Project2 Lokeswari Umakanthan(LXU190000)

Real Time Sentiment Analysis of Tweets

Libraries required for the project

```
#Libraries to be imported
library(rtweet)

## Warning: package 'rtweet' was built under R version 3.6.3

library(sentimentr)

## Warning: package 'sentimentr' was built under R version 3.6.3

library(SentimentAnalysis)

## Warning: package 'SentimentAnalysis' was built under R version 3.6.3

##
## Attaching package: 'SentimentAnalysis'

## The following object is masked from 'package:base':
##
##     write

library(dplyr)

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

library(tidyverse)

## Warning: package 'tidyverse' was built under R version 3.6.2

## -- Attaching packages ---------------------------------------------------
------------- tidyverse 1.3.0 --

## v ggplot2 3.2.1     v purrr   0.3.3
## v tibble  2.1.3     v stringr 1.4.0
## v tidyr   1.0.0     v forcats 0.4.0
## v readr   1.3.1
## Warning: package 'purrr' was built under R version 3.6.2

## -- Conflicts ---------------------------------------------------------------
------- tidyverse_conflicts() --
## x dplyr::filter()  masks stats::filter()
## x purrr::flatten() masks rtweet::flatten()
## x dplyr::lag()     masks stats::lag()

library(ggplot2)
library(knitr)
library(tm)

## Warning: package 'tm' was built under R version 3.6.3

## Loading required package: NLP

##
## Attaching package: 'NLP'

## The following object is masked from 'package:ggplot2':
##
##     annotate

library(wordcloud)

## Warning: package 'wordcloud' was built under R version 3.6.3

## Loading required package: RColorBrewer
```

Loading 2000 Tweets related to the query here the query is “`72% of Americans`” posted in English. Tweets can also be loaded based on the tweets posted on a particular time in english language.

Removing all the Hperlinks,Punctuation, Control Symbols and digits.

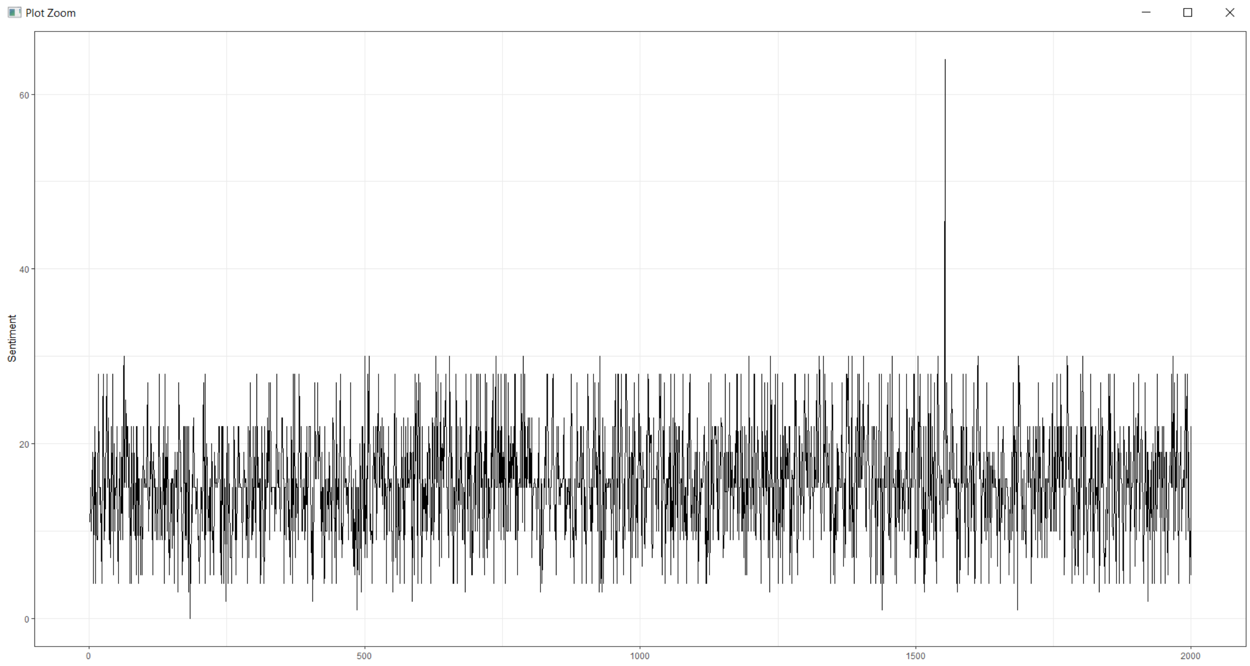
```
#Search for the tweets and load 2000 tweets on coronavirus
Twitter <- search_tweets("72% of Americans", n =2000, include_rts = TRUE,lang
="en")
# Process the tweets by removing the hyperlinks,punctuation,whitespaces,stopw
ords and convert to lowercase
Twitter$stripped_text <- gsub("http.*","",Twitter$text)
Twitter$stripped_text <- gsub("https.*","",Twitter$stripped_text)
Twitter$stripped_text <- gsub('[[:punct:]]',"" ,Twitter$stripped_text)
Twitter$stripped_text <- gsub('[[:cntrl:]]', "" ,Twitter$stripped_text)
Twitter$stripped_text <- gsub('\\d+',"" ,Twitter$stripped_text)
tweet_list <- Twitter$stripped_text
```

Creating a function to remove the english Stopwords, specila symbols/characters, whitespace, punctuation from every tweet and convert every tweet to lower. The Cleaned tweet has the list of 2000 pre-processed tweets.

```
clean <- function(x){
  x <- gsub("http.*","",x)
  x <- str_replace_all(x, "[^[:alnum:]]", " ")
  x <- tolower(x)
  x <-removeWords(x,stopwords('en'))
  x <-removeWords(x,c('americans','Americans'))
  x <-removePunctuation(x)
  x <-stripWhitespace(x)
  return(x) }

cleaned_tweets <- clean(tweet_list)
length(cleaned_tweets)

## [1] 2000
```

The analyzeSentiment() function of the SentimentAnalysis package is used for estimating the sentiment of each tweet from the range -1 to +1. A graph has been plotted based on these values of the sentiments for each 2000 tweets.

