

# Final Project

*Jiawei Li*

*2017/12/18*

## Text analysis of

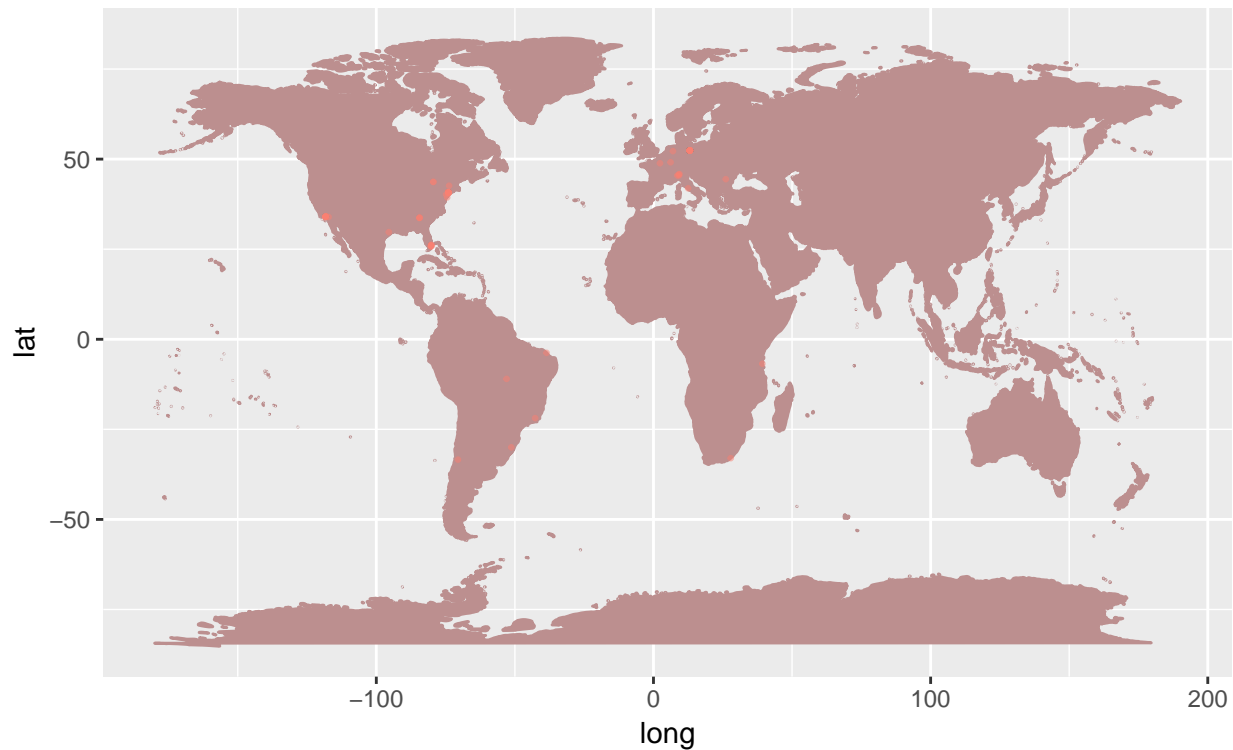
Topic #Rihanna Topic #Taylor Swift

### Search Topic

First search 5000 results of the topic #Rihanna from Twitter. It looks like most of the data points I collected are in the area of Europe and United States.

```
## [1] 85
```

### The Map of the Data Points–Rihanna



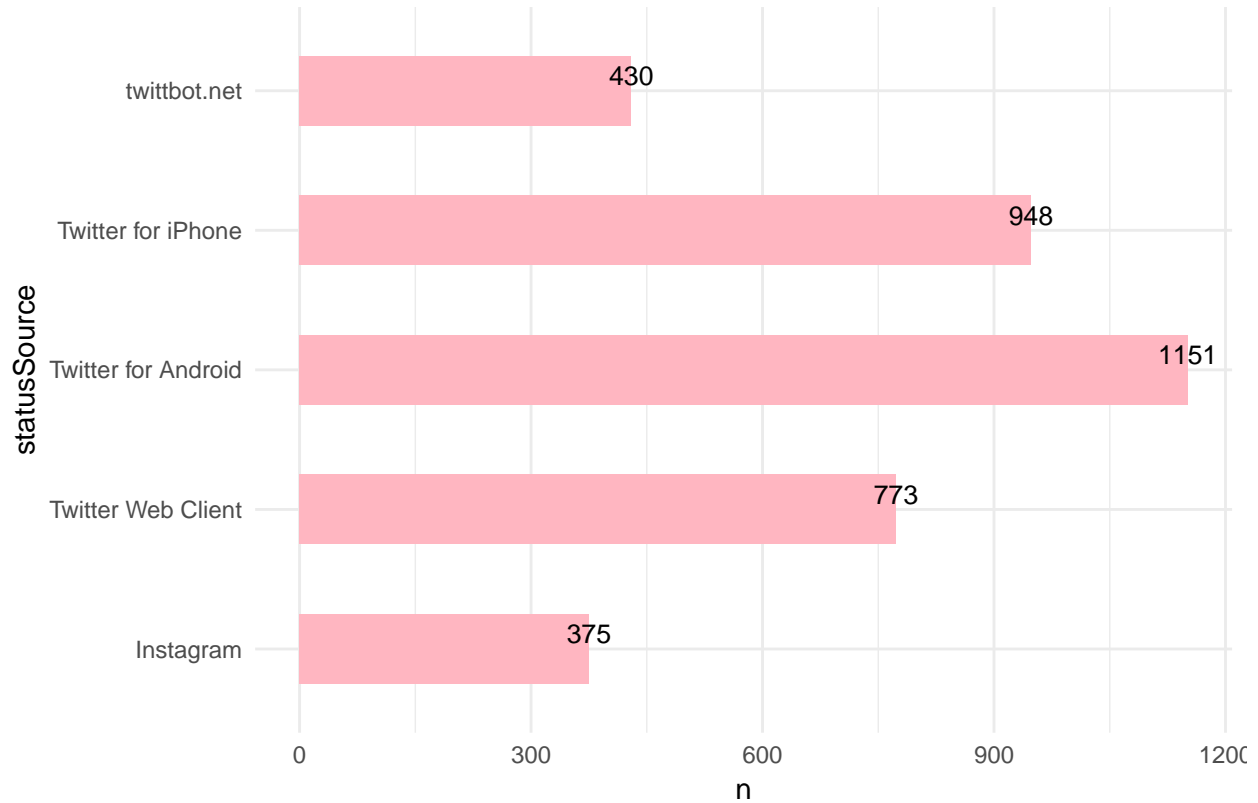
### Calculate the count of tweets by platform.

