Kritagya Nepal(kxn190007)
CS 6301.003
Assignment3

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.

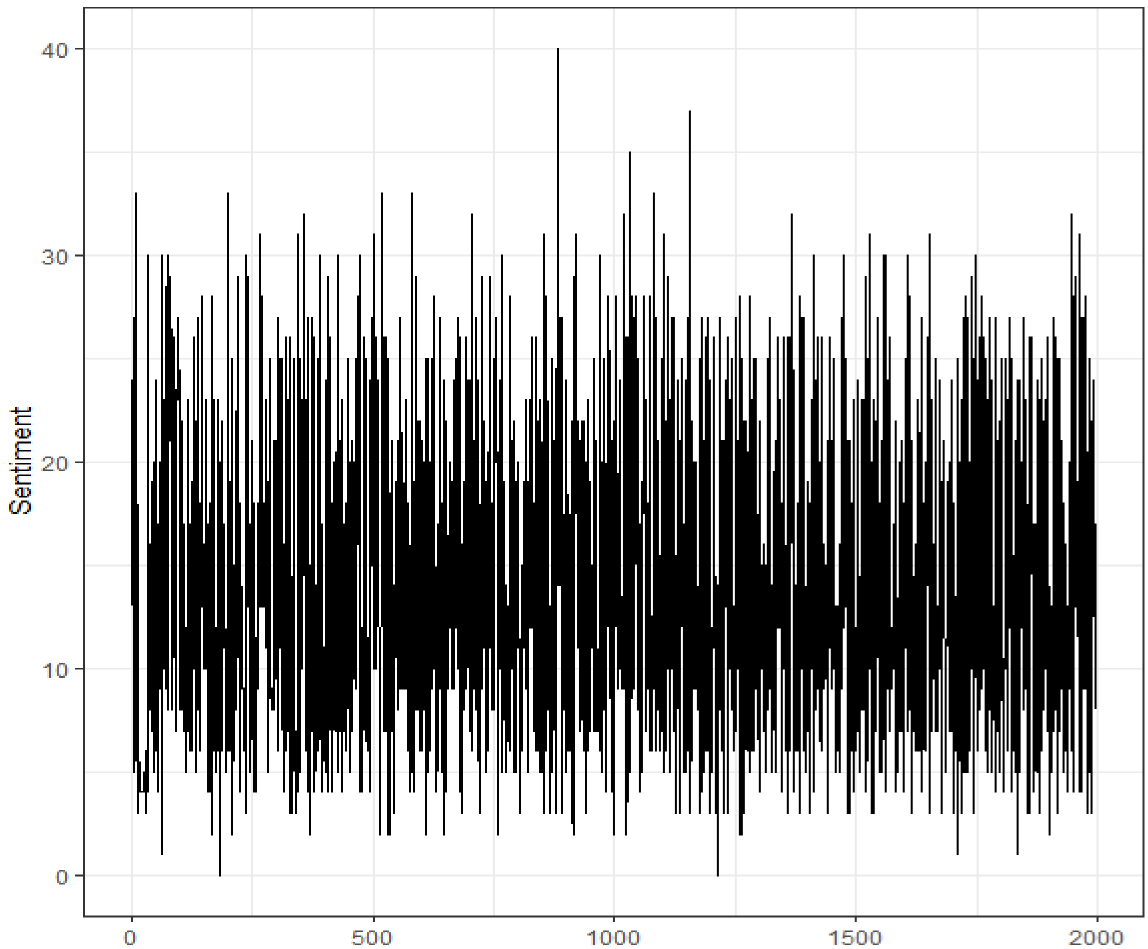
```
Twitter$stripped_text <- gsub("http.*","",Twitter$text)
Twitter$stripped_text <- gsub("https.*","",Twitter$stripped_text)
Twitter$stripped_text <- gsub('[[:punct:]]','' ,Twitter$stripped_text)
Twitter$stripped_text <- gsub('[[:cntrl:]]', "" ,Twitter$stripped_text)
Twitter$stripped_text <- gsub('\\d+',''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



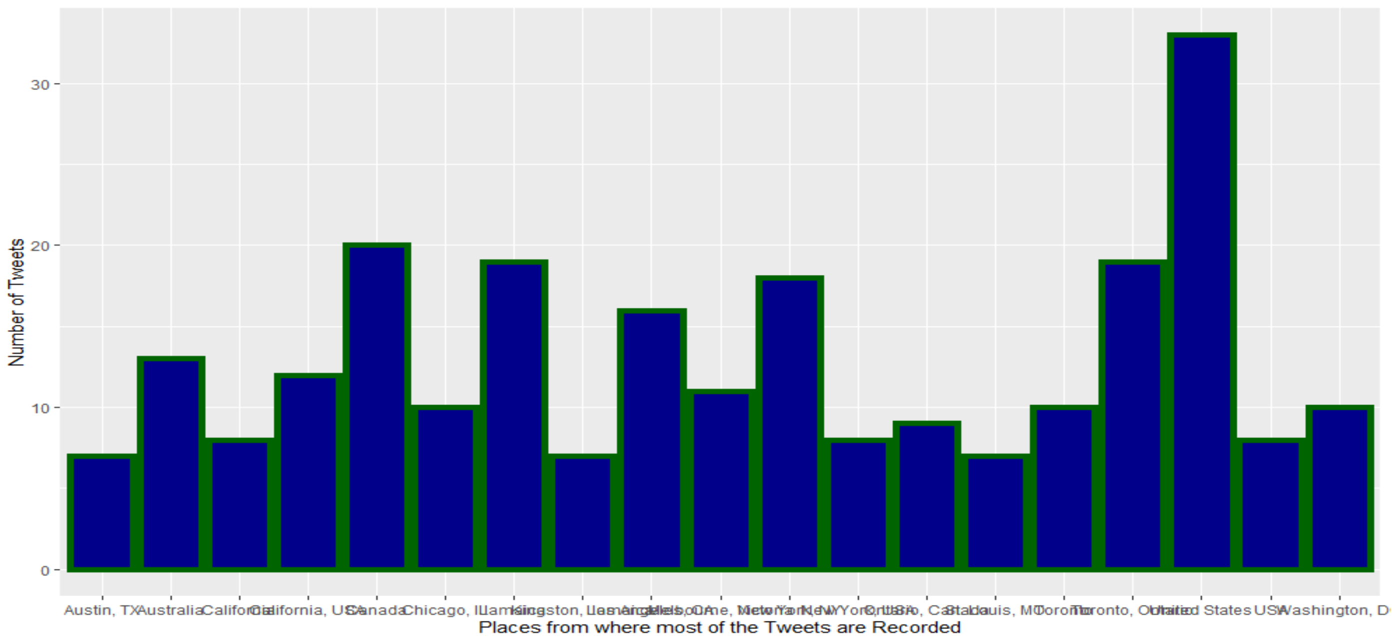
The x-axis represents the number of tweets and the y-axis represents the sentiments of each tweets. The