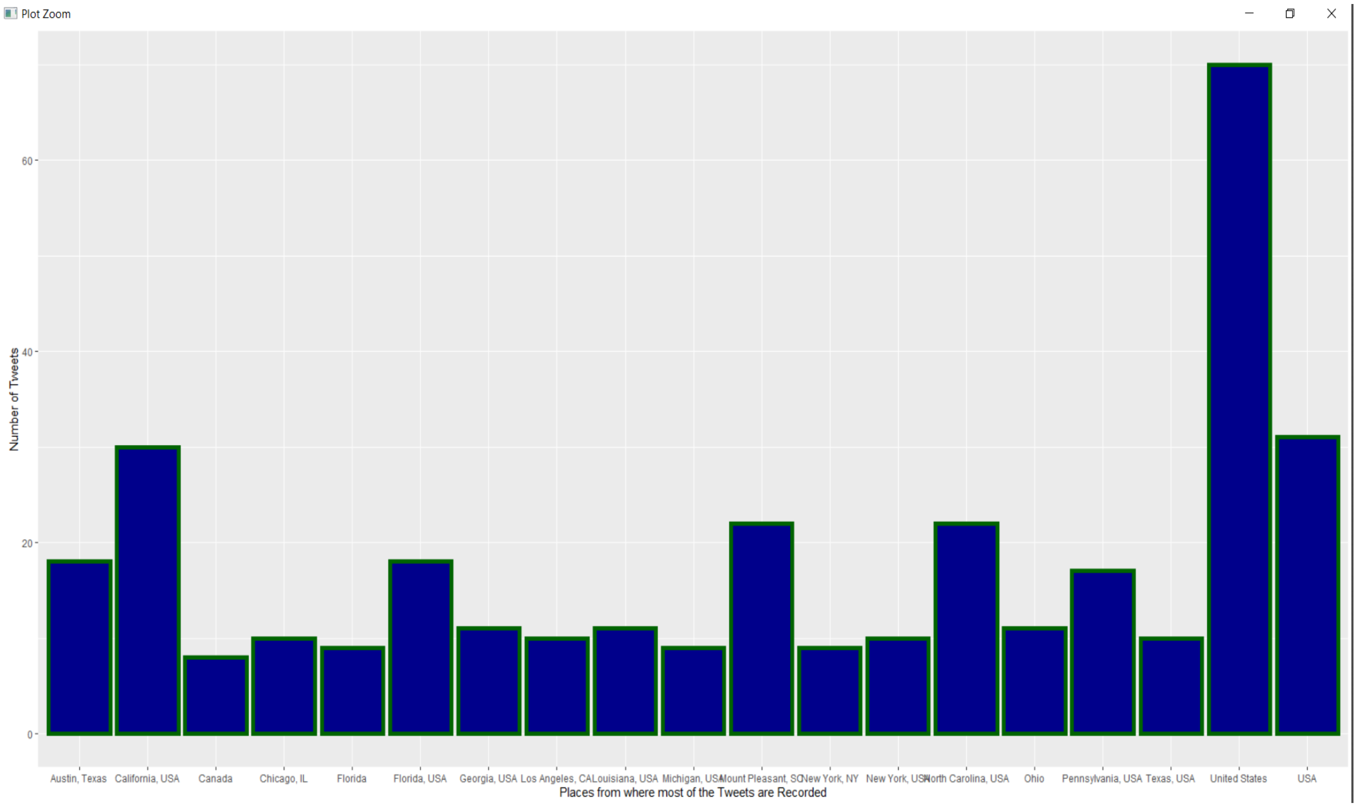
```
sentiment <- analyzeSentiment(cleaned_tweets)
plotSentiment(sentiment, x = NULL, cumsum = FALSE, xlab = "",
              ylab = "Sentiment")
```



Observation : The x-axis is the number of tweets and the y-axis is the cumulative sum of the sentiment of each tweet ranging from -1 to +1. Since there is a mix of positive and negative tweets the cumulative sum is maintained within a certain range. The point where there is a higher value means there are higher number of positive tweets at that range.

The Location of each tweets are recorded, and the top 20 locations are plotted against the count of the number of tweets recorded from that location. This gives the visualization that which of the countries has most recorded tweets.

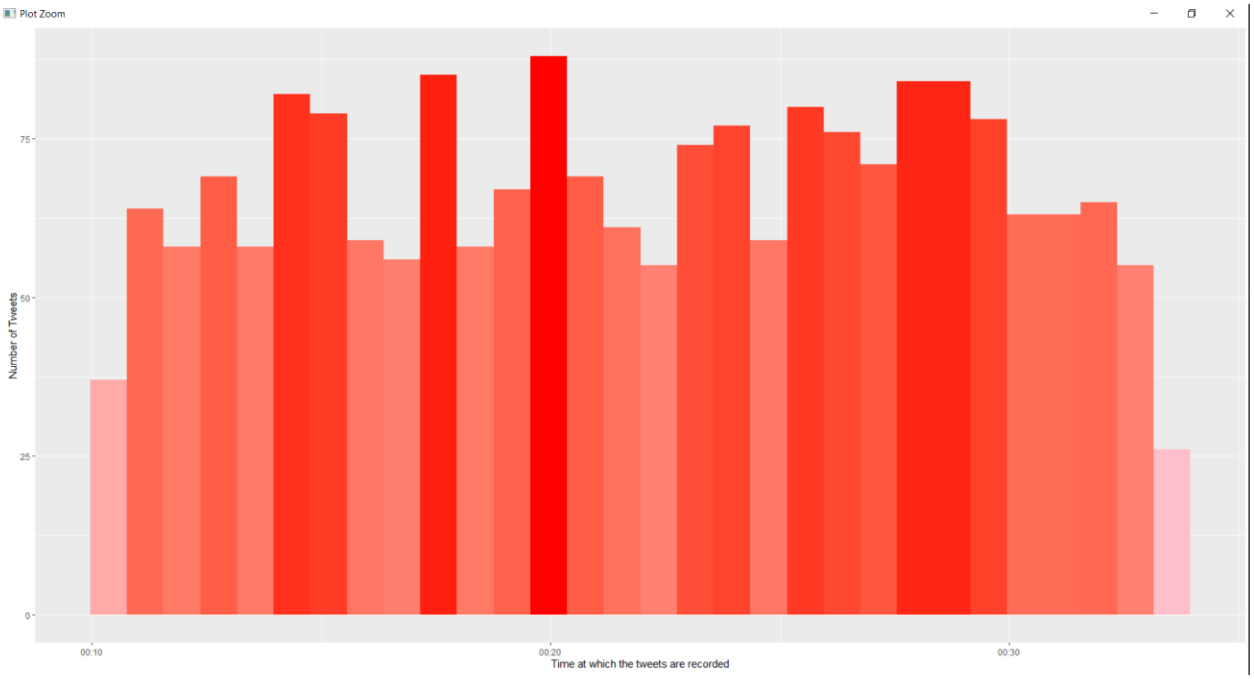
```
location_names <- unique(unlist(Twitter$location),use.names =FALSE)
location_count <- tabulate(match(Twitter$location, unique(Twitter$location)))
location_df <- data.frame("names" = unlist(location_names),"count" = unlist(l
ocation_count))
location_df <- location_df %>%
  arrange(desc(location_count)) %>% top_n(20)

## Selecting by count

location_df <- location_df[2:20,]
maxcount <- max(location_df$count)

ggplot(data = location_df,aes(x =names, y=count))+
  geom_col(col='dark green', fill ='dark blue', size = 2)+  theme(legend.posi
tion = "none")+
  ylim(0,maxcount)+
  xlab("Places from where most of the Tweets are Recorded")+ ylab("Number of
Tweets")
```



