4)

TERM 4

"Assessing Consumer Financial Complaints-Driving insights using Sentiment Analysis"

GROUP B16

Ritwika Majumdar	*B17162*
Sreela Basu Roy	*B17174*
Priyadarshi Dasgupta	*H17155*
Sanchit Agarwal	*H17163*

Business Management
XLRI Jamshedpur

Contents

INTRODUCTION

In this project we are analysing the customer complaints for various financial services such as Credit card or prepaid card, mortgage, vehicle loan or lease, debt collection, checking or savings account and so on received in 2018 in USA. The dataset was obtained from the website (https://catalog.data.gov/) which has public records made available.

Data Source: https://catalog.data.gov/dataset/consumer-complaint-database

The data acquired was cleaned to arrive in the format:

Date received	The date that the complaint was received on
Product	The type of product/service the customer identified
Sub-product	The type of sub product identified by the customer
Issue	The issue identified by the customer in the complaint
Sub-issue	The sub issue identified by the customer in the complaint
Consumer complaint narrative	Complaint recorded verbatim in text format
Company public response	The company's public response to the consumer complaint
Company	The company the complaint is against
State	Customer's state as per mailing address
ZIP code	Customer's ZIP code as per mailing address
Tags	Data that supports easier searching on behalf of consumers
Consumer consent provided?	If the consumer consents to publishing the complaint narrative
Submitted via	Medium of submission of the complaint
Date sent to company	The date that the complaint is forwarded to the company
Company response to consumer	How the company responded to the complaint
Timely response?	If the response by the company was given in appropriate time
Complaint ID	Unique identification number of a complaint

We will be doing sentiment analysis on the consumer complaints narrative to find out about the specific nature of complaints via text mining. Since these are complaints, most/all of the sentiments will be negative. However, we shall look at the magnitudes across categories to understand the dissatisfaction. We will also be looking at the frequencies of complaints based on different states and across different product categories to understand the major problem areas that a financial service provider faces.

METHODOLOGY

1. Data Preparation

The dataset received from the site required data cleaning and preparation before any analysis could be performed. The data was filtered for 2018 data, and irrelevant columns were removed. Rows with corrupted or blanks cells were also removed, thus bringing down 5 Lakh records to around 1 Lakh.

2. Sentiment Analysis

Sentiment analysis is the process of identifying and categorizing opinions expressed in a piece of text to determine broadly what the author's attitude towards that particular subject is. Lots of major companies employ algorithms to translate consumer feedback into sentiments for gaining insight into a consumer's thought process. This approach helps them to understand what the consumer expects and how product/service matches up to those expectations.

Generally, it is done as a binary distinction between positive and negative sentiments, but it may also be more fine-tuned to more specific sentiments felt by the author such as anger, fear, joy, etc. Of course, it is not always a 100% accurate, but it is a useful way to measure the emotions portrayed in large blocks of text.

The following libraries were loaded to help us perform this analysis:

```
library(tidyverse)
library(tidytext)
library(glue)
library(stringr)
```

The tidytext package performs sentiment analysis using the following method:

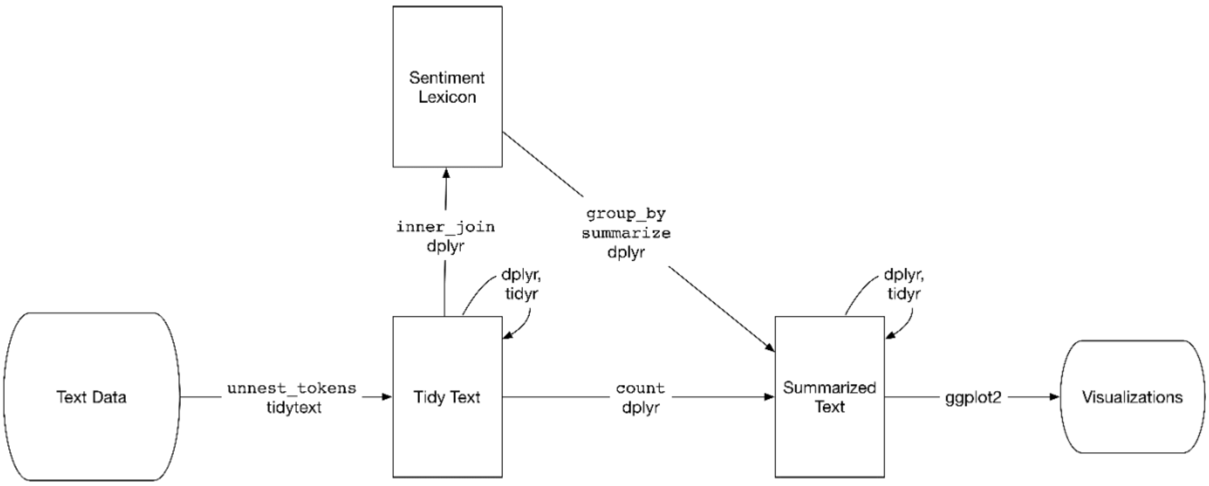


Figure: A flowchart of a typical text analysis that uses tidytext for sentiment analysis.

The code used on R to perform analysis for the same is as follows:

```
#Sentiment Analysis
#Extract complaint narrative
dataset_main1<-dataset_main[!(dataset_main$Consumer.complaint.narrative==""),]

#Find unique product types
product = unique(dataset_main1[c("Product")])

#Clean data for sentiment analysis and write complaints for each product to separate text file
CleanData = function(complaint, fileName){
  complaint = gsub("[[:punct:]]","", complaint)
  complaint = gsub("@\\w+","", complaint)
  complaint = gsub("[[:digit:]]","", complaint)
  complaint = gsub("  "," ", complaint)
  complaint = gsub("[ \t]{2,}"," ", complaint)
  complaint = gsub("XX"," ", complaint)
  complaint = gsub("XXX"," ", complaint)
  complaint = gsub("XXXX","", complaint)
  complaint = gsub("XXXXXXXX","", complaint)
  complaint <- tolower(complaint)
  complaint <- complaint[!is.na(complaint)]
  complaint <- gsub("\\s+", " ", complaint)
  complaint <- trimws(complaint)
  filePath = paste("input/",fileName,".txt", sep="")
  write.table(complaint, file=filePath,sep = "\t", row.names = FALSE)
}

#Loop through products and call CleanData function
for(i in 1:9){
  complaint= dataset_main2[,6]
  CleanData(complaint,product[i,1])
 # CleanData(sentiment_dataset,product[i,1])
}

#Store list of filenames (complaint files)
files <- list.files("C://Users/sreel/OneDrive/XLRI/Term 4/BIDM/Project/input")
```

```
# write a function that takes the name of a file and returns the # of postive
# sentiment words, negative sentiment words, and the difference
GetSentiment <- function(file){
  # get the file
  fileName <- glue("input/", file, sep = "")
  # get rid of any sneaky trailing spaces
  fileName <- trimws(fileName)

  df1 <- read.table(fileName, header=TRUE, fill = TRUE)

  # read in the new file
  fileText <- glue(read_file(fileName))
  # remove any dollar signs (they're special characters in R)
  fileText <- gsub("\\$", "", fileText)



  # tokenize
  tokens <- data_frame(text = fileText) %>% unnest_tokens(word, text)

  # get the sentiment from the first text:
  sentiment <- tokens %>%
    inner_join(get_sentiments("bing")) %>% # pull out only sentimen words
    count(sentiment) %>% # count the # of positive & negative words
    spread(sentiment, n, fill = 0) %>% # made data wide rather than narrow
    mutate(sentiment = positive - negative) %>% # # of positive words - # of negative owrds
    mutate(file = file) %>% # add the name of our file
    mutate(product = file) # add product



  # return our sentiment dataframe
  return(sentiment)
}

# file to put our output in
sentiments <- data_frame()

# get the sentiments for each file in our datset
for(i in files){
  sentiments <- rbind(sentiments, GetSentiment(i))
}

#plot overall sentiment for each product type
ggplot(sentiments, aes(x = file, y = sentiment, color = file)) +
  geom_hline(yintercept = 0, color = "red") +
  theme(axis.text.x = element_blank())+
  geom_boxplot() # draw a boxplot for each product
```

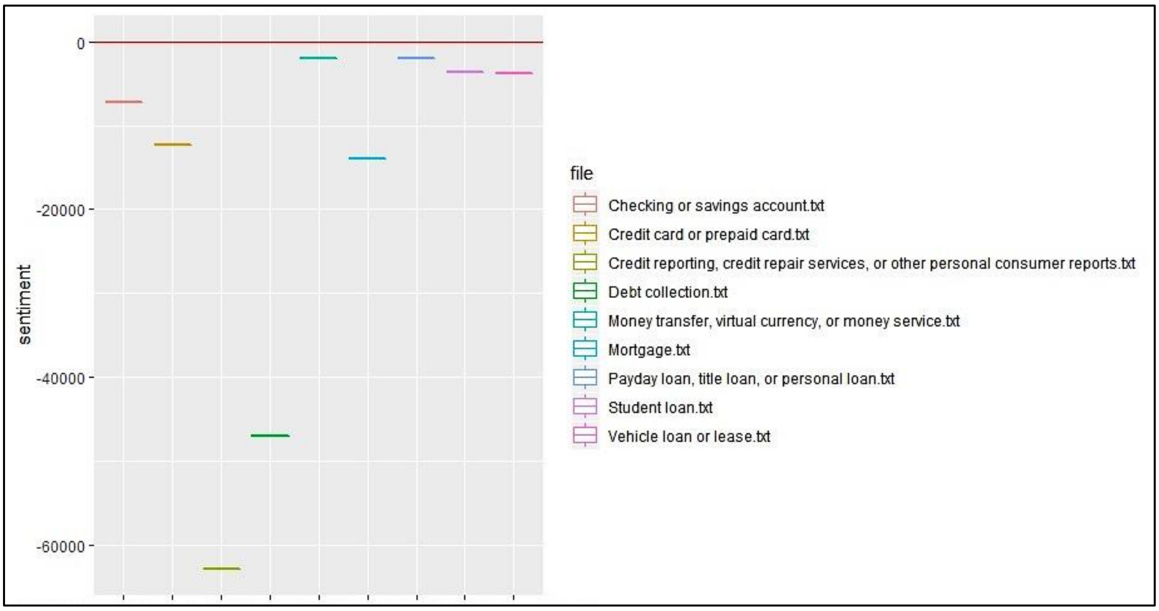
1.1. Observations

The **consumer complaint narratives** for each product type were analysed as follows:

1.1.1. Sentiment Analysis- This is the classification by emotion, each complaint is classified by emotion and this is the histogram representing each group of emotion, using the bing lexicon (part of the tidytext package)

1.1.2. Word Cloud- A word cloud represents all the words available in the database, the size of each word tells its frequency in the database, the bigger the word is, the more frequent that word appears in the database.