graph shows the mixture of positive and negative sentiments with-in a certain range. Higher value indicates there is a higher number of positive tweets at the given range.

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:

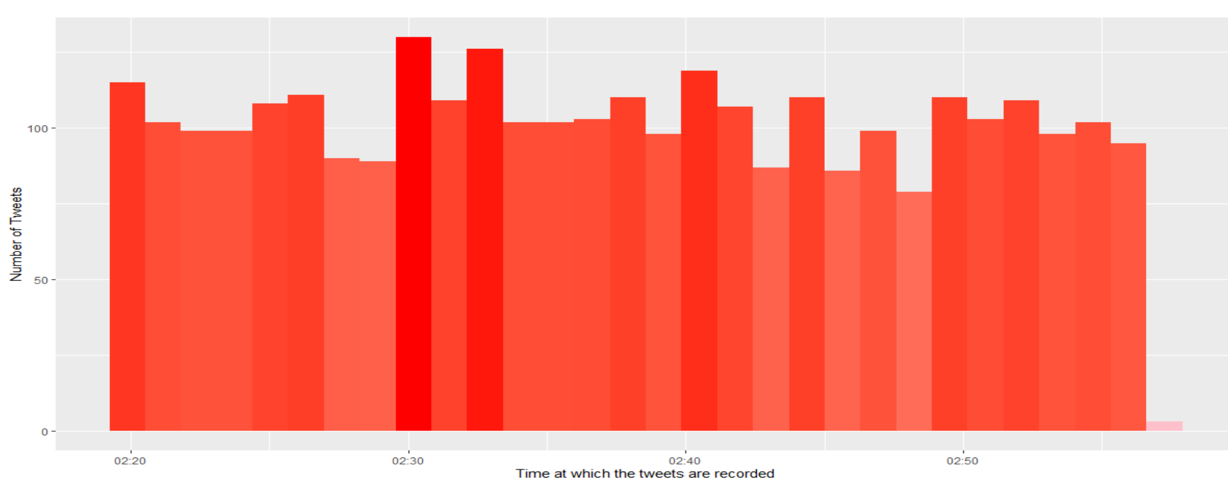
```
ggplot(data = location_df,aes(x =names, y=count))+
  geom_col(col='dark green', fill ='dark blue', size = 2)+  theme(legend.position = "none")+
  ylim(0,maxcount)+
  xlab("Places from where most of the Tweets are Recorded")+ ylab("Number of Tweets")
```



The plot show that the highest number of tweets are from 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 location.