Observation : The location of the tweets are recorded for every 2000 tweets. This plot shows the count of tweets recorded in that location. The location plotted in x-axis are the top 19 location with the highest number of tweets. The first two highest number of tweets are recorded under the location named “United States” and “USA”. The third highest number of tweets is recorded under “California”. This shows that most of the people from California are involved/interested in this event/issue or happening of “72% of Americans”

The second plot is regarding the visualization of the time when the highest number of tweets recorded. The number of tweets is scaled with the time at which it is posted. This gives us the time at which the particular topic has been talked and tweeted about.

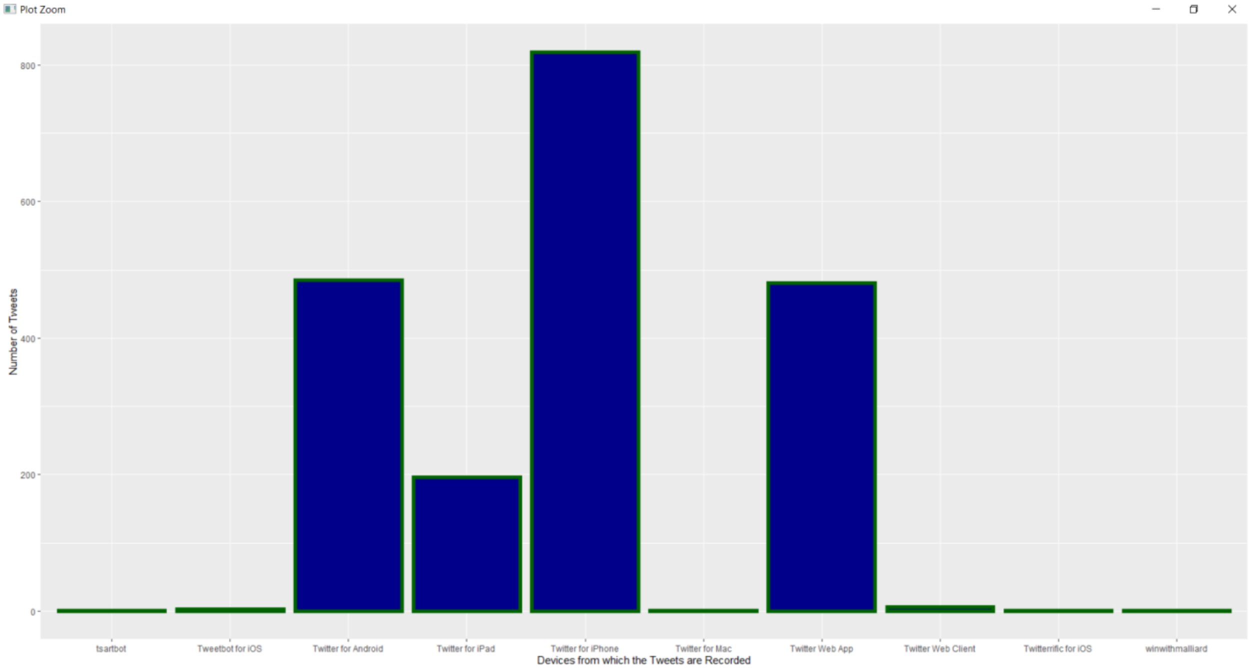
```
#Plot for analysing the time when the highest number of tweets are recorded
ggplot(data = Twitter,aes(x = Twitter$created_at))+
  geom_histogram(aes(fill = ..count..))+
  theme(legend.position = "none")+
  xlab("Time at which the tweets are recorded")+ ylab("Number of Tweets")+
  scale_fill_gradient(low = 'pink',high ='red')

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Observation : The time and date for every tweet has been recorded. This gives us the data that which is the time the incident/happening broke or initiated or getting crucial. We can conclude from the graph that the highest number of tweets has been recorded at 4/9/2020 00:20AM. So, with this analysis we can make an assumption that the time at which the highest number of tweets are recorded is the time at which the incident or the event has become critical and important. Since the presidential election has been going on actively, the citizens are interested in deciding the way to vote during this Corona Virus Crisis. The 72% of the Americans wanted to vote through mail because of this major outbreak of Covid19.

The Source is the device used by the person to tweet that tweet. The tweet has been tweeted at that time and using a device. The plot has been visualized based on the number of tweets and devices used to tweet that tweet. The source and the count of tweets have been retrieved using unique and tabulate function. From the plot we can visualize most of the people use Iphone or Android phone to posts their tweets.

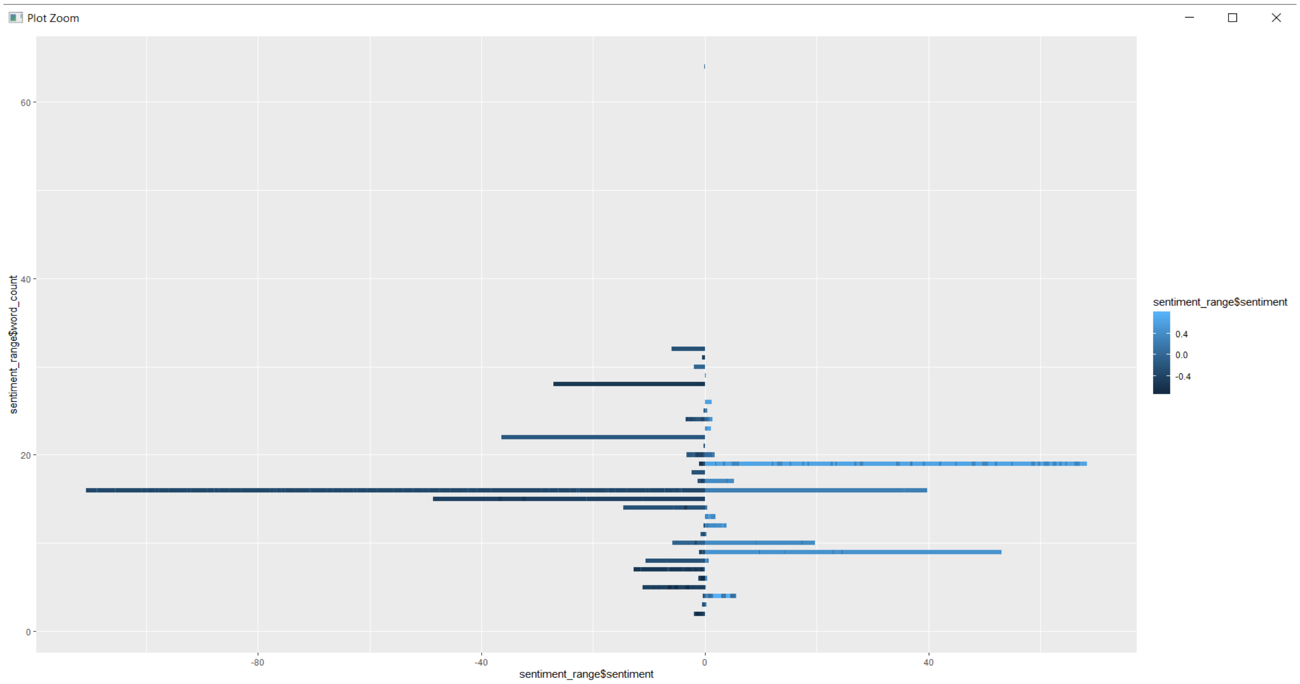
```
source_uniq <- unique(unlist(Twitter$source),use.names =FALSE)
source_count <- tabulate(match(Twitter$source, unique(Twitter$source)))
source_uniq <- source_uniq[1:10]
source_count <- source_count[1:10]
source_df <- data.frame("names" = unlist(source_uniq),"count" = unlist(source
_count))
max_count <- max(source_count)


