

## Sentiment Analysis

The sentiment related to the word Vaccine is analyzed on this report. The dataset is taken from the tweets where the word vaccine is mentioned. The following library are used for performing this analysis.

```
library(rtweet)
library(sentimentr)
library(SentimentAnalysis)
library(dplyr)
library(tidyverse)
library(tidyr)
library(ggplot2)
library(tm)
library(wordcloud)
```

We look at 3000 tweets made in real time related to the query word Vaccine. From the tweets, the stop words, hyperlinks and punctuation are removed using the following syntax. In addition, the clean function is used to remove special character and to create a list of 3000 pre-processed tweets.

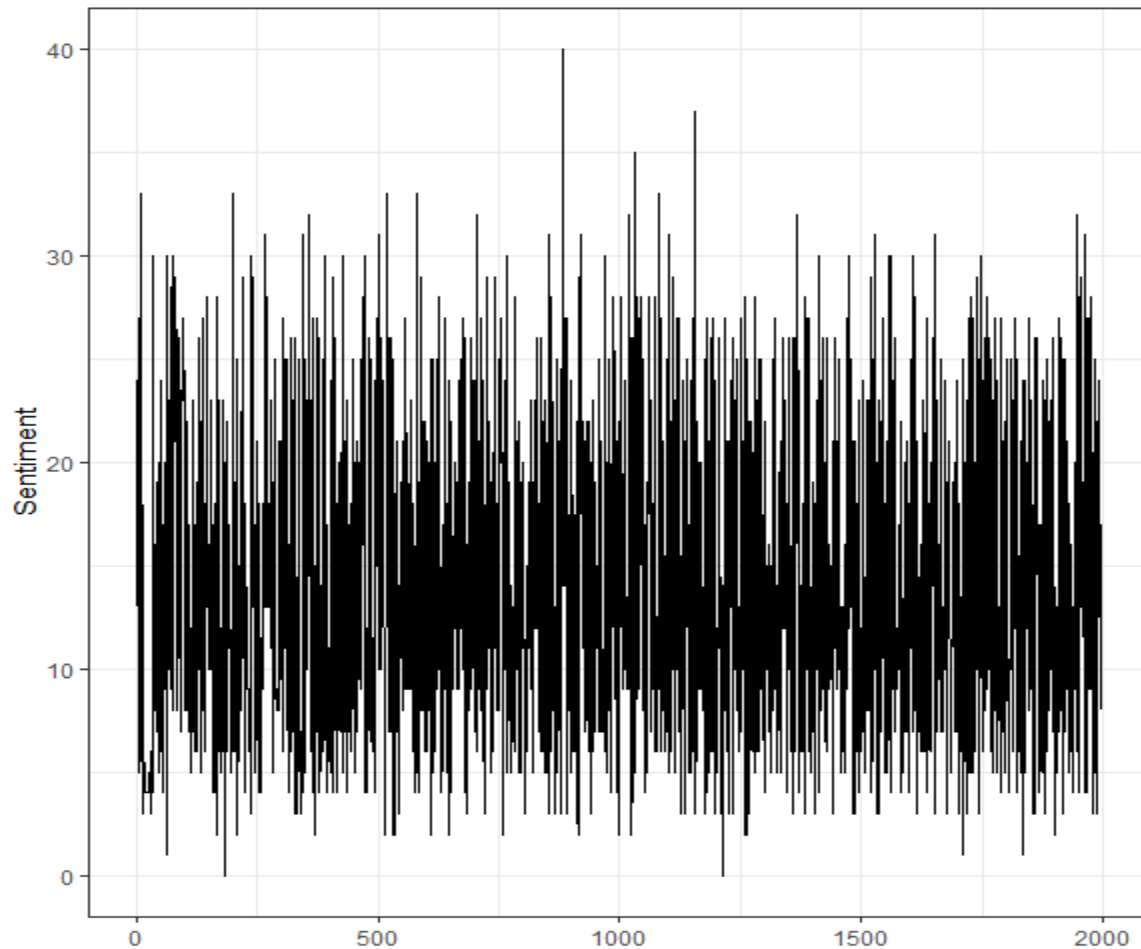
```
Twitter$stripped_text <- gsub("http.*", "", Twitter$text)
Twitter$stripped_text <- gsub("https.*", "", Twitter$stripped_text)
Twitter$stripped_text <- gsub('[:punct:]', "", Twitter$stripped_text)
Twitter$stripped_text <- gsub('[:cntrl:]', "", Twitter$stripped_text)
Twitter$stripped_text <- gsub('\\d+', "", Twitter$stripped_text)
Twitter$stripped_text <- tolower(Twitter$stripped_text)
tweet_list <- Twitter$stripped_text
view(tweet_list)

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)
#View(cleaned_tweets)
length(cleaned_tweets)

> length(cleaned_tweets)
[1] 3000
> |
```