Here is the barplot of the count of the source. Most of the tweets are from Android, and then is Iphone.

statusSource	n	percent_of_tweets
Twitter for Android	1151	0.230
Twitter for iPhone	948	0.190
Twitter Web Client	773	0.155

statusSource	n	percent_of_tweets
twittbot.net	430	0.086
Instagram	375	0.075

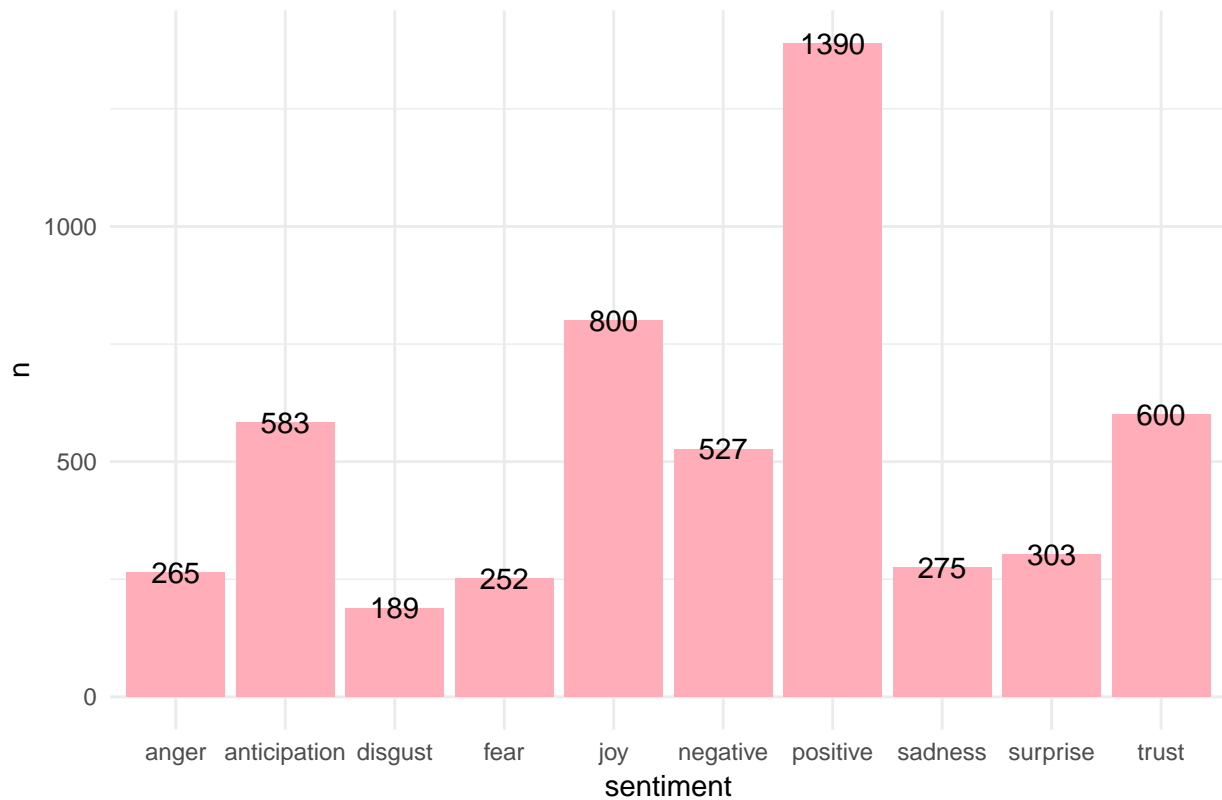
The Counts of the Data From Different Platform–Rihanna



### Sentiment of total data