#Source from which the Tweet has been posted
#Gives the information which digital device people are using most
ggplot(data = source_df,aes(x =source_uniq, y=source_count))+
  geom_col(col='dark green', fill ='dark blue', size = 2)+  theme(legend.posi
tion = "none")+
  ylim(0,max_count)+
  xlab("Devices from which the Tweets are Recorded")+ ylab("Number of Tweets"
)
```



Observation : The Source i.e. the device from which the tweets has been posted are recorded. This gives us the information that which devices most of the people use to connect with the social media. From the graph we can summarize that the most of the people use iPhone or Android phones to upload their tweets. This can also be framed as the people who use iPhone and Android actively tweet about the current issues or incidents.

The sentiment for each tweet has been retrived using the function sentiment() of the sentimentr package.The sentiment obtained for each tweet is stored as an array in sentiment_range. The range of sentiments from(-1 to +1) has been plotted as a bar graph which visualizes the value/intensity of sentiment for each 2000 tweets.The positive and negative tweets have been plotted with respect to the number of words in each tweet.

```
sentiment_range <- sentiment(cleaned_tweets)

count_tweets <- table(sign(sentiment_range$sentiment))

count_tweets

## -1  0  1

## 1289 98 613

## Warning: Each time `sentiment` is run it has to do sentence boundary disam
biguation when a
## raw `character` vector is passed to `text.var`. This may be costly of time
and
## memory.  It is highly recommended that the user first runs the raw `charac
ter`
## vector through the `get_sentences` function.