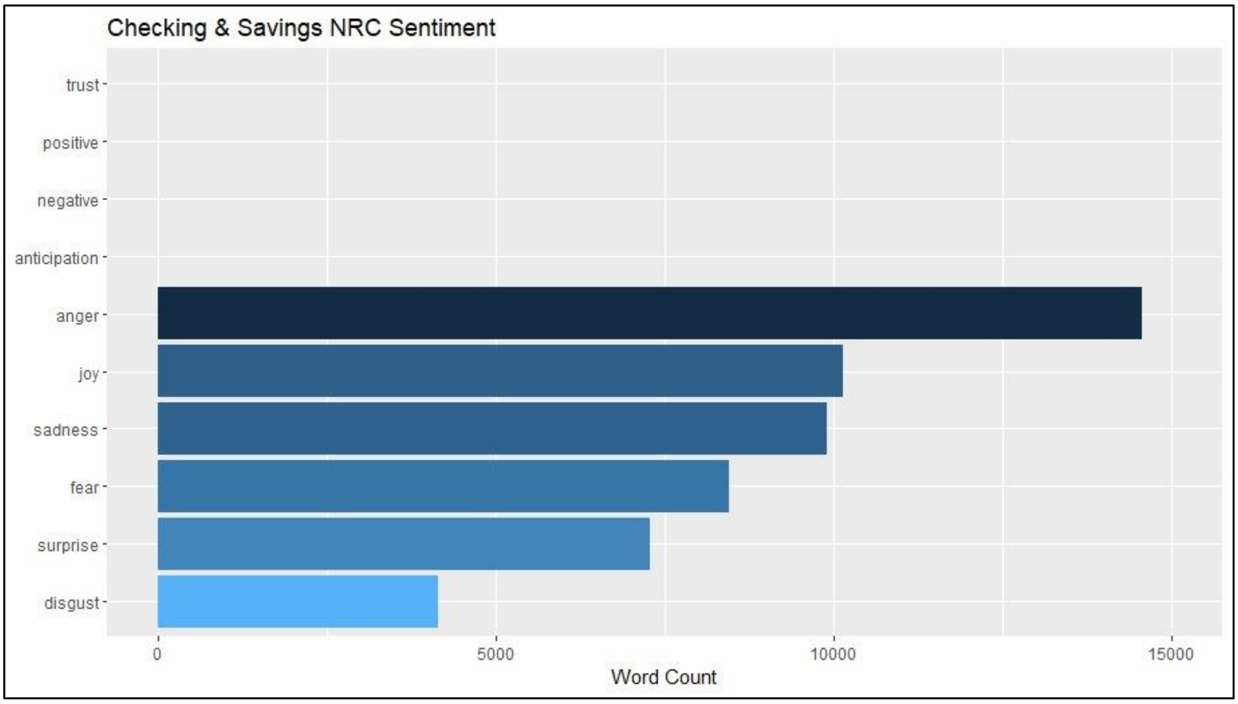
Furthermore, we did sentiment analysis and assessed word clouds on the complaints for each different type of financial service that is provided.



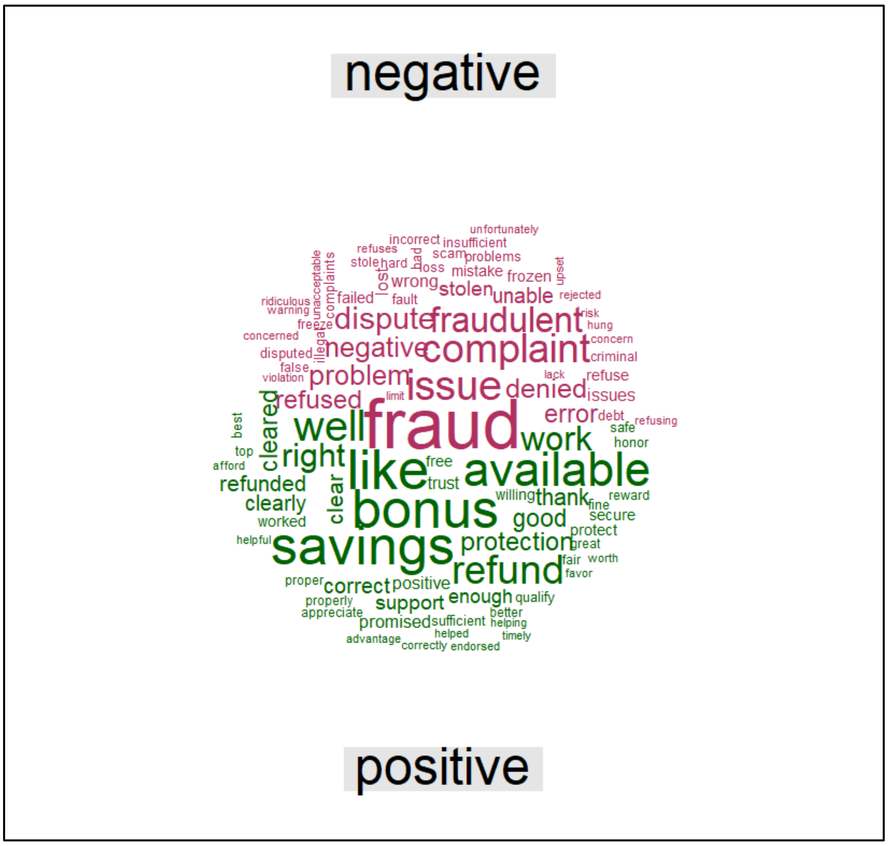
The sentiment analysis of the consumer complaint narratives using the **tidytext** packages shows that all the narratives are primarily negative, with Credit Reporting, credit repair services or other personal consumer reports being most negative. **The NRC lexicon** (part of the **tidytext** package) was used to classify the words in the narratives into positive and negative, and an overall score for each product type was arrived at.

Following are the sentiment analysis histograms and word clouds for complaints on each financial service:

A. For checking and savings accounts:

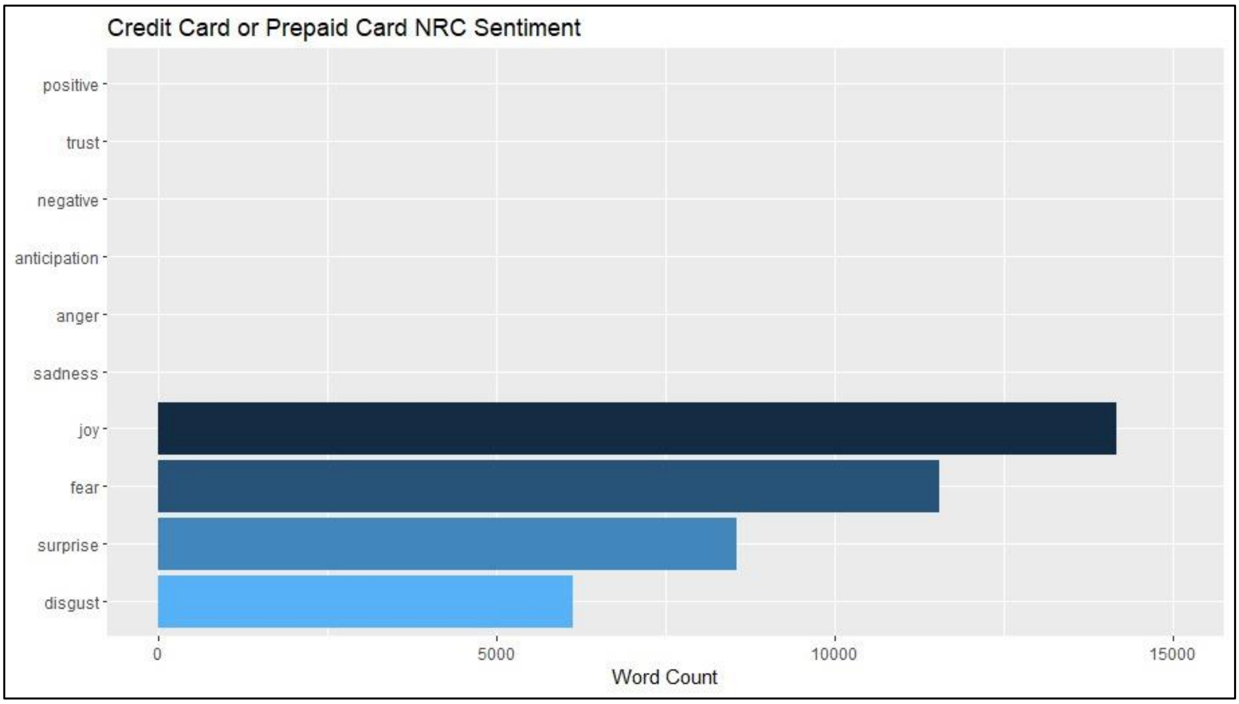


Here, anger is the emotion with most complaints followed by joy, sadness, fear and so on.

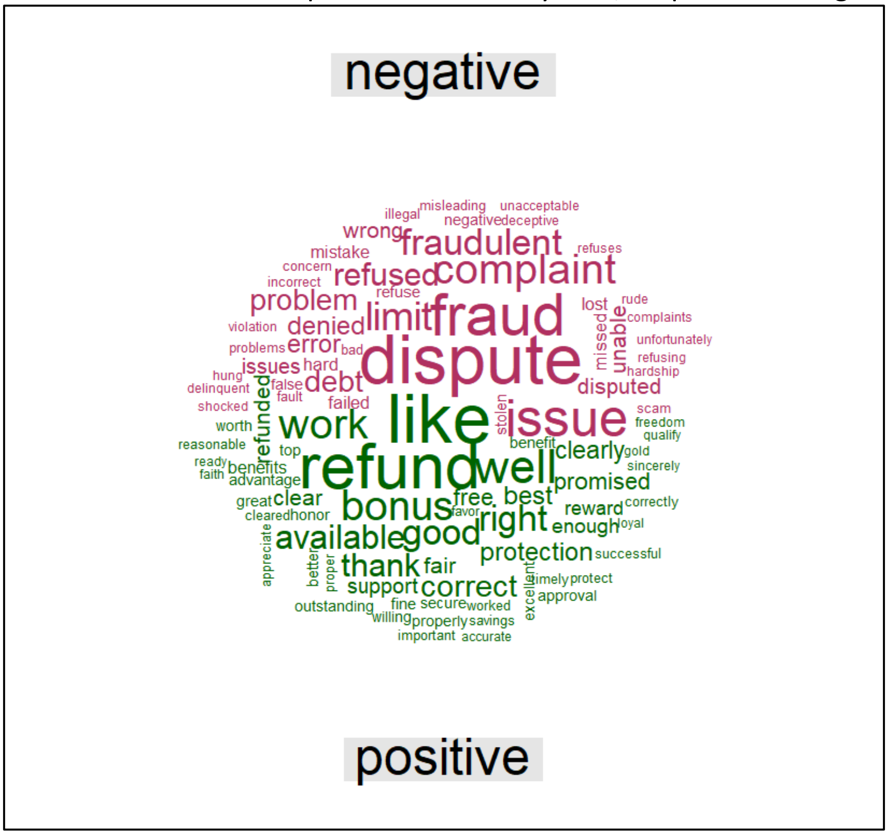


"Fraud" is the word with the highest frequency in our given database, with respect to the Checking and Savings Accounts.