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')
```

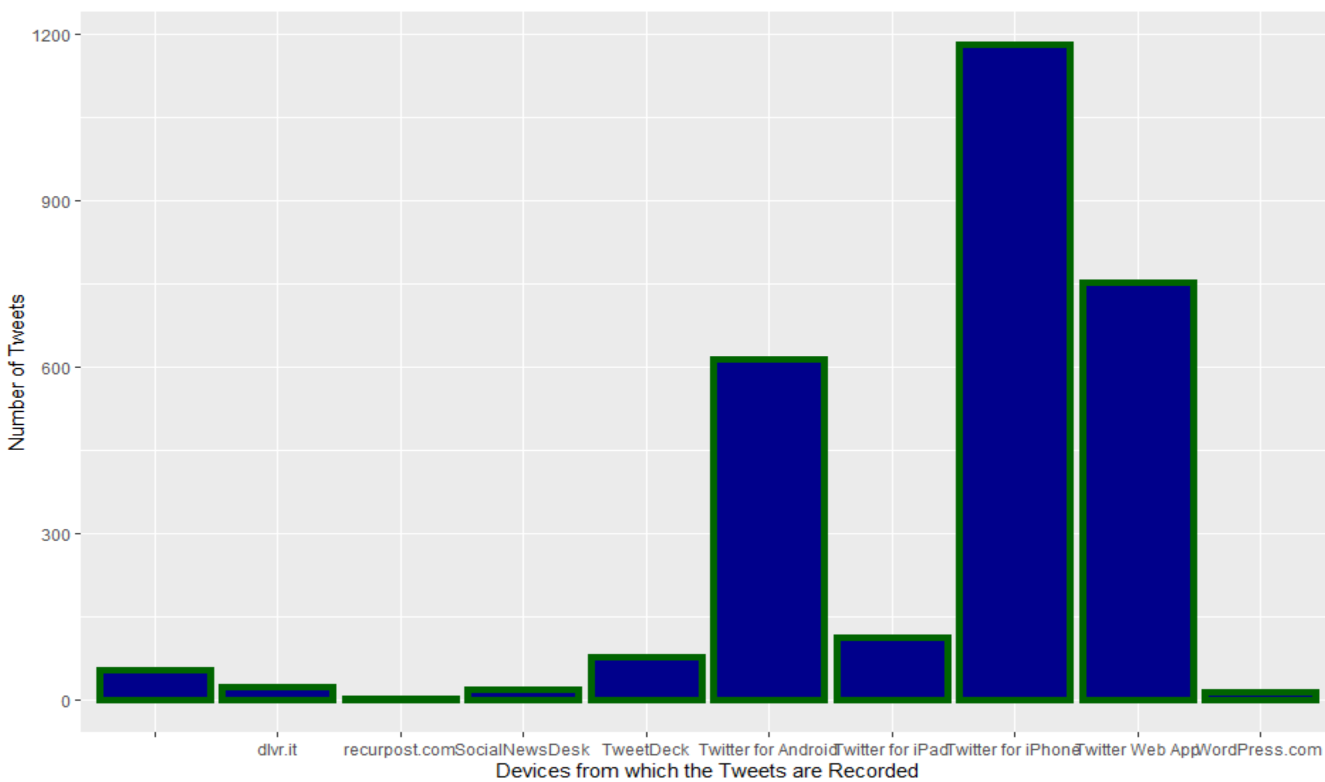


The above plot explains the time and date of the tweet being posted. From the graph above it is clear that the time from 2:30PM to 2:40 PM is crucial as there were many tweets about the word vaccine that other time.