#Range of sentiment from -1 to +1 for each tweets with respect to their word_
count using sentimentr package
ggplot(sentiment_range, aes(x=sentiment_range$word_count, y=sentiment_range$s
entiment)) +
  geom_bar(stat='identity', aes(fill=sentiment_range$sentiment), width=.5)+
  coord_flip()
```


Plot Zoom
sentiment_range$word_count
0
20
40
60
-80
-40
0
40
sentiment_range$sentiment
sentiment_range$sentiment
0.4
0.0
-0.4

Observation : The sentiment given for each tweet ranges between -1 to +1. The sentiment for each tweet has been plotted against the number of words of each tweet data. This gives us the graph that the number of positive-negative tweets and the conclusion that the negative tweets have a greater number of words that the positive tweets. From the graph we can conclude that people has expressed the negative tweets in a more elaborate way than positive tweets.

The emotion for each tweet has been obtained using the emotion() function using sentimentr package. The emotion gives one of the following categories : anger, anger_negated, anticipation, anticipation_negated, disgust, disgust_negated, fear, fear_negated, joy, joy_negated, sadness, sadness_negated, surprise, surprise_negated, trust, trust_negated. Each tweets have been categorized into one of the above mentioned categories and plotted based on the occurrences in 2000 tweets.

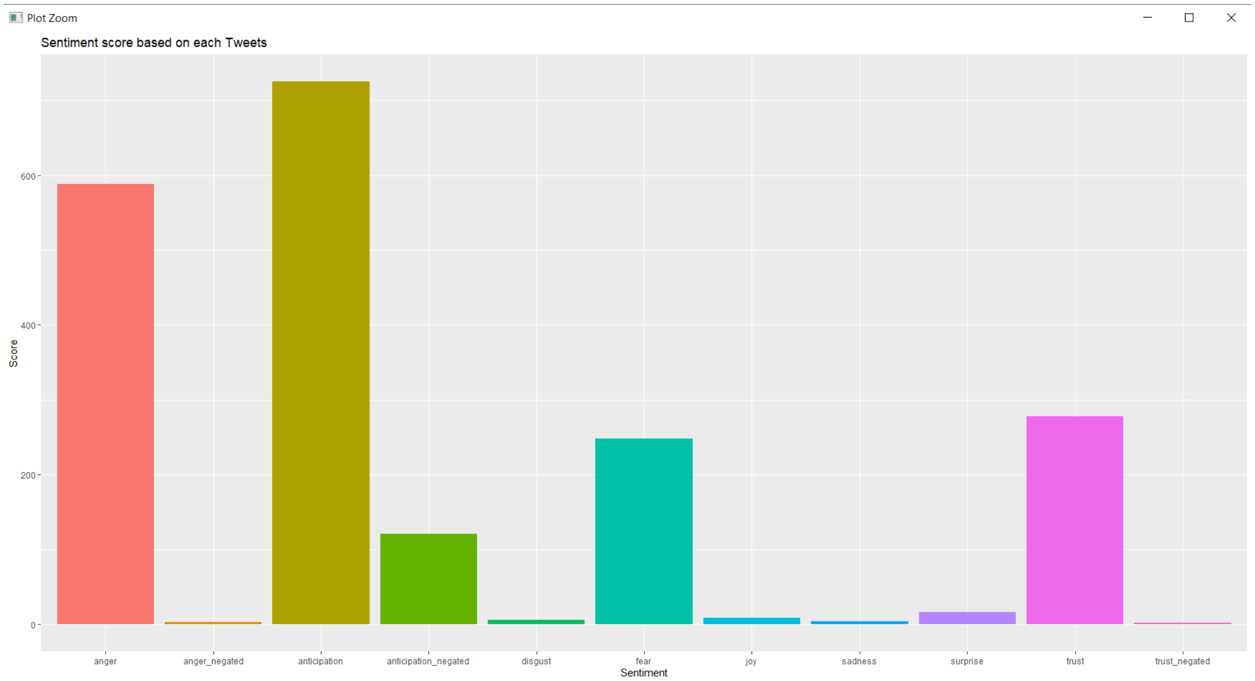
```
#Getting the emotions for each sentence i.e. each tweet using sentimentr pack
age
emotion_range <- emotion(cleaned_tweets)
```

```
## Warning: Each time `emotion` is run it has to do sentence boundary disambi
guation when a
## raw `character` vector is passed to `text.var`. This may be costly of time
and
## memory.  It is highly recommended that the user first runs the raw `charac
ter`
## vector through the `get_sentences` function.
```

```
emotion_range <- emotion_range %>% group_by(element_id) %>% filter(emotion_co
unt == max(emotion_count)) %>% slice(1)
Twitter$sentiment <- emotion_range$emotion_type

emotion_unique <- unique(unlist(emotion_range$emotion_type),use.names =FALSE)
emotion_count <- tabulate(match(emotion_range$emotion_type, unique(emotion_ra
nge$emotion_type)))
emotion_unique <- emotion_unique[1:10]
emotion_count <- emotion_count[1:10]
emotion_df <- data.frame("names" = unlist(emotion_unique),"count" = unlist(em
otion_count))
```

Plotting Histogram which includes the visualization of emotions obtained for each tweet using sentimentr package and their number of occurrences in total 2000 tweets. Their count shows the common emotion in the highest number of tweets.

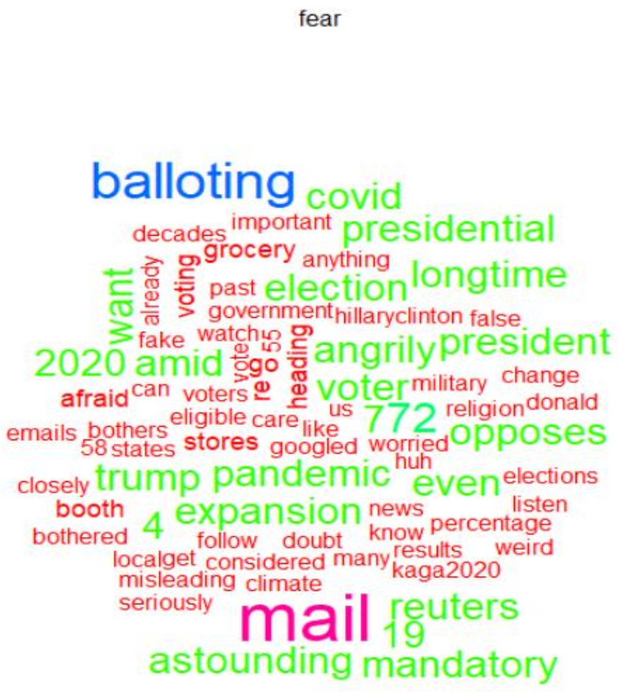
```
#Histogram based on the emotion analysed by sentimentr package
ggplot(data = emotion_df, aes(names, count)) +
  geom_bar(aes(fill = names),stat = "identity")+
  theme(legend.position = "none") +
  xlab("Sentiment") + ylab("Score") + ggtitle("Sentiment score based on each
Tweets")
```



Observation: The Tweets on the 72% of Americans wants to vote through mail is mostly of anticipation sentiment. Most of the people tweets with anticipation/hope that the event will happen. Thus, the highest number of tweets has been recorded with anticipation sentiment. The next higher number of tweets recorded is of anger sentiment. The third mostly recorded sentiment of tweets is trust. These sentiments have been estimated using the emotion() function of sentiment package.

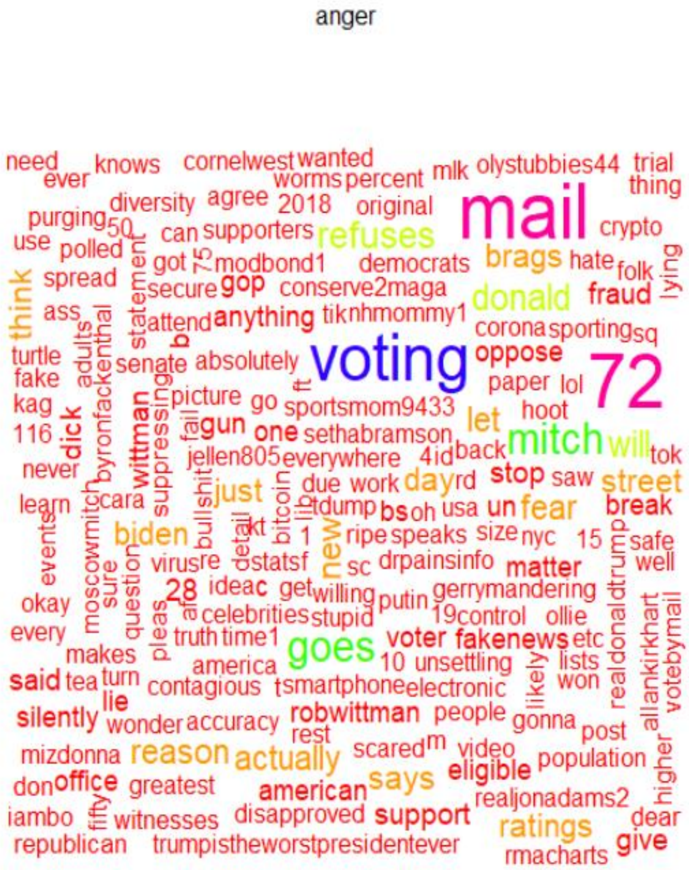
Retrieving the most frequently occuring words in each emotions/sentiment. The tweets are grouped according to the emotions/sentiments and the most frequently occuring words in those tweets are visualized as wordcloud. Each emotion is visualized as a wordcloud.