B. For credit card or prepaid cards:

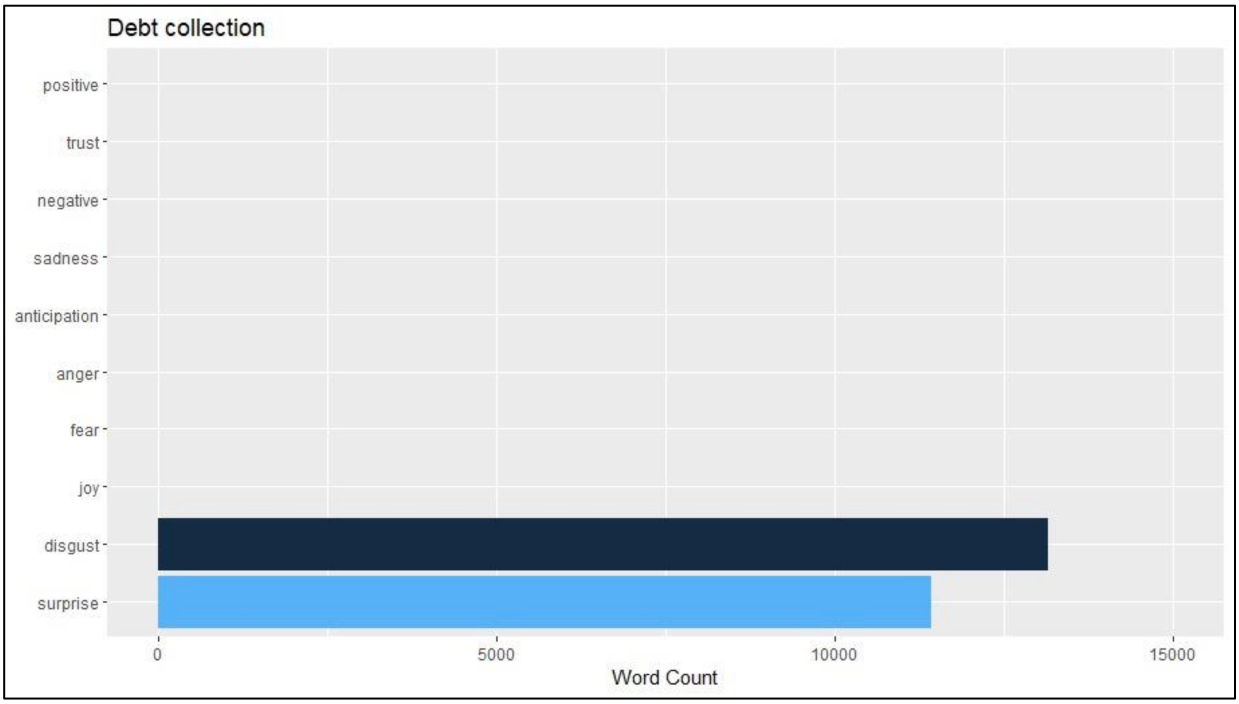


Here, joy is the emotion with most complaints followed by fear, surprise and disgust.

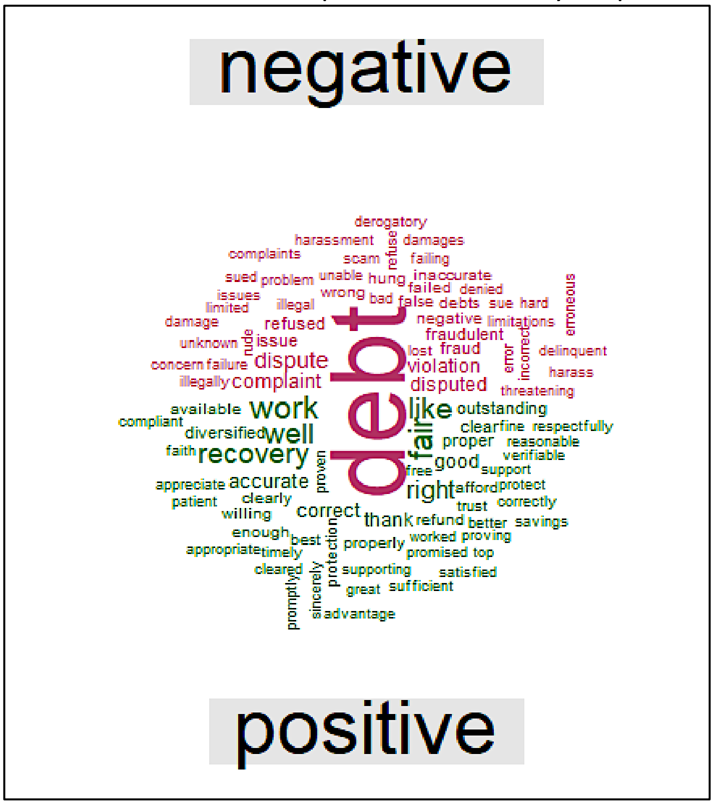


“Dispute” is the word with the highest frequency in our given database, with respect to the Credit or Prepaid cards.

C. For debt collection:

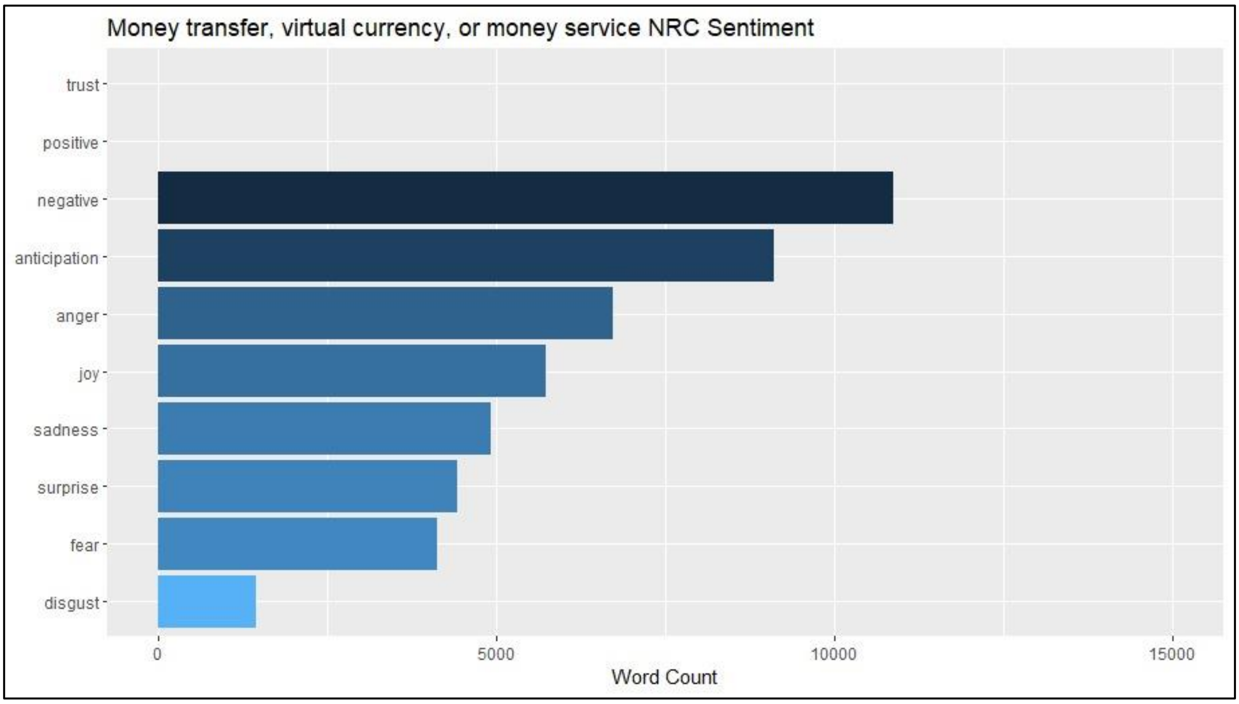


Here, disgust is the emotion with most complaints followed by surprise.

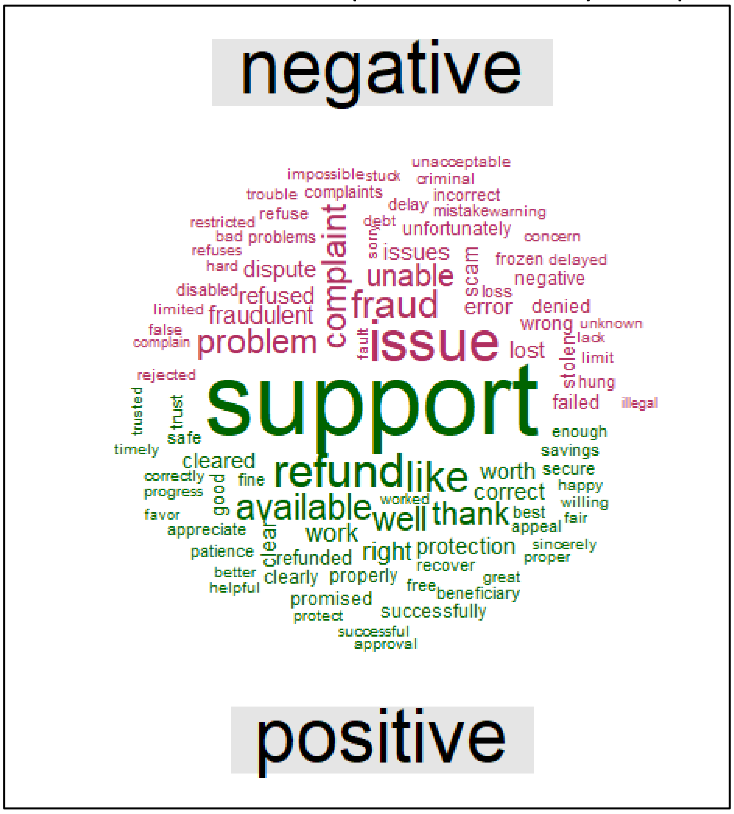


"Debt" is the word with the highest frequency in our given database, with respect to the Debt Collection.