The function analyzeSentiments from the SentimentAnalysis package is used to estimate the sentiments of each tweets. Graph has been plotted based on these sentiments:



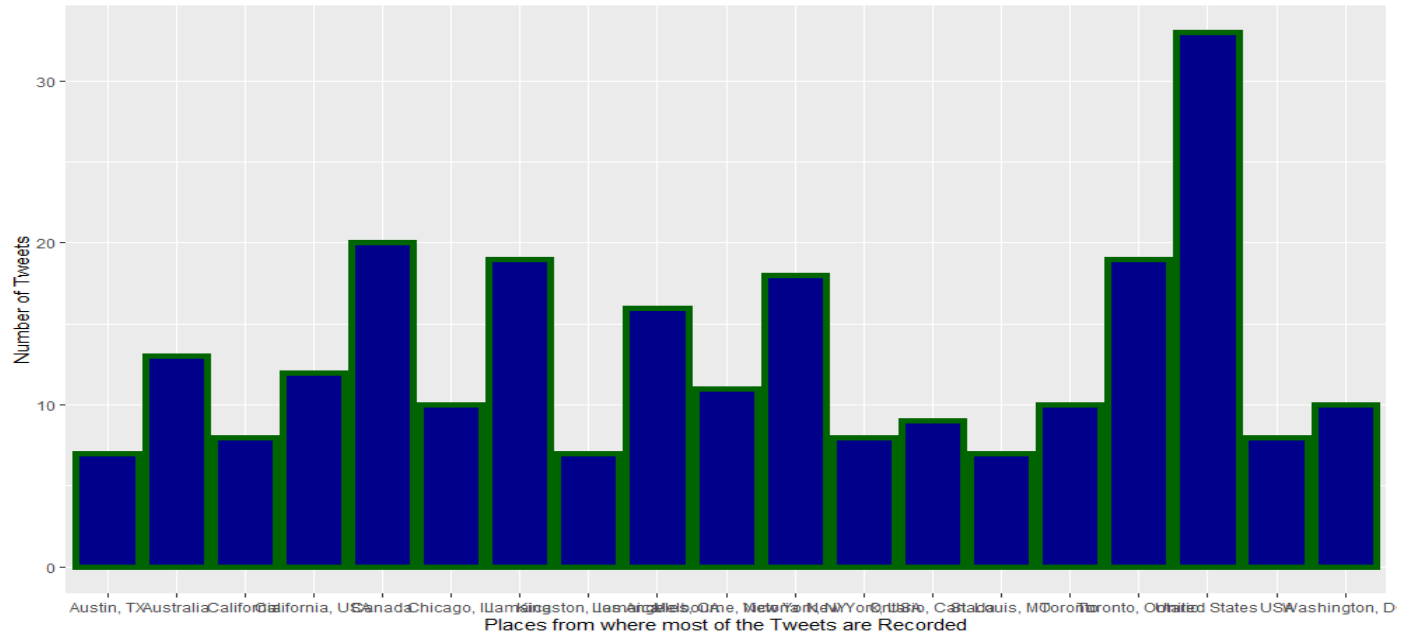
The x-axis represents the number of tweets and the y-axis represents the sentiments of each tweets. The graph shows the mixture of positive and negative sentiments with-in a certain range. Higher value indicates there is a higher number of positive tweets at the given range.