```r
#Words related to positive and negative tweets or based on emotions

for(un in unique(unlist(Twitter$sentiment)))
{
  Tweet_sentiment <- data.frame()
  sentiment_text <- data.frame()

  Tweet_sentiment <- Twitter[Twitter$sentiment == un,]
  sentiment_text <- Tweet_sentiment$text
  sentiment_text <- clean(sentiment_text)

  line_by <- function(x){
    words_list <-  tibble("word"= unlist(strsplit(x," ")))
    freq_words <- words_list %>%
      group_by(word) %>%
      summarise(count = n()) %>%
      arrange(desc(count)) %>% top_n(100)
    return (freq_words)
  }

  freq <- line_by(sentiment_text)
  layout(matrix(c(1, 2), nrow=2), heights=c(1, 4))
  par(mar=rep(0, 4))
  plot.new()
  text(x=0.5, y=0.5, un)
  wordcloud(freq$word,freq = freq$count,color = rainbow(10),scale = c(3,0.5),
min.freq = 1, random.color = FALSE,main = "Title")

}

## Selecting by count

## Warning in wordcloud(freq$word, freq = freq$count, color = rainbow(10), :
## mail could not be plotted. It will not be plotted.

## Selecting by count
```



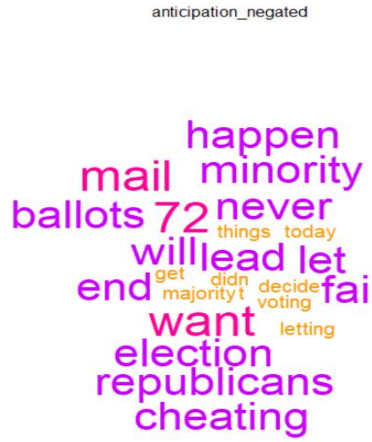
Observation : This is the word cloud for the tweets which are recorded with the sentiment fear. The tweets with the emotion of fear has words like mandatory, balloting, opposes, trump, Hillary Clinton which concludes that this event is related to politics or the effect of this event will reflect in the results of the politics. People fear that if the voting through mail which 72% of Americans hope doesn't happen, then it will reflect in the election results. They also fear if this doesn't happen then it may be too difficult for the citizens to vote during this Covid Crisis Outbreak.

```
## Warning in wordcloud(freq$word, freq = freq$count, color = rainbow(10), :
## mail could not be fit on page. It will not be plotted.

## Warning in wordcloud(freq$word, freq = freq$count, color = rainbow(10), :
## trump could not be fit on page. It will not be plotted.

## Selecting by count
```



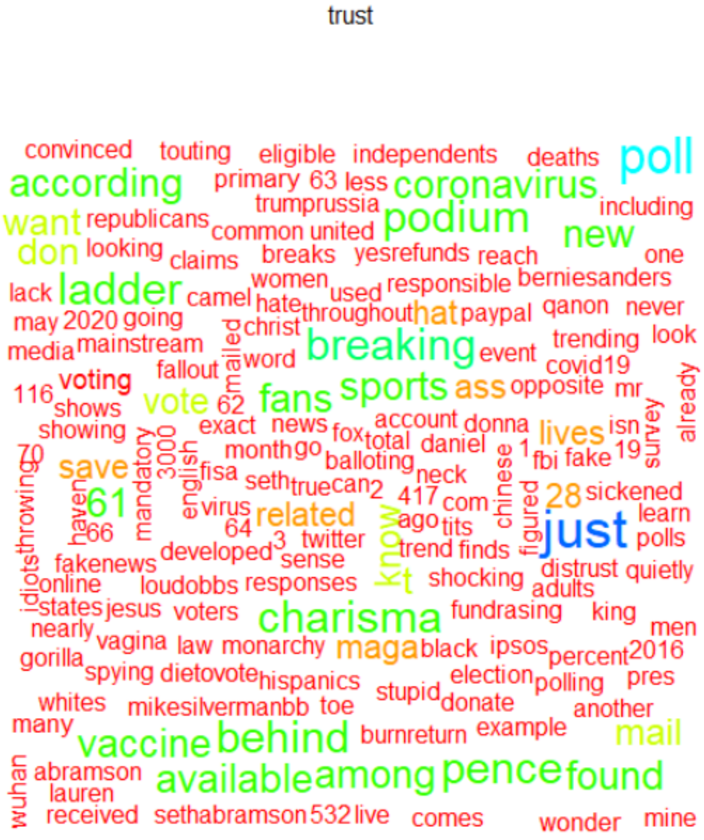
Observation: The word cloud is formed from the tweets which executes the emotion of anger. This shows that the tweets which sentiments anger includes words that opposes and blames the current ruling party, Corona Virus Outbreak, the other political celebrities and the rest 28% people who doesn't agree with this suggestion. The number of tweets with the emotion of anger is the second highest number of tweets recorded.