D. For money transfer, vital currency or money services:

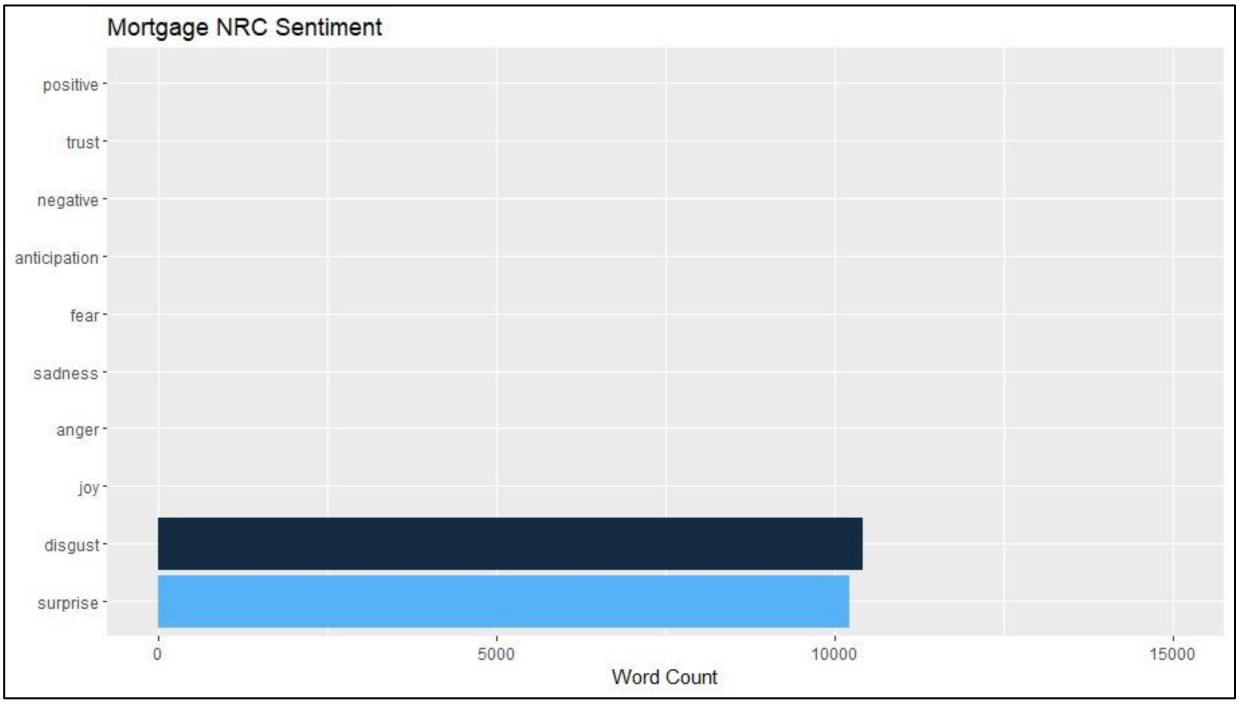


Here, negative is the emotion with most complaints followed by anticipation, anger and so on.

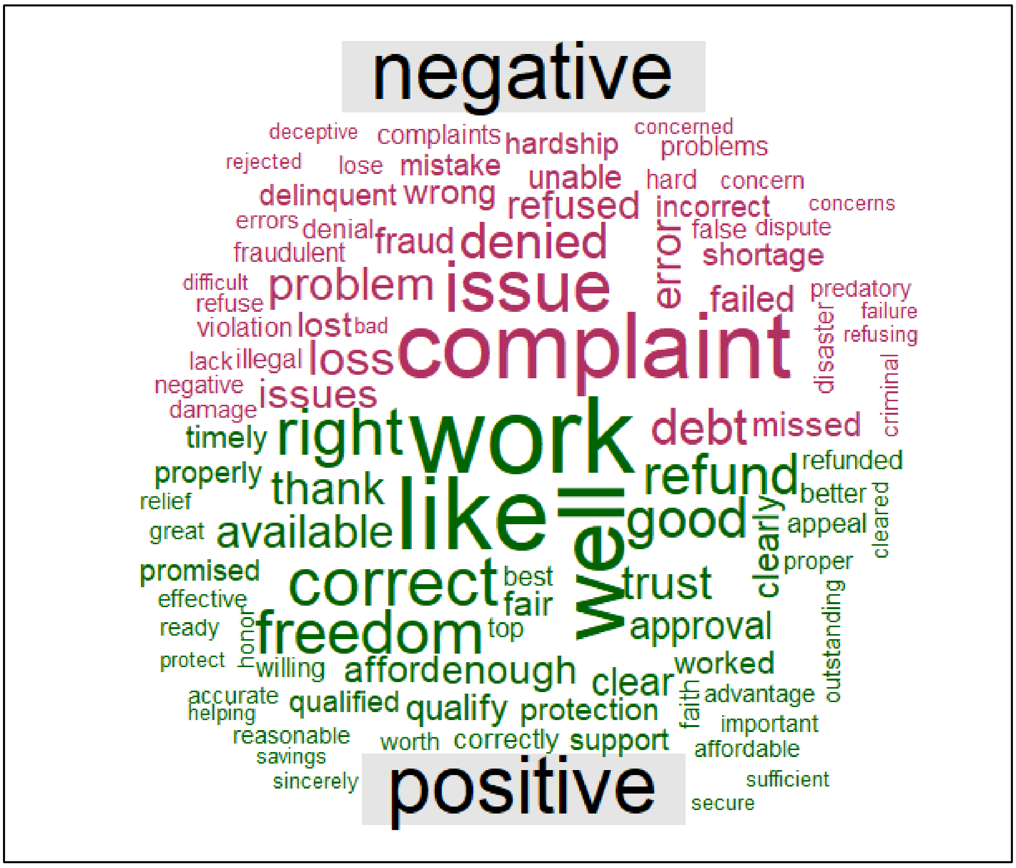


“Support” is the word with the highest frequency in our given database, with respect to the money transfer, vital currency or money services.

E. For mortgage:

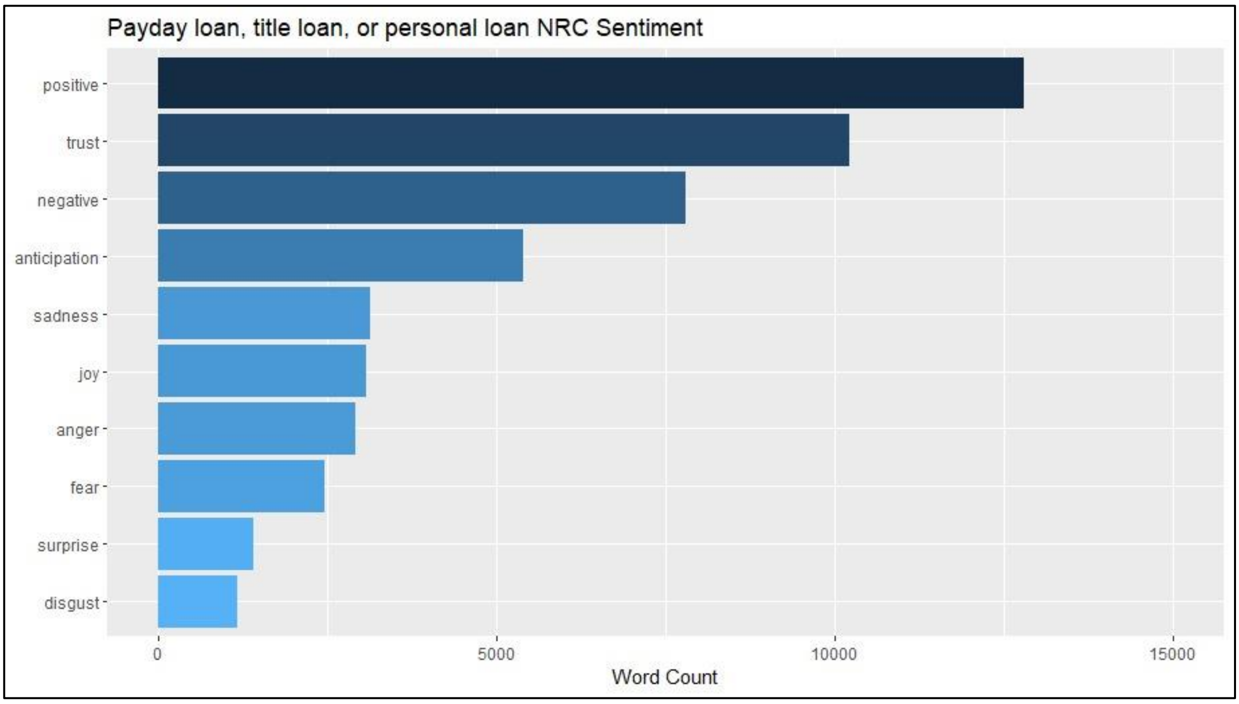


Here, disgust is the emotion with most complaints followed by surprise.

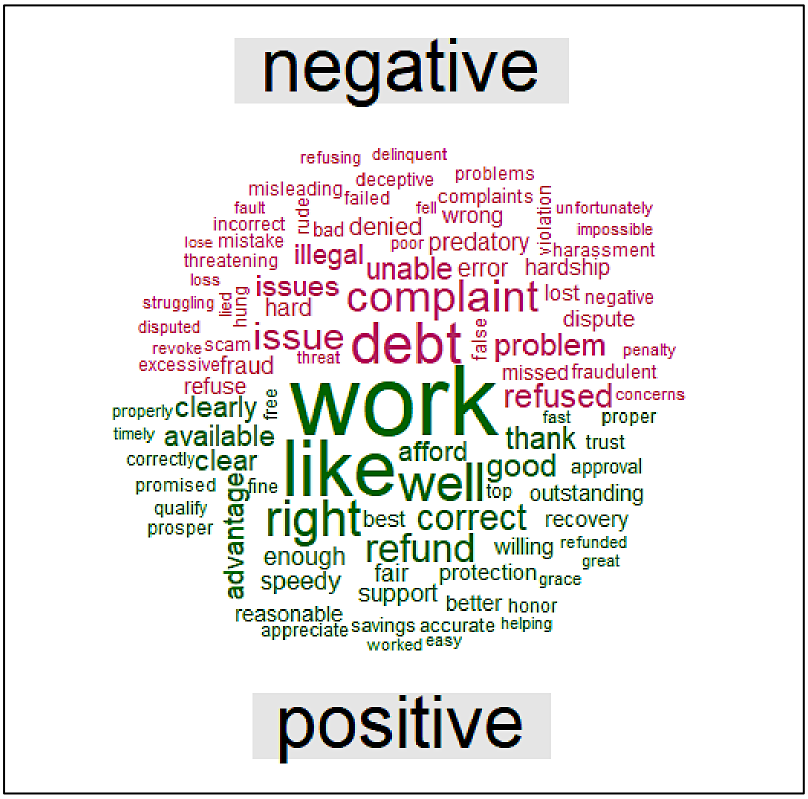


“Work”, “Complaint”, “Like” are some of the words with the highest frequency in our given database, with respect to Mortgage.