We now take a loot 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")
```

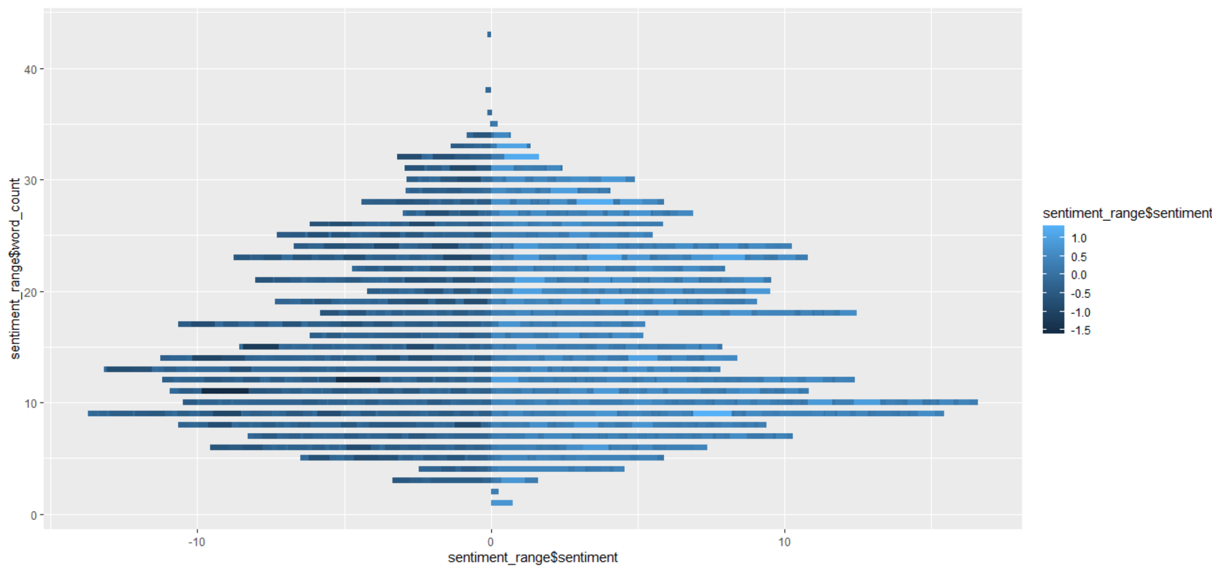


Here, we can see the highest tweets were made from Twitter for iPhone whereas the second and third sources were Twitter Web App and Twitter for Androids respectively. This implies that most people prefer Twitter for iPhone to post tweets compared to Web App and Android.

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()
```



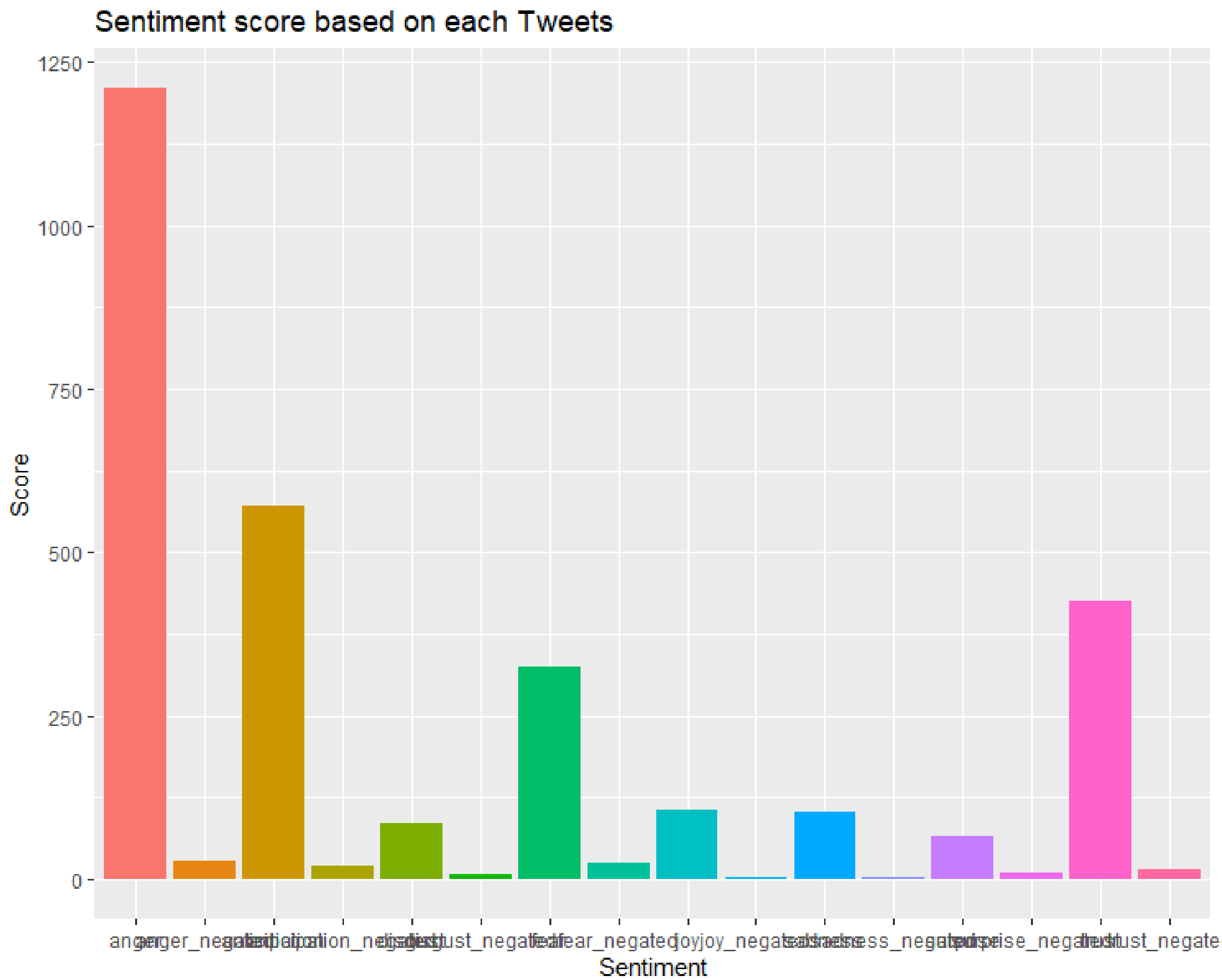
The sentiments for each tweet is plotted with number of each words.This gives us the conclsion that