```
## Selecting by count
```



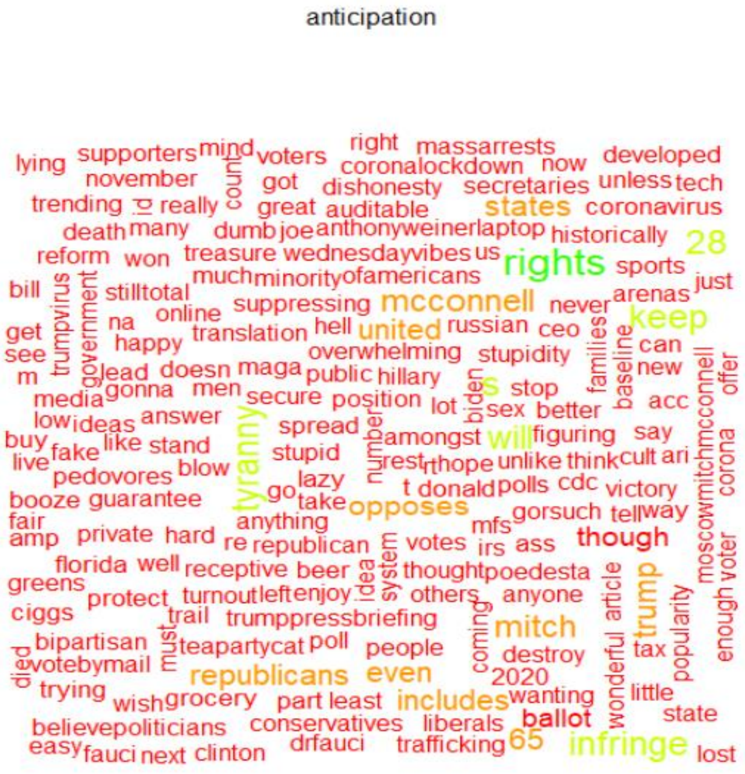
Observation : The tweets with the sentiment anticipation_negated says that the people wants the event to happen but they feel that because of some reason or politics this will not happen. The negation of the hope that the 72% of the Americans wanted to vote through mail. These tweets convey that they have reasons or issues which could stop the happening of the event.

```
## Selecting by count
```



Observation: Trust is the third sentiment with highest number of tweets recorded. People who supports the 72% of Americans to vote by mail have trust that the government will approve this event to happen. They have the trust because of the reasons which we can conclude from the word cloud which includes words like vaccine for coronavirus, breaking news, trump and Russia conflict, China-wuhan countries and some political reasons. They think like it might take some longer time to find vaccine and get a complete cure from coronavirus, conflict between the countries with US and may be since this will affect the political election results, they have a trust that this vote through mail would happen.

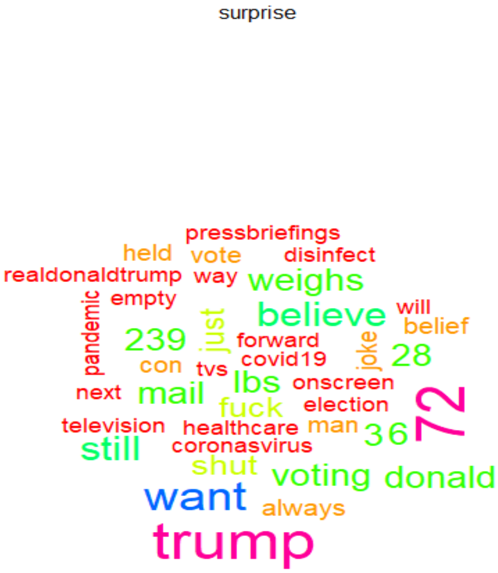
```
## Warning in wordcloud(freq$word, freq = freq$count, color = rainbow(10), :
## moscowmitchmctreason could not be fit on page. It will not be plotted.

## Warning in wordcloud(freq$word, freq = freq$count, color = rainbow(10), :
## republicans could not be fit on page. It will not be plotted.

## Selecting by count
```



Observation:

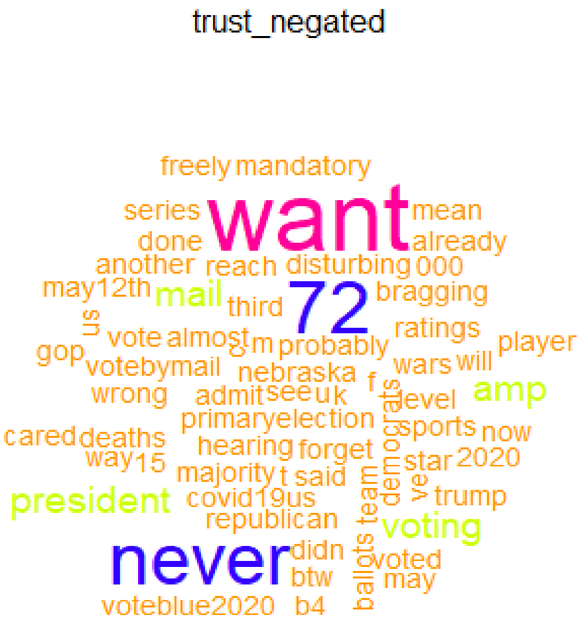
The wordcloud is formed with the tweets with the sentiment of anticipation/hope. They hope that the voting through mail will happen and should happen to overcome this crisis. These tweets involve political parties, celebrities, rights of citizenship, incident of Senate Leader Mitch McConnell during the presidential candidate selection and covid outbreak events that increases the hope that there’re people and reasons to support the voting through mail idea. This sentiment has the highest number of tweets recorded.