F. For payday loan:

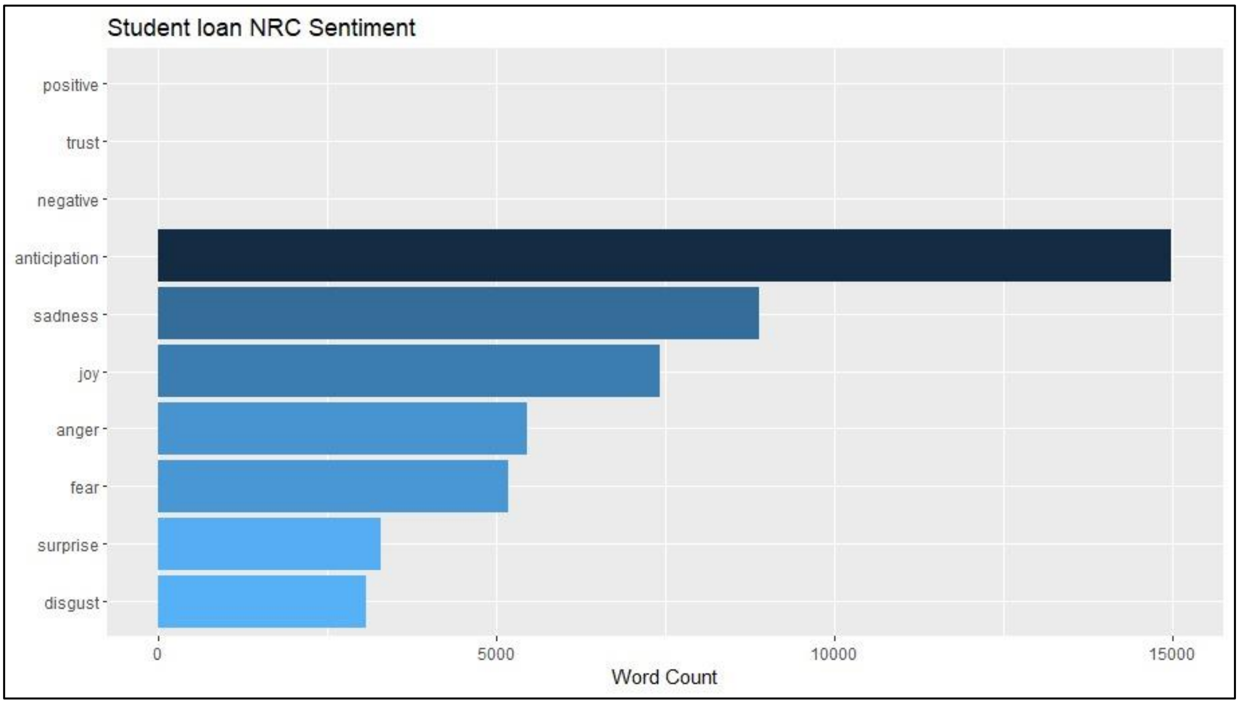


Here, positive is the emotion with most complaints followed by trust, negative, anticipation, sadness, joy etc.

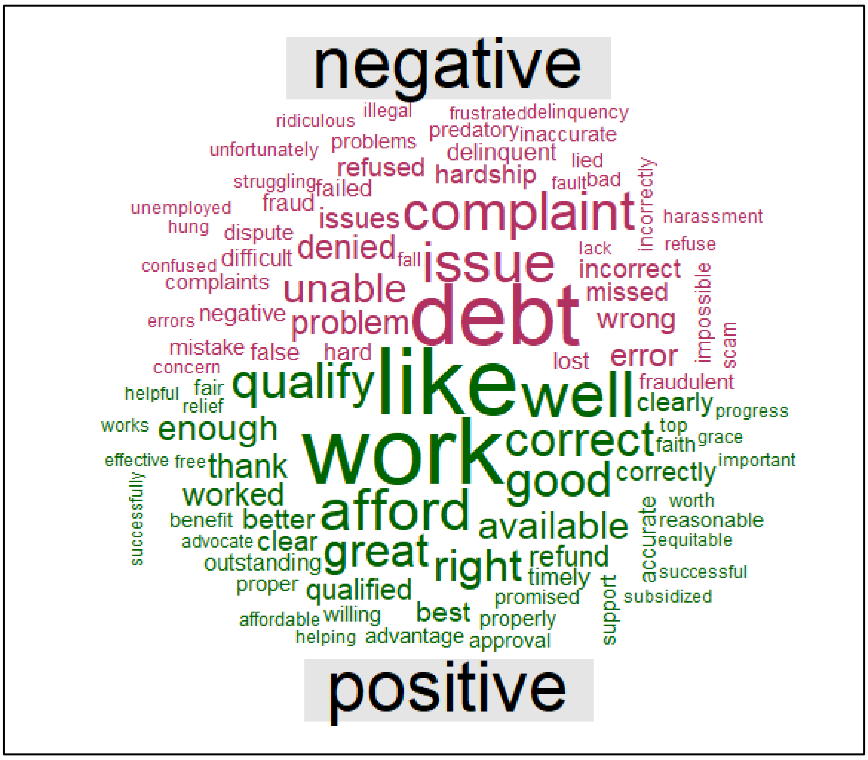


“Work”, “Like” are some of the words with the highest frequency in our given database, with respect to Payday Loan, Title Loan, or Personal Loan.

G. For student loan:

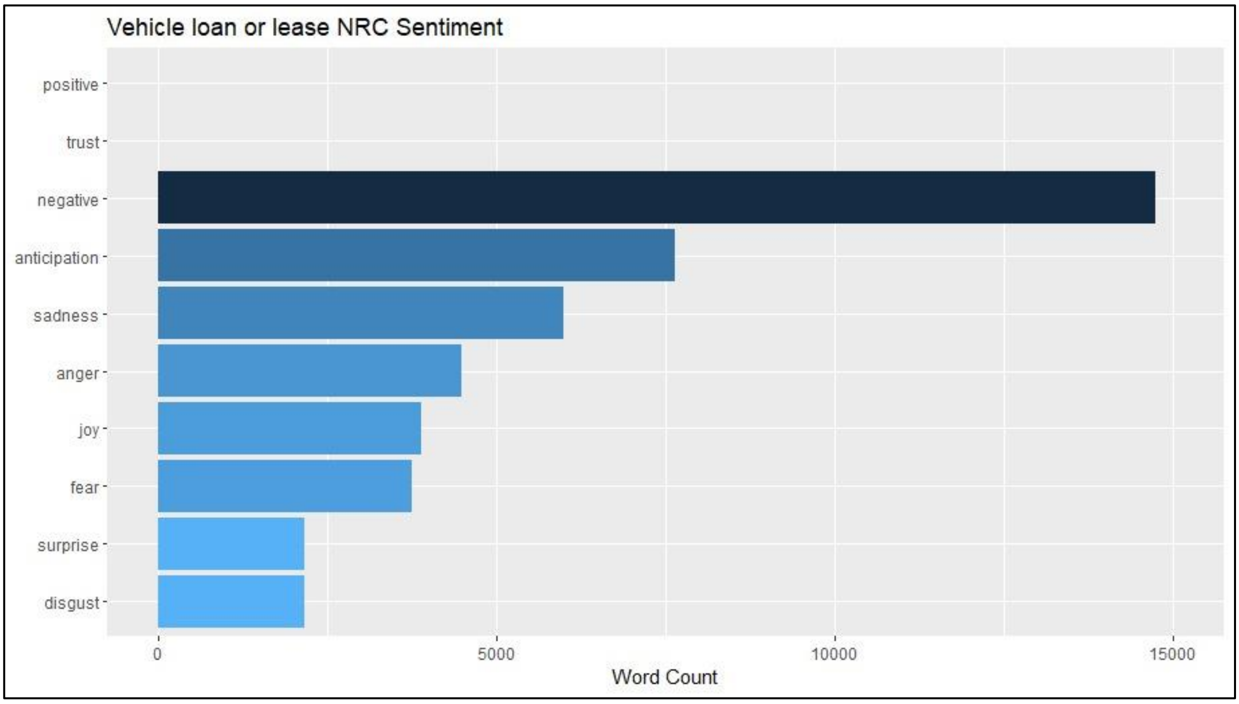


Here, anticipation is the emotion with most complaints followed by joy, anger, fear, surprise and disgust.

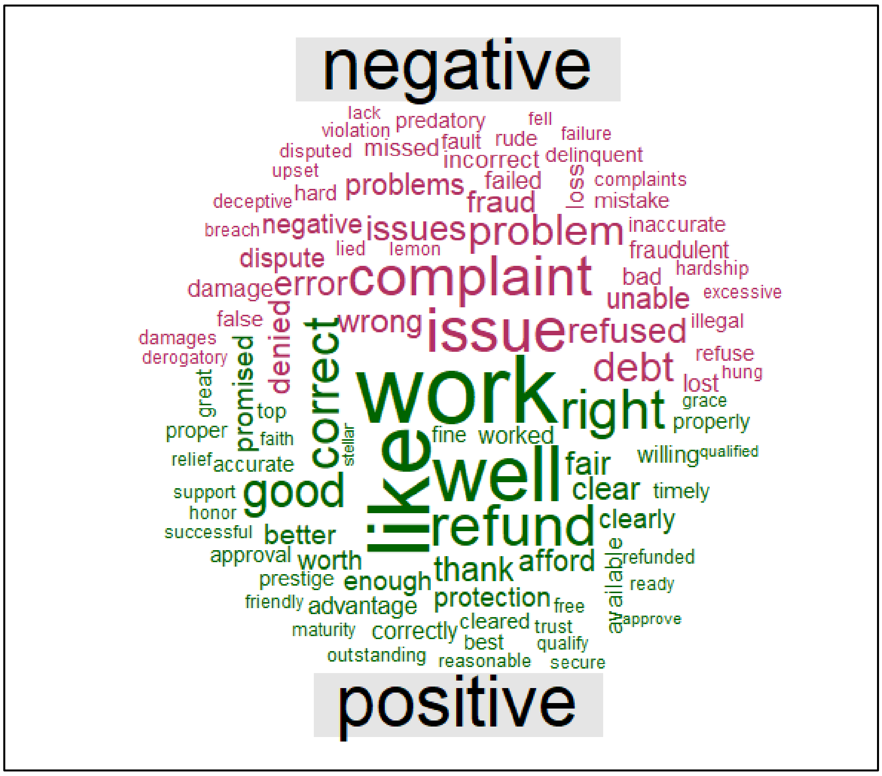


"Work", "Like", "Afford", "Debt" are some of the words with the highest frequency in our given database, with respect to Student Loan.

H. For vehicle loan or lease:



Here, negative is the emotion with most complaints followed by anticipation, sadness, anger and so on.



“Work”, “Like”, “Refund” are some of the words with the highest frequency in our given database, with respect to Vehicle Loan or Lease.