negative tweets have higher number of words than positive tweets. This implies people express negative tweets more elaborately than positive tweets.

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")
```

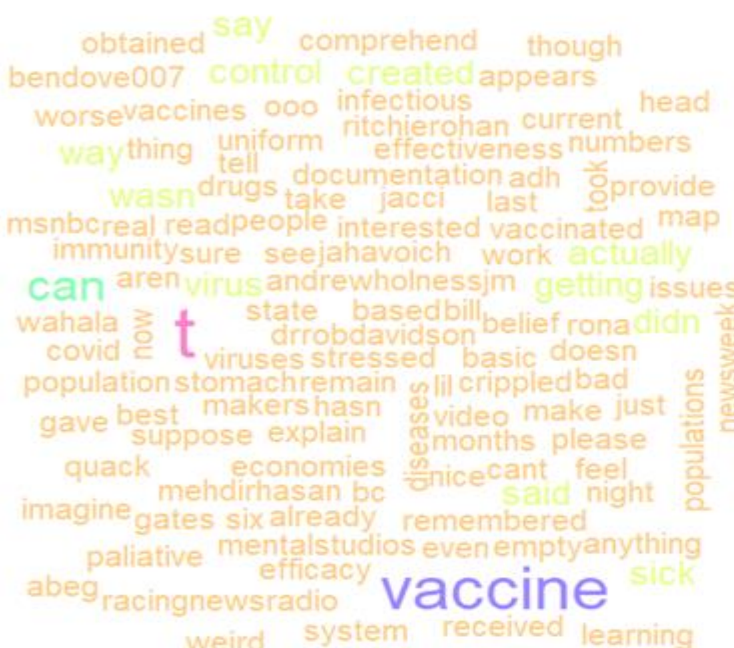


The graph shows that the tweet related to vaccine mostly fall under anger. This implies that most of the people are not satisfied with vaccine that are available. This might be due the fact that the vaccine has some temporary side effects. The second highest is the anticipation and the third is trust.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