After plotting the sentiments we want to know where the maximum sentiments are coming from, thus we use the following syntax to record the top 20 places. The plots are count of tweets versus the location.

```
location_names <- unique(unlist(Twitter$location),use.names =FALSE)
location_count <- tabulate(match(Twitter$location, unique(Twitter$location)))
location_df <- data.frame("names" = unlist(location_names),"count" = unlist(location_count))
location_df <- location_df %>%
  arrange(desc(location_count)) %>% top_n(20)
location_df <- location_df[2:20,]
maxcount <- max(location_df$count)
```

The plot is shown below:

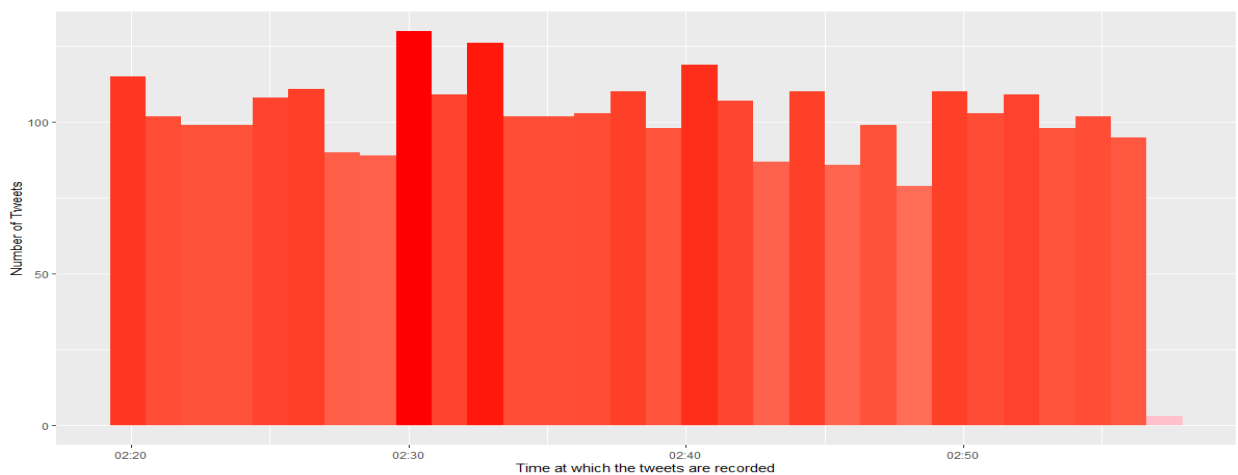
```
ggplot(data = location_df,aes(x =names, y=count))+
  geom_col(col='dark green', fill = 'dark blue', size = 2)+ theme(legend.position = "none")+
  ylim(0,maxcount)+
  xlab("Places from where most of the Tweets are Recorded")+ ylab("Number of Tweets")
```



The plot shows that the highest number of tweets are from the United States and the second highest is Canada. This implies that people from these countries are more interested or concerned with the word vaccine compared to other locations.