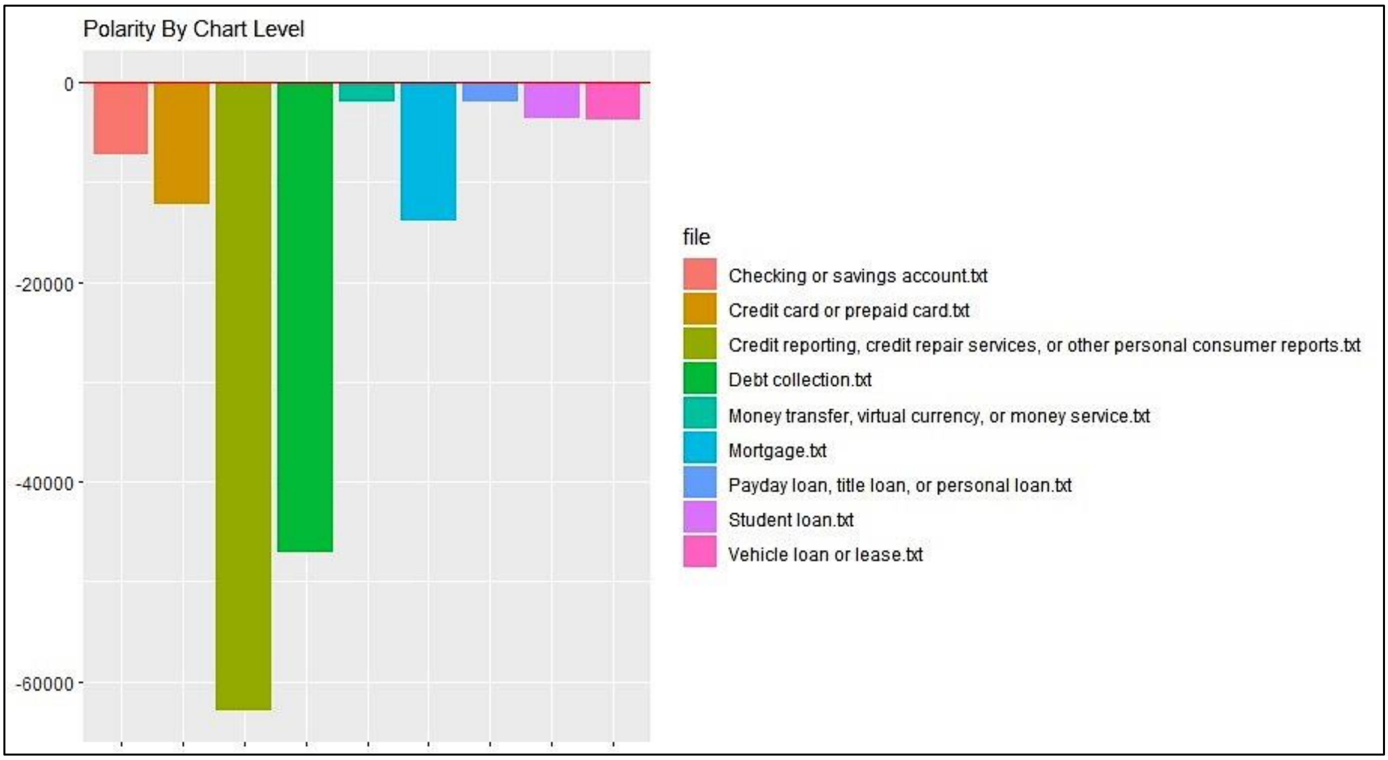
```r
#plot nrc plot
undesirable_words <- c("dear","sir","madam")

#Store list of filenames (complaint files)
files <- list.files("C://Users/sreel/OneDrive/XLRI/Term 4/BIDM/Project/input")

#glue path to file name
fileName <- glue("C://Users/sreel/OneDrive/XLRI/Term 4/BIDM/Project/input/", files[9], sep = "")

#Remove whitespace
fileName <- trimws(fileName)

fileText <- glue(read_file(fileName))

tokens <- data_frame(text = fileText) %>% unnest_tokens(word, text)
tokens %>%
  filter(!nchar(word) < 3) %>% #remove words with less than 3 characters
  filter(!word %in% undesirable_words) %>%
  anti_join(stop_words)


savings_nrc <- tokens %>%
  inner_join(get_sentiments("nrc"))

nrc_plot <- savings_nrc %>%
  group_by(sentiment) %>%
  summarise(word_count = n()) %>%
  ungroup() %>%
  mutate(sentiment = reorder(sentiment, word_count)) %>%
  #Use `fill = -word_count` to make the larger bars darker
  ggplot(aes(sentiment, word_count, fill = -word_count)) +
  geom_col() +
  guides(fill = FALSE) + #Turn off the legend
  labs(x = NULL, y = "Word Count") +
  scale_y_continuous(limits = c(0, 15000)) + #Hard code the axis limit
  ggtitle("Vehicle loan or lease NRC Sentiment") +
  coord_flip()

plot(nrc_plot)
```

1.1.3. Polarity-

Additionally, we found polarity of the sentiments used in all these categories with results as in the table below:

```
> sentiments
# A tibble: 9 x 5
  negative positive sentiment file
     <dbl>    <dbl>     <dbl> <chr>
1    20214    12994     -7220 Checking or savings account.txt
2    32347    20063    -12284 Credit card or prepaid card.txt
3   111110    48239    -62871 Credit reporting, credit repair services, or ot~
4    73950    26913    -47037 Debt collection.txt
5     8994     7006     -1988 Money transfer, virtual currency, or money serv~
6    40257    26318    -13939 Mortgage.txt
7     5221     3270     -1951 Payday loan, title loan, or personal loan.txt
8    12933     9360     -3573 Student loan.txt
9     9011     5294     -3717 Vehicle loan or lease.txt
```

This led to the following polarity comparison between the categories of complaints:



Through the polarity analysis using bing lexicon in tidytext, we can infer that Credit Reporting, credit repair services or other personal consumer reports cause the maximum negative complaints, followed by Debt collection & Mortgage

1.2. Business Insights

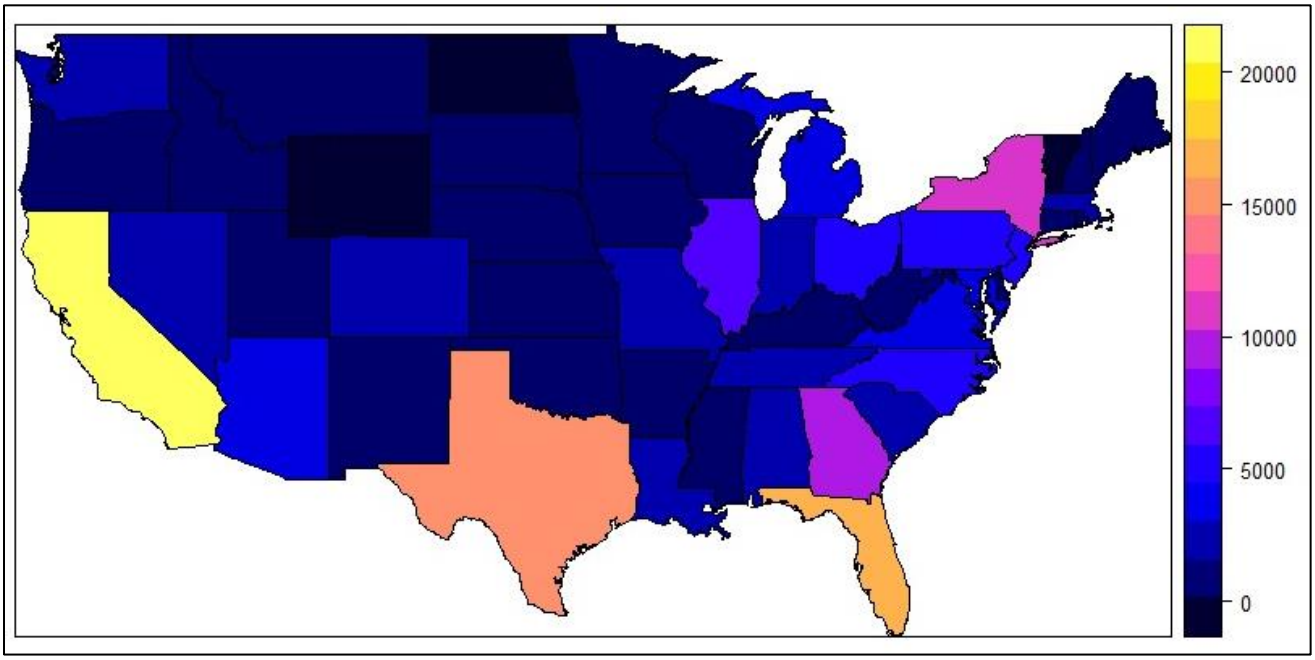
1. **Credit Reporting, credit repair services, or other personal consumer reports and Debt Collection** create the largest number of complaints amongst the users. Companies should focus on these major criteria so that they are able to track and handle issues in an effective manner.
2. From the word cloud and sentiment analysis, "**Fraud**" is the leading opinion that runs amongst the public. So it is imperative to address these concerns in order to better reach the people and solve their problems better.

2. Frequency Analysis

The frequency of the different types of complaints according to various different kinds of parameters might help us understand problem areas. The frequency of different type of products would also help us understand which products are more problematic in general. We will be able to see what kinds of complaints arise in which areas.

2.1 Observations

Total heat map of the complaints received across the 50 states:

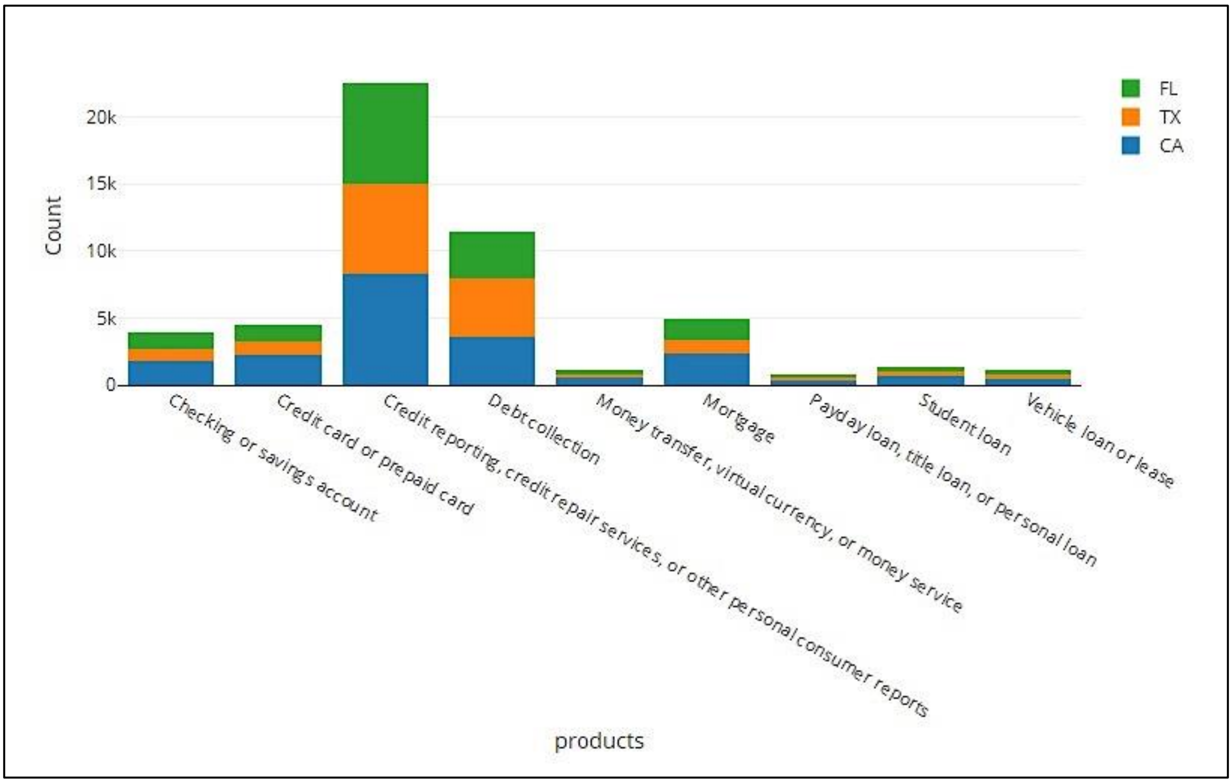


From this, we notice that the top 3 states in terms of complaints are:

- **Florida (FL)**
- **Texas (TX)**
- **California (CA)**

Taking a look at the break-up of the complaints in this area, we see:

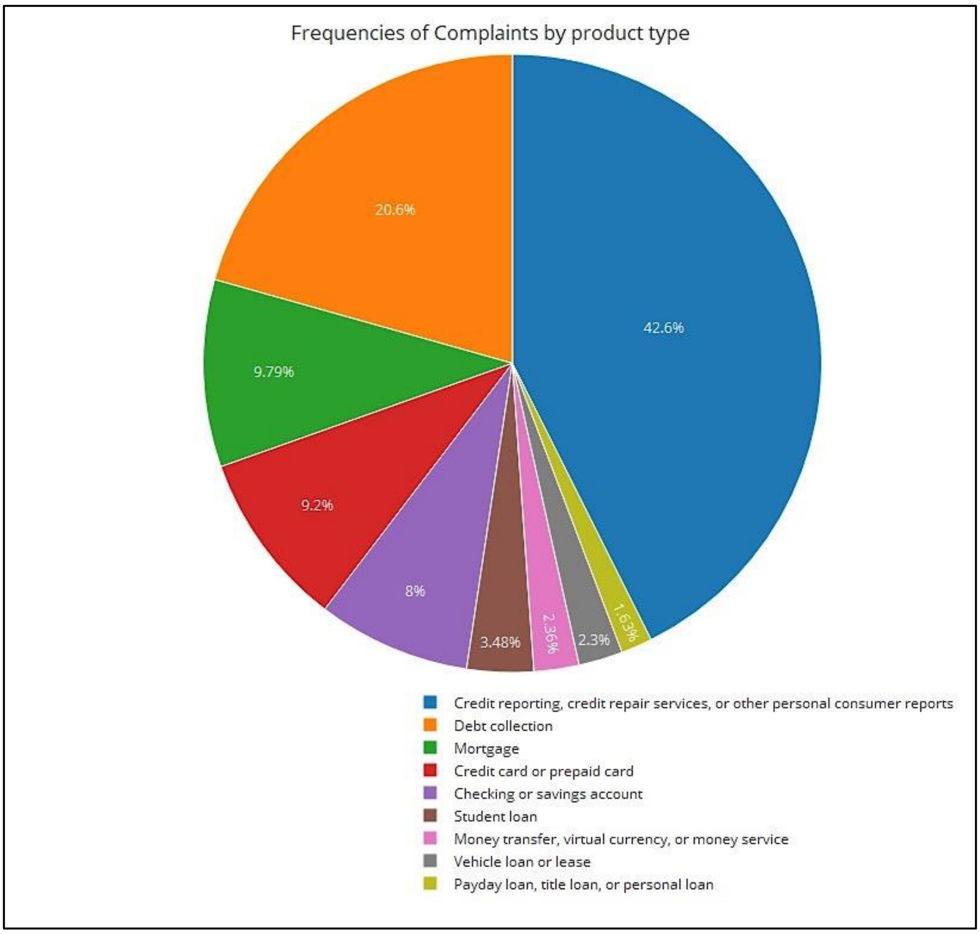
For all types of financial complaints, California has the highest number of complaints. Overall, Credit Reporting, credit repair services, or other personal consumer reports have registered the highest complaints across all the three states. Whereas, Payday loan, title loan, or personal loan has the least amount of complaints across these regions of states.




```
#plot complaint frequences on the US map
library(maps)
library(maptools)
library(sp)
library(plotly)
b<-as.data.frame(table(dataset_main$State))
str(b)
names(b)[1] <-'state.abb'
str(b)
b$states <- tolower(state.name[match(b$state.abb,  state.abb)])
str(b$states)
mapUSA <- map('state',  fill = TRUE,  plot = FALSE)
nms <- sapply(strsplit(mapUSA$names,  ':'),  function(x)x[1])
USApolygons <- map2SpatialPolygons(mapUSA,  IDs = nms,  CRS('+proj=longlat'))

idx <- match(unique(nms),  b$states)
dat2 <- data.frame(value = b$Freq[idx], state = unique(nms))
row.names(dat2) <- unique(nms)

USAsp <- SpatialPolygonsDataFrame(USApolygons,  data = dat2)
spplot(USAsp['value'])
```

```
#plot complaint frequencies by location
ca_data=subset(dataset_main, State == "CA")
ca_prod<-as.data.frame(table(ca_data$Product))
ca_prod_freq = ca_prod$Freq

tx_data=subset(dataset_main, State == "TX")
tx_prod<-as.data.frame(table(tx_data$Product))
tx_prod_freq = tx_prod$Freq

fl_data=subset(dataset_main, State == "FL")
fl_prod<-as.data.frame(table(fl_data$Product))
fl_prod_freq = fl_prod$Freq

products = ca_prod$Var1
data = data.frame(products,ca_prod_freq,tx_prod_freq)

p <- plot_ly(data, x = ~products, y = ~ca_prod_freq, type = 'bar', name = 'CA') %>%
  add_trace(y = ~tx_prod_freq, name = 'TX') %>%
  add_trace(y = ~fl_prod_freq, name = 'FL') %>%
  layout(yaxis = list(title = 'Count'), barmode = 'stack')
p
```

Overall the break-up of complaints across the product categories are:

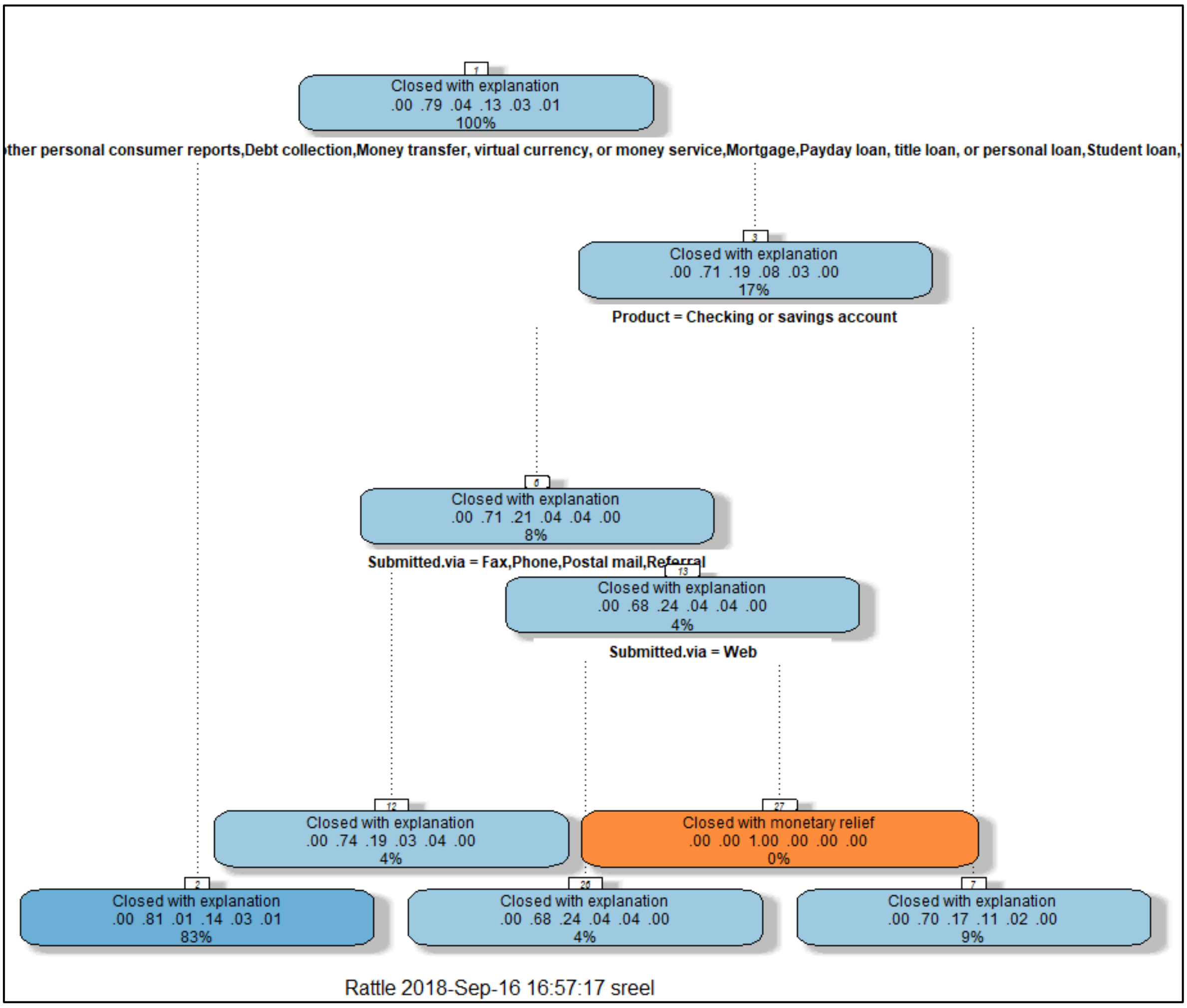


From the above pie chart we can infer that maximum number of complaints arise from the Credit reporting, credit repair services, or other personal consumer reports category, followed by debt collection and mortgage.

2.2 Business Insights

1. Considering the top 3 states with the highest number of complaints are Florida, Texas and California, companies need to be more proactive in these regions, especially if they are looking to open up new branches or areas of service.
2. Amongst these regions, special attention should be given to Credit Reporting, credit repair services, or other personal consumer reports and Debt Collection, which contribute to the highest portion of complaints.

3. Decision Tree using RPart



The decision tree classifies the consumer complaints on the basis of product type, method of submission to arrive at how the consumer responded to customers about their complaints.