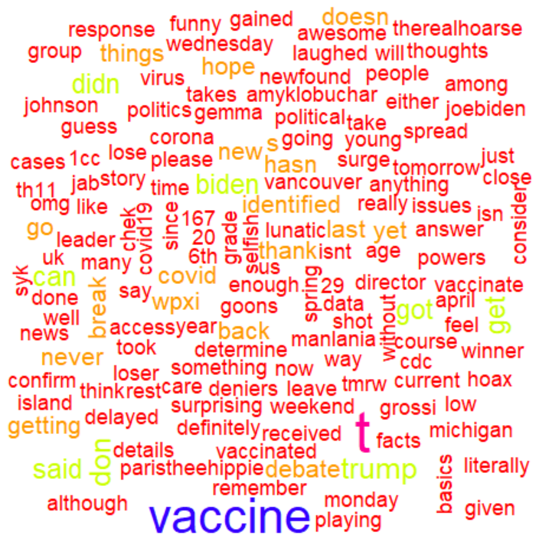
Now we look at the most occurring words in each emotion. The tweets are grouped and the most used words are used to make a word cloud. Each of the clouds are shown on next page.

joy_negated

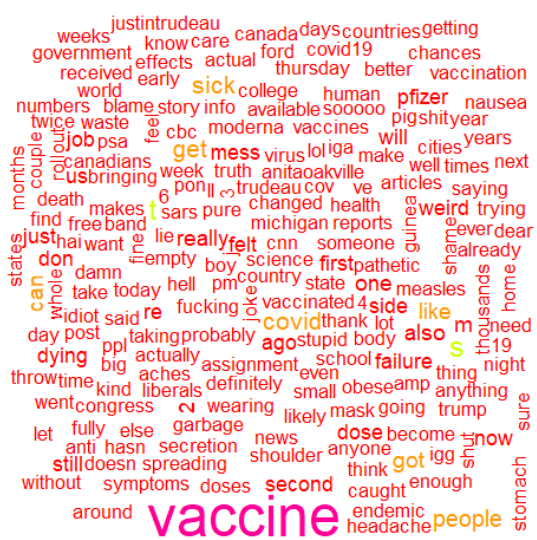
disgust_negated



surprise_negated



disgust



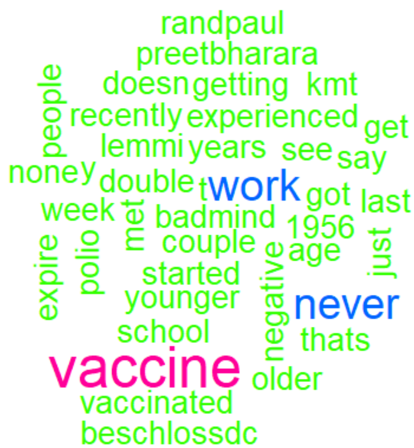
anticipation_negated

fear_negated

joy

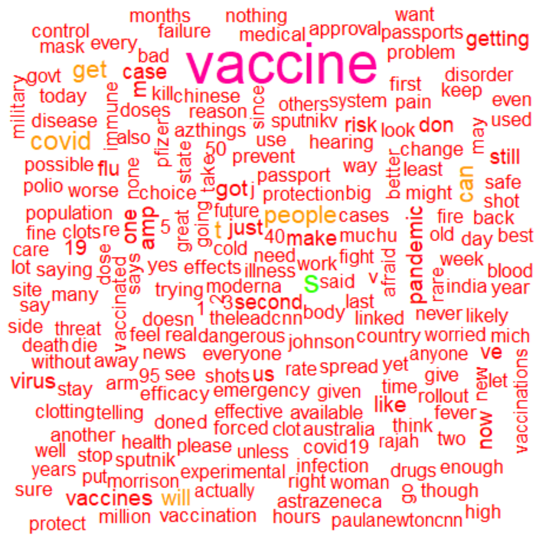
anger_negated

sadness_negated



sadness

fear



anticipation

surprise

anger

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.