We make another plot visualizing the time when the highest number of tweets are recorded. The number of tweets is scaled with the time at which it is posted. This result explains us the time at which the particular topic has been 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')
```

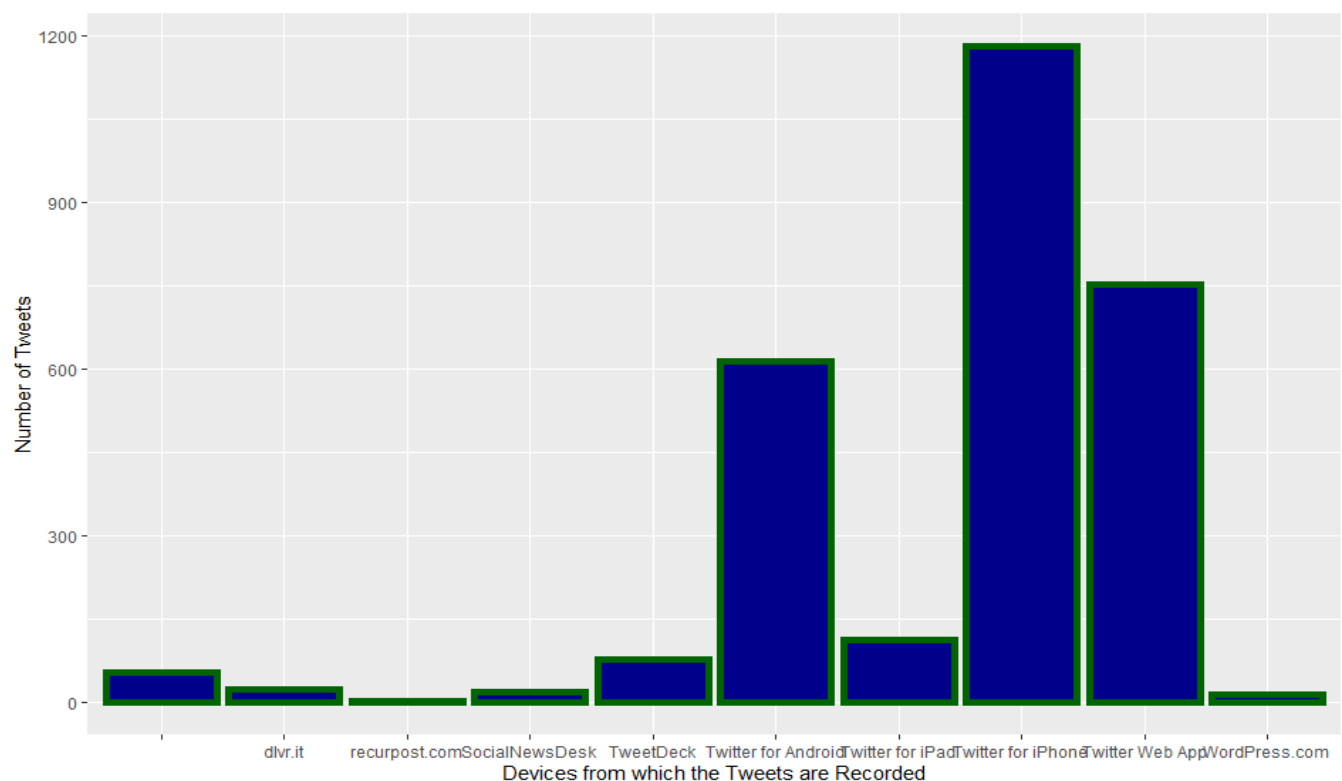


The above plot explains the time and date of the tweet being posted. From the graph above it is clear that the time from 2:30PM to 2:40 PM is crucial as there were many tweets about the word vaccine that other time.

We now take a look at the source through which the user made the post. The following syntax and the plot interpret this result.

```
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]
max_count <- max(source_count)
source_df <- data.frame("names" = unlist(source_uniq),"count" = unlist(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.position = "none")+
  ylim(0,max_count)+
  xlab("Devices from which the Tweets are Recorded")+ ylab("Number of Tweets")
```

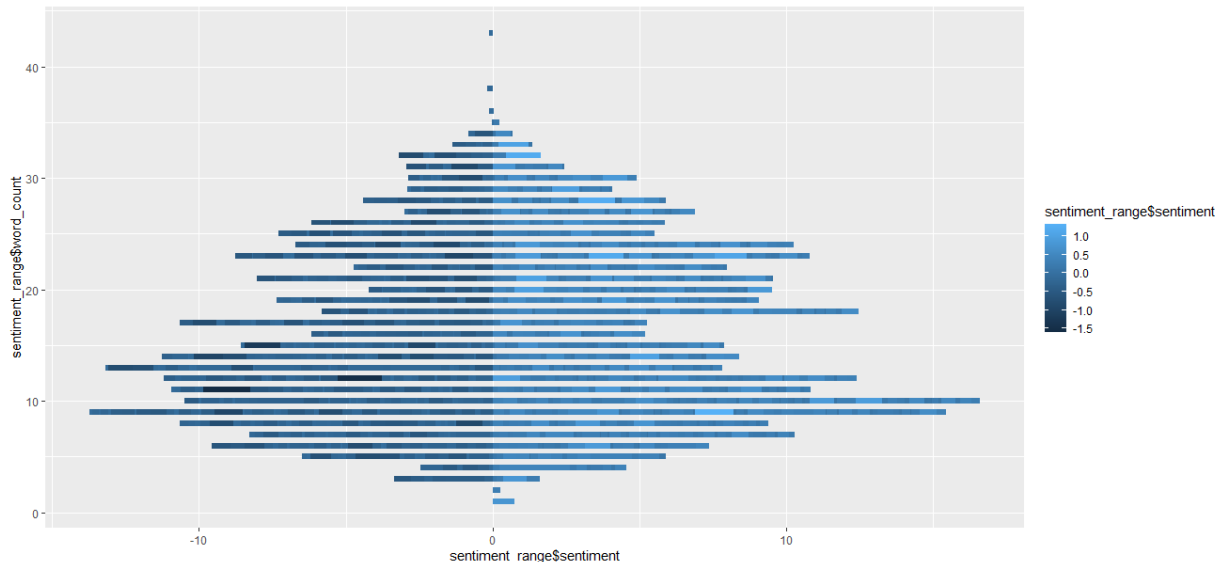


Here, we can see the highest tweets were made from Twitter for iPhone whereas the second and third sources were Twitter Web App and Twitter for Androids respectively. This implies that most people prefer Twitter for iPhone to post tweets compared to Web App and Android.

Now we move towards the sentiment of each tweet in the range of -1 to 1. We use the function `sentiment()` from the `sentiment` package in R. The result obtained from this function is stored in an array of `sentiment_range`. The following syntax and plot show the sentiments of each tweets.

```
sentiment_range <- sentiment(cleaned_tweets)|
count_tweets <- table(sign(sentiment_range$sentiment))
count_tweets
```

```
ggplot(sentiment_range, aes(x=sentiment_range$word_count, y=sentiment_range$sentiment)) +
  geom_bar(stat='identity', aes(fill=sentiment_range$sentiment), width=.5)+
  coord_flip()
```



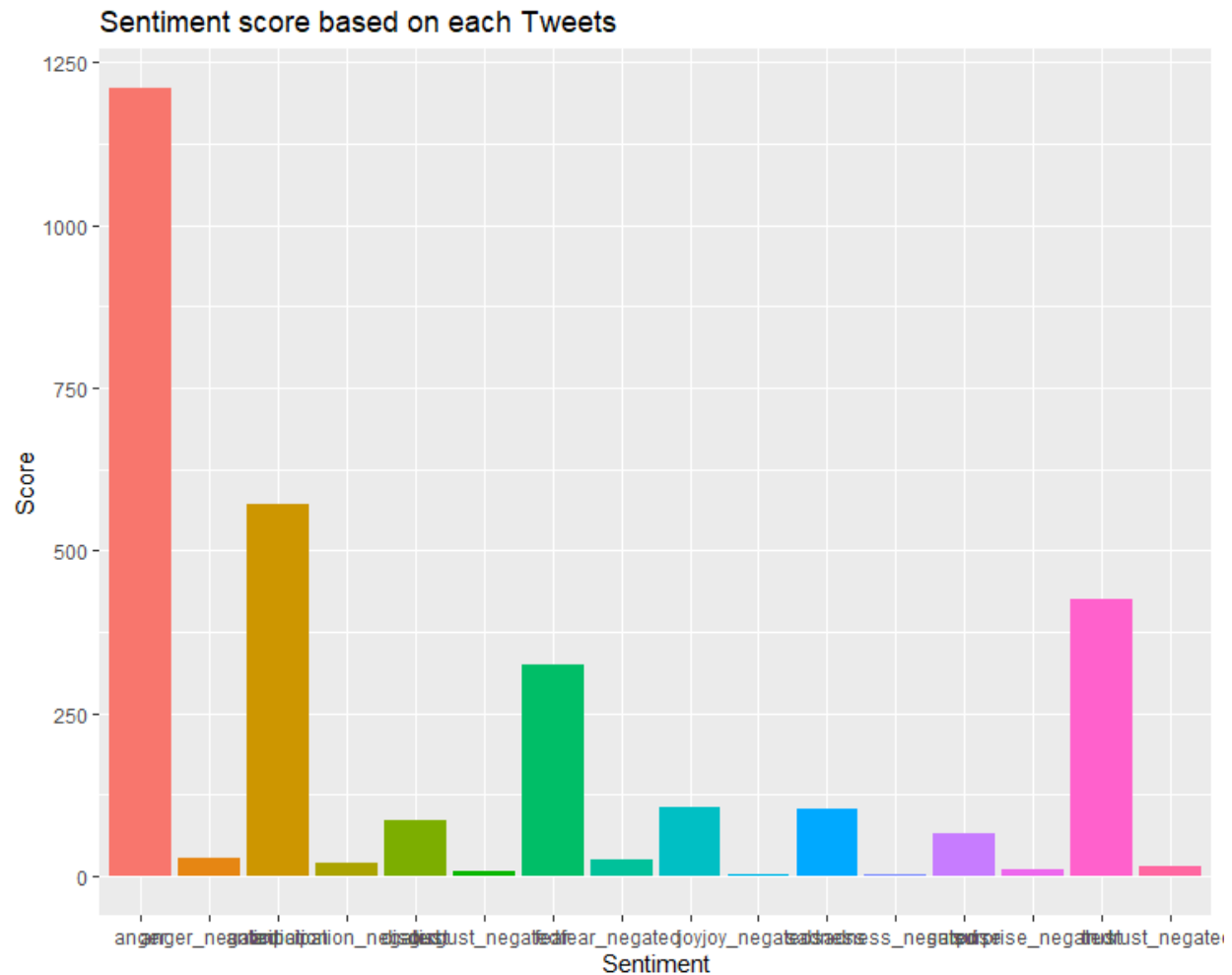
The sentiments for each tweet is plotted with number of each words. This gives us the conclusion that negative tweets have higher number of words than positive tweets. This implies people express negative tweets more elaborately than positive tweets.

We now look at emotion function from sentimentr package. The function categories each tweet into one of these emotions: anger, anger\_negated, anticipation, anticipation\_negated, disgust, disgust\_negated, fear, fear\_negated, joy, joy\_negated, sadness, sadness\_negated, surprise, surprise\_negated, trust, trust\_negated.

```
emotion_range <- emotion(cleaned_tweets)
emotion_range[1:20]
emotion_range <- emotion_range %>% group_by(element_id) %>% filter(emotion_count == 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_range$emotion_type)))
emotion_df <- data.frame("names" = unlist(emotion_unique), "count" = unlist(emotion_count))

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")
```



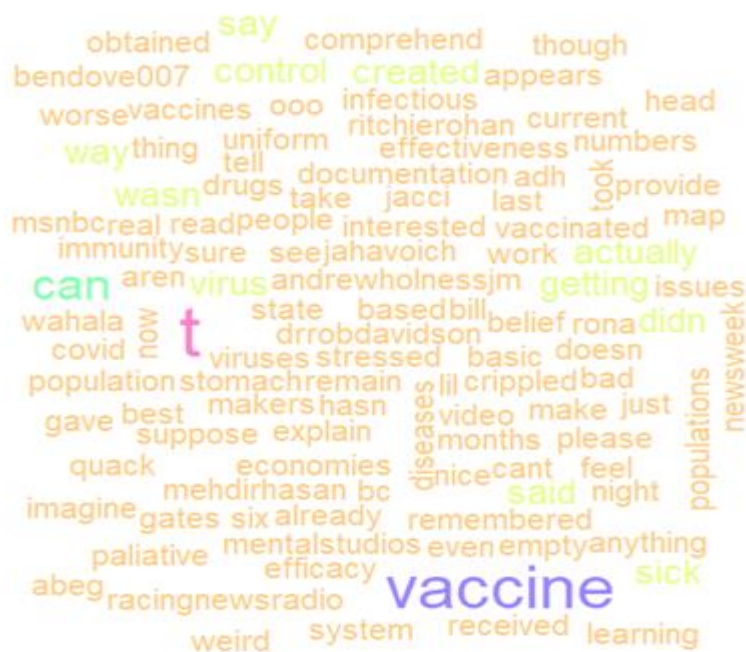
The graph shows that the tweet related to vaccine mostly fall under anger. This implies that most of the people are not satisfied with vaccine that are available. This might be due the fact that the vaccine has some temporary side effects. The second highest is the anticipation and the third is trust. Following after is fear sentiment, which implies most of the people are unsure of taking vaccine.

Now we look at the most occurring words in each emotion. The tweets are grouped and the most used words are used to make a word cloud. Each of the clouds are shown on next page.

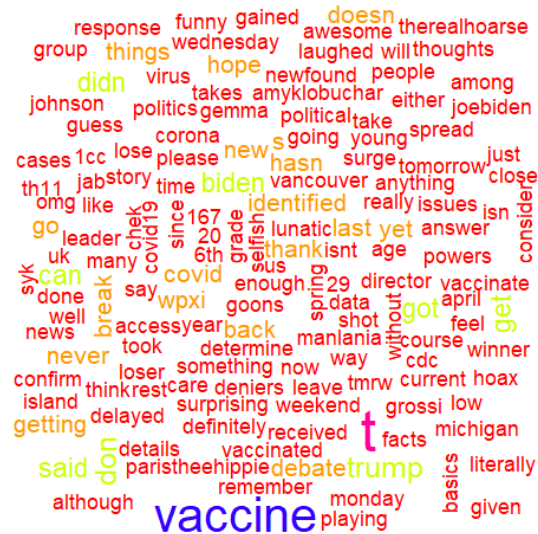
joy\_negated



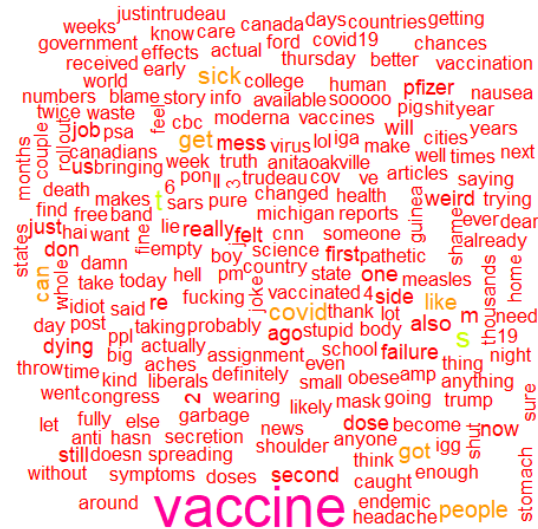
disgust\_negated



surprise\_negated



disgust





## anticipation\_negated



## fear\_negated



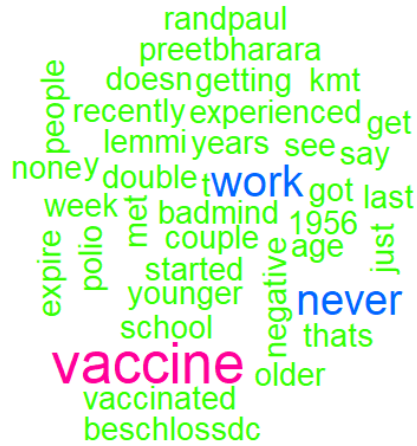
joy

canada research beard thank coronavirus frm yes  
china number completely c autoimmune study better  
literature मणवाद needs seats baba worked  
shot minutes dose special stay nap effects australia  
love get biden amp believe side us evening sitting  
able will available went returning people kids true  
differently week south live due appointment dolly 100 15 deal  
thankful we favorite much oh county sure trump happy  
talk antibodies health 60 chinese food mask iridden hope man  
left anti 60 well now age cent freemoved month 30  
m antibodies grateful roll necessary look billion enough nurses cbsnews  
saheb normal can lesstials fda said re need fine world light  
uk baby efficacy want treatment some 2 50 6 already world beach  
sydney data safe really see told photo read sphule variant already take say  
may help days thanks lol go indonesia save feel put tattoo  
protect area think kaz day m loving god like lot laugh  
news india florida स्व 4 months though yet per second parton  
doctors even appointments lives young a second way  
pfizer helpful soldiers ontario working hug just know  
without break biontech uwillow lucky helping palm country  
passport share mahatma टीकाकरण darling gets made still  
don first virus rondesantis effectiveness 19 vaccine  
celebrate covaxin

anger\_negated

right makes rennysonbrick disappointment purchase let  
already personnel documentation due never jake  
jwong global news didn fortunate can prevent make calibrate florida  
body matters start track whose michigan force away 20 white  
science anybody consists cons levinedrew anything disease military poisoned  
soon nursing shit though thanked number vaping going  
states news back family 5 cns notufirst run multitask  
less oh powder skinny re wwe lie criteria t  
one bcndp work cell faux deaths end lot fall cnn  
will pop drtessat work like selfless live t use raw after mama  
loool numbers health narwhal lord duck shots nypost line  
lack crap fight home care means gotd poison ny post everyone little  
trail filled now major jaketapper 3 restored evgt bind governor  
infertile smoking vaccinesaf thank gene hit 16 wouldn  
wen sue access ve hesitant mask doesnt god aside  
dying president opinion s sprayed interest know idiots  
texas except yes globaled monton truly bs still world  
immunity heroin expected bout taxpayers meet  
blood weapon forward shoot jumped visited grab really  
keep federal johnson social others want labs  
benshapiro appointment connaught biden kurokawa m  
lunus of gay takes thousands think gloated matter bad  
either huge clots sore

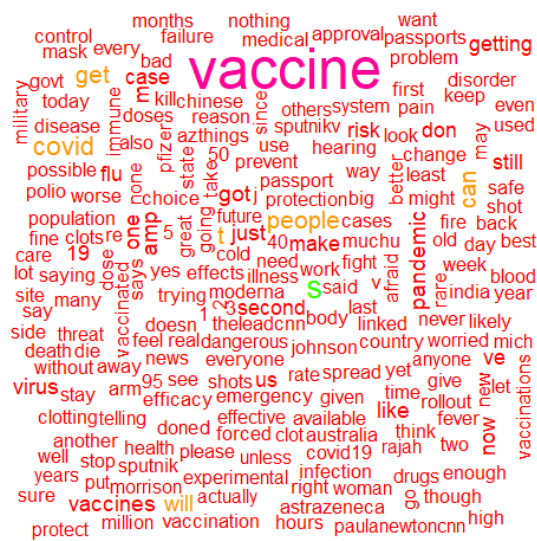
sadness\_negated



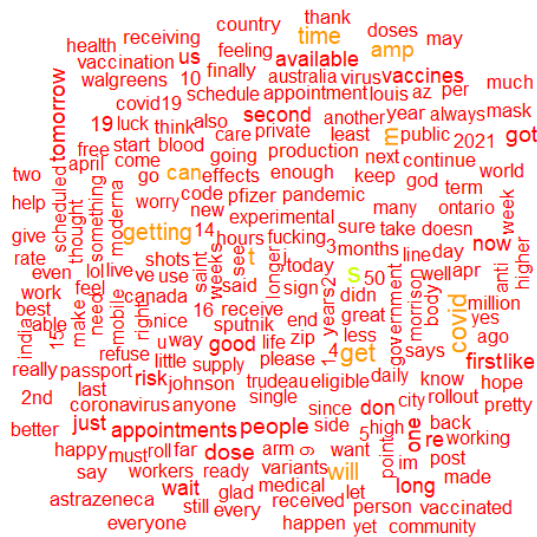
sadness



fear



anticipation



surprise

red opportunistic master excelsior ukmeh state  
article considering globeandmail nytimes  
oann acute gop takes anyways ur don tk oh  
aids supporters 19 approved joke metallic  
fever means west taking d based 13 fucking fda  
crisis needs seems punk russian still except won  
inc jeff society insurance india asked much  
seen rest 16 therapeutics t gave cells  
im 80 gov variable shawty j reg usa mandatory  
m hold take 3 son 0 reddy 1 2 freedom folks  
months astrazeneca arrest cruise man 90  
brrrr studies guess elusive rules grocery bar  
floating aca tomilahren typed v  
S charge dr promising antibodies nj fatigue  
ve wild deccanherald unheard nod africa  
pin percentage remembered fox sucks residents  
wants

anger

may australia t something disease  
m covid19 last someone people virus  
fuck way 1 news appointments great go  
ass hit early blood theleadcnn know  
S taking re 4 providence u week us say d  
passports 3 little deaths without want  
yeah already two isn myturn got ve please y  
first make look 5 end health let  
days yet moderna uk ll best roll  
2 anti gov 7 19 j ca will 04 pm 2021 v positive  
don many every anyone federal mrna  
mask soon 9 going adverse 50  
weeks next months come vaccinated else  
need much made available really im  
12 says experimental sure now buggy

trust



## **Conclusion:**

Based on the analysis we can conclude that people tweeted more briefly about the negative aspects of vaccine. Most of the tweets have anger associated with it. This might be due to improper management of vaccine supply or the temporary side effects that are caused through it. In a gist people seems to have negative sentiment towards the word vaccine.