The unique product types are as follows:

[1] "Checking or savings account "

[2] "Credit card or prepaid card "

[3] "Credit reporting, credit repair services, or other personal consumer reports "

[4] "Debt collection "

[5] "Money transfer, virtual currency, or money service "

[6] "Mortgage "

[7] "Payday loan, title loan, or personal loan "

[8] "Student loan "

[9] "Vehicle loan or lease "

- For **Products 3-9**, the 83% responses from the company are “Closed with Explanation”.
- For **Product 2**, 9% responses from the company are “Closed with Explanation”.
- For **Product 1**, 8% responses from the company are “Closed with Explanation”. The breakup of this 8% is obtained from the following sub-cases:
    - 4% responses from the company are “Closed with Explanation”- those submitted via Fax, Phone, Postal Mail and Referral
    - 4% responses from the company are “Closed with Explanation”- those submitted via Web
- For **Product 1**, those which are submitted via e-mail have negligible (~0%) cases where monetary relief is given.

```
#decision tree

library(dplyr)
dataset_main3 <- dataset_main
decision_data = select(dataset_main3, Product, Submitted.via, Company.response.to.consumer)
str(decision_data)

library(rpart)
library(rattle)
ac<-rpart(Company.response.to.consumer~.,data=decision_data,method="class",control =rpart.control(minsplit =1,minbucket=1, cp=0))
summary(ac)
plot(ac, uniform=TRUE, main="Classification tree for Company Response to consumers")
text(ac,pretty=0)

fancyRpartPlot(ac)
```

REFERENCES

1. https://www.tidytextmining.com/sentiment.html
2. https://stackoverflow.com

CODES

```
library(tm)
library(RColorBrewer)
library(NLP)
library(wordcloud)
library(SnowballC)
library(plyr)
library(stringr)
library(Rstem)
library(tidyverse)
library(tidytext)
library(glue)
library(reshape2)
library(dplyr)
setwd('C://Users/sreel/OneDrive/XLRI/Term 4/BIDM/Project')

dataset_main = read.csv('2018.csv',header = T)

issue = dataset_main[,4]
complaint = dataset_main[,6]
# issue is taking data only for column 4 which has the issue in complaints.
#We have clustered only issue but any other column can be clustered as well.
ab=Corpus(VectorSource(issue))
#inspect(ab)

ac= tm_map(ab,tolower)
ac = tm_map(ac, removeWords, stopwords("english"))
ac = tm_map(ac, removePunctuation)
ac = tm_map(ac, removeNumbers)
ac = tm_map(ac, stripWhitespace)

ac = tm_map(ac, stemDocument, language = "english")
#statements to prepare the document

wordcloud(ac,min.freq = 4)
dtm = DocumentTermMatrix(ac)
#head(inspect(dtm1))
dtm1 = as.matrix(dtm)
#dtm1
kk = colSums(dtm1)
kk1 = sort(kk, decreasing = TRUE)
#head(kk1)

wtt = weightTfIdf(dtm)
#inspect(wtt)
wtt1 = as.matrix(wtt)
```

```r
#wtt1


f1= function(x)
{
  (sum(x^2))^.5
}
f2 = function(y)
{
  wtt1/apply(y, 1, f1)
}
am = f2(wtt1)

abc = 10

# for (i in 2:10)
# {
#   abc[i]<-kmeans(am, centers = i, nstart = 10, iter.max = 10)$tot.withinss
# }
# plot(1:10,abc,type = "b")
ak = kmeans(am,5)
#ak

dd<-cbind(dataset_main,ak$cluster)
head(dd)






#Sentiment Analysis
#Extract complaint narrative
dataset_main1<-dataset_main[!(dataset_main$Consumer.complaint.narrative==""),]

#Find unique product types
product = unique(dataset_main1[c("Product")])

#Clean data for sentiment analysis and write complaints for each product to separate text file
CleanData = function(complaint, fileName){
  complaint = gsub("[[:punct:]]","", complaint)
  complaint = gsub("@\\w+","", complaint)
  complaint = gsub("[[:digit:]]","", complaint)
  complaint = gsub("  "," ", complaint)
```

```r
  complaint = gsub("[ \t]{2,}"," ", complaint)
  complaint = gsub("XX"," ", complaint)
  complaint = gsub("XXX"," ", complaint)
  complaint = gsub("XXXX","", complaint)
  complaint = gsub("XXXXXXXX","", complaint)
  complaint <- tolower(complaint)
  complaint <- complaint[!is.na(complaint)]
  complaint <- gsub("\\s+", " ", complaint)
  complaint <- trimws(complaint)
  filePath = paste("input/",fileName,".txt", sep="")
  write.table(complaint, file=filePath,sep = "\t", row.names = FALSE)
}

#Loop through products and call CleanData function
for(i in 1:9){
  complaint= dataset_main2[,6]
  CleanData(complaint,product[i,1])
 # CleanData(sentiment_dataset,product[i,1])
}

#Store list of filenames (complaint files)
files <- list.files("C://Users/sreel/OneDrive/XLRI/Term 4/BIDM/Project/input")

# write a function that takes the name of a file and returns the # of postive
# sentiment words, negative sentiment words, and the difference
GetSentiment <- function(file){
  # get the file
  fileName <- glue("input/", file, sep = "")
  # get rid of any sneaky trailing spaces
  fileName <- trimws(fileName)

  df1 <- read.table(fileName, header=TRUE, fill = TRUE)

  # read in the new file
  fileText <- glue(read_file(fileName))
  # remove any dollar signs (they're special characters in R)
  fileText <- gsub("\\$", "", fileText)



  # tokenize
  tokens <- data_frame(text = fileText) %>% unnest_tokens(word, text)

```

```r
  # get the sentiment from the first text:
  sentiment <- tokens %>%
    inner_join(get_sentiments("bing")) %>% # pull out only sentimen words
    count(sentiment) %>% # count the # of positive & negative words
    spread(sentiment, n, fill = 0) %>% # made data wide rather than narrow
    mutate(sentiment = positive - negative) %>% # # of positive words - # of negative owrds
    mutate(file = file) %>% # add the name of our file
    mutate(product = file) # add product



  # return our sentiment dataframe
  return(sentiment)
}

# file to put our output in
sentiments <- data_frame()

# get the sentiments for each file in our datset
for(i in files){
  sentiments <- rbind(sentiments, GetSentiment(i))
}

#plot overall sentiment for each product type
ggplot(sentiments, aes(x = file, y = sentiment, color = file)) +
  geom_hline(yintercept = 0, color = "red") +
  theme(axis.text.x = element_blank())+
  geom_boxplot() # draw a boxplot for each product


#Polarity by product
plot1 <- sentiments %>%
  ggplot( aes(x = file, sentiment, fill = file)) +
  geom_col() +
  geom_hline(yintercept = 0, color = "red") +
   theme(plot.title = element_text(size = 11)) +
  xlab(NULL) + ylab(NULL) +
  theme(axis.text.x = element_blank())+
  ggtitle("Polarity By Chart Level")
plot1


#plot nrc plot
undesirable_words <- c("dear","sir","madam")

```

```r
#Store list of filenames (complaint files)
files <- list.files("C://Users/sreel/OneDrive/XLRI/Term 4/BIDM/Project/input")

#glue path to file name
fileName <- glue("C://Users/sreel/OneDrive/XLRI/Term 4/BIDM/Project/input/", files[9], sep = "")

#Remove whitespace
fileName <- trimws(fileName)

fileText <- glue(read_file(fileName))

tokens <- data_frame(text = fileText) %>% unnest_tokens(word, text)
tokens %>%
  filter(!nchar(word) < 3) %>% #remove words with less than 3 characters
  filter(!word %in% undesirable_words) %>%
  anti_join(stop_words)


savings_nrc <- tokens %>%
  inner_join(get_sentiments("nrc"))

nrc_plot <- savings_nrc %>%
  group_by(sentiment) %>%
  summarise(word_count = n()) %>%
  ungroup() %>%
  mutate(sentiment = reorder(sentiment, word_count)) %>%
  #Use `fill = -word_count` to make the larger bars darker
  ggplot(aes(sentiment, word_count, fill = -word_count)) +
  geom_col() +
  guides(fill = FALSE) + #Turn off the legend
  labs(x = NULL, y = "Word Count") +
  scale_y_continuous(limits = c(0, 15000)) + #Hard code the axis limit
  ggtitle("Vehicle loan or lease NRC Sentiment") +
  coord_flip()

plot(nrc_plot)


#wordcloud
wordCloud_plot = function(file){
  fileName <- glue("input/", file, sep = "")
  # get rid of any sneaky trailing spaces
  fileName <- trimws(fileName)

```

```r
  df1 <- read.table(fileName, header=TRUE, fill = TRUE)

  # read in the new file
  fileText <- glue(read_file(fileName))
  # remove any dollar signs (they're special characters in R)
  fileText <- gsub("\\$", "", fileText)



  # tokenize
  tokens <- data_frame(text = fileText) %>% unnest_tokens(word, text)

  tokens %>%
    inner_join(get_sentiments("bing")) %>%
    count(word, sentiment, sort = TRUE) %>%
    acast(word ~ sentiment, value.var = "n", fill = 0) %>%
    comparison.cloud(colors = c("maroon", "darkgreen"),
                     max.words = 100)

}

wordCloud_plot(files[9])

#decision tree

library(dplyr)
dataset_main3 <- dataset_main
decision_data = select(dataset_main3, Product, Submitted.via, Company.response.to.consumer)
str(decision_data)

library(rpart)
library(rattle)
```

```r
ac<-rpart(Company.response.to.consumer~.,data=decision_data,method="class",control =rpart.control(minsplit =1,minbucket=1, cp=0))
summary(ac)
plot(ac, uniform=TRUE, main="Classification tree for Company Response to consumers")
text(ac,pretty=0)

fancyRpartPlot(ac)


a<-as.data.frame(table(dataset_main$Product))
str(a)
pie(a$Freq,lbls = a$Var1)
a
library(ggplot2)
library(scales)
# Barplot
bp<- ggplot(a, aes(x="", y=a$Freq, fill=Var1))+
  geom_bar(width = 1, stat = "identity")
bp
pie<- bp+coord_polar("y", start=0)+theme(axis.text.x = element_blank())+ xlab(NULL) + ylab(NULL)
pie

str(dataset_main)

#pie chart
p <- plot_ly(a, labels = ~Var1, values = ~Freq, type = 'pie',insidetextfont = list(color = '#FFFFFF'),marker = list(colors = colors,
                           line = list(color = '#FFFFFF', width = 1))) %>%
  layout(title = 'Frequencies of Complaints by product type') %>%
  layout(legend = list(x = 100, y = -10))
p


#plot complaint frequences on the US map
library(maps)
library(maptools)
library(sp)
library(plotly)
b<-as.data.frame(table(dataset_main$State))
str(b)
names(b)[1] <-'state.abb'
str(b)
b$states <- tolower(state.name[match(b$state.abb,  state.abb)])
str(b$states)
mapUSA <- map('state',  fill = TRUE,  plot = FALSE)
nms <- sapply(strsplit(mapUSA$names,  ':'),  function(x)x[1])
```

```r
USApolygons <- map2SpatialPolygons(mapUSA,  IDs = nms,  CRS('+proj=longlat'))

idx <- match(unique(nms),  b$states)
dat2 <- data.frame(value = b$Freq[idx], state = unique(nms))
row.names(dat2) <- unique(nms)

USAsp <- SpatialPolygonsDataFrame(USApolygons,  data = dat2)
spplot(USAsp['value'])



#plot complaint frequencies by location
ca_data=subset(dataset_main, State == "CA")
ca_prod<-as.data.frame(table(ca_data$Product))
ca_prod_freq = ca_prod$Freq

tx_data=subset(dataset_main, State == "TX")
tx_prod<-as.data.frame(table(tx_data$Product))
tx_prod_freq = tx_prod$Freq

fl_data=subset(dataset_main, State == "FL")
fl_prod<-as.data.frame(table(fl_data$Product))
fl_prod_freq = fl_prod$Freq

products = ca_prod$Var1
data = data.frame(products,ca_prod_freq,tx_prod_freq)

p <- plot_ly(data, x = ~products, y = ~ca_prod_freq, type = 'bar', name = 'CA') %>%
  add_trace(y = ~tx_prod_freq, name = 'TX') %>%
  add_trace(y = ~fl_prod_freq, name = 'FL') %>%
  layout(yaxis = list(title = 'Count'), barmode = 'stack')
p
```