```
## Selecting by count
```



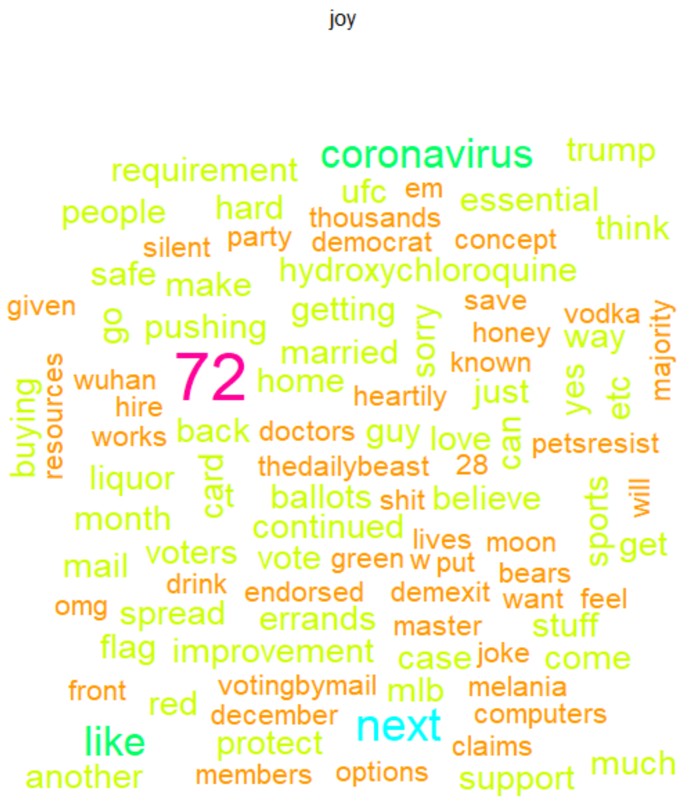
Observation: Some tweets express surprise emotion that the voting through mail idea can be a astonishing event for some people. This includes trump, healthcare and television-press.

```
## Selecting by count
```



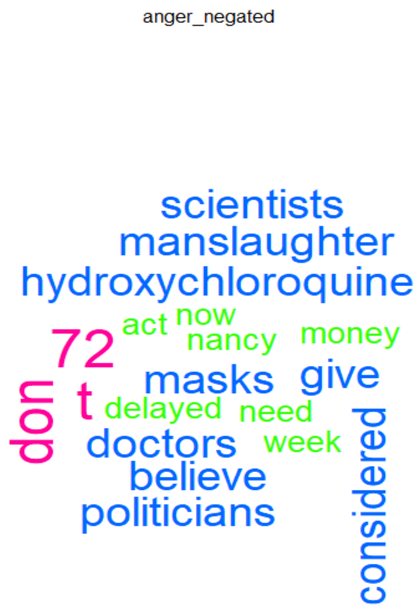
Observation : The trust_negated word cloud is formed from the tweets which includes some facts about mortality rate of covid, republican, democrats and some political reasons. Due to these political reasons their trust may be negated.

```
## Selecting by count
```



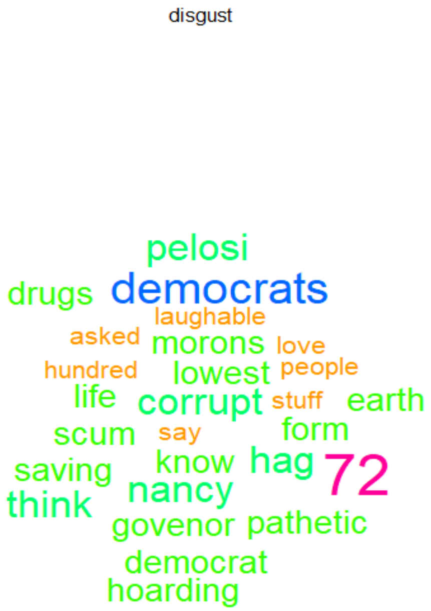
Observation : People express their joy if this event happens. Citizens of America will be happier if they we're given the right to put their vote by mail which 72% of the Americans support. This would create joy for the people like doctors, cops who are struggling to keep the crisis affecting other people. This word cloud includes reasons like computer – work from home professionals, doctors, medical fact – petsresist which can make this event happen.

```
## Warning in wordcloud(freq$word, freq = freq$count, color = rainbow(10), :
## hydroxychloroquine could not be fit on page. It will not be plotted.

## Selecting by count
```



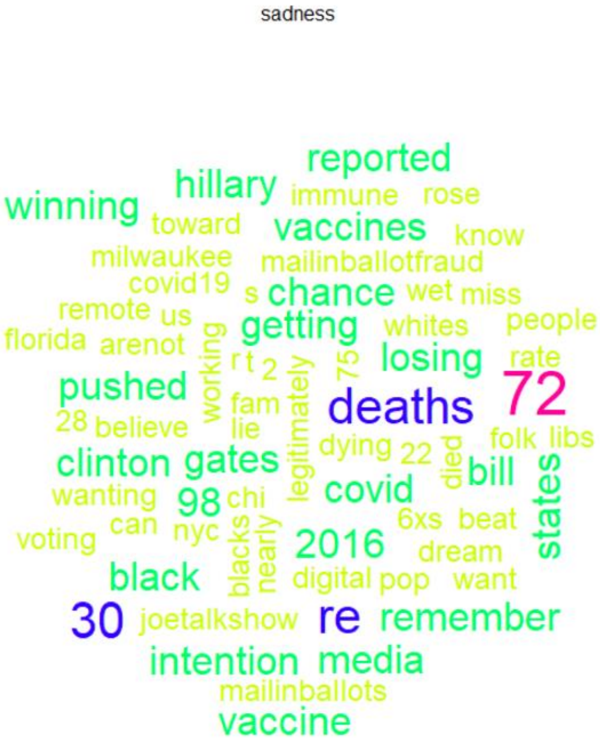
Observation : Anger_negated word cloud includes tweets which discusses about doctors, scientists, malaria curing medicine – hydroxychloroquine and politicians involved in Covid crisis. The people’s anger might be negated because these reason can be support the happening of the event.

```
## Selecting by count
```



Observation : Disgust word cloud includes the tweets which discusses about the Nancy Pelosi’s opinion of not reopening the lockdown, some democrats, governor and Covid drugs. People find these events are disgusting and not an appropriate way to happen.

```
## Selecting by count
```



Observation : The Sadness word cloud includes tweets which discusses about the Political celebrities of US, Hillary Clinton, Covid Vaccine, Media, Businessman and co-founder of Microsoft Bill Gates- Sad thoughts on the Corona and they are afraid that the Mail in Ballot can become a fraud. These tweets expresses the sadness that this idea of vote by mail can become a wrong idea.

CONCLUSION : Based on the analysis of these tweets, we can conclude that the people tweeted with anticipation that this event of 72% of Americans vote by mail will happen. The causes and actions for this event is mainly because of various political events and Corona Virus Outbreak. There are 1289 number of negative tweets, 98 neutral tweets and 613 positive tweets. The number of negative tweets is more than the positive tweets. People consider this issue or event or the topics related to this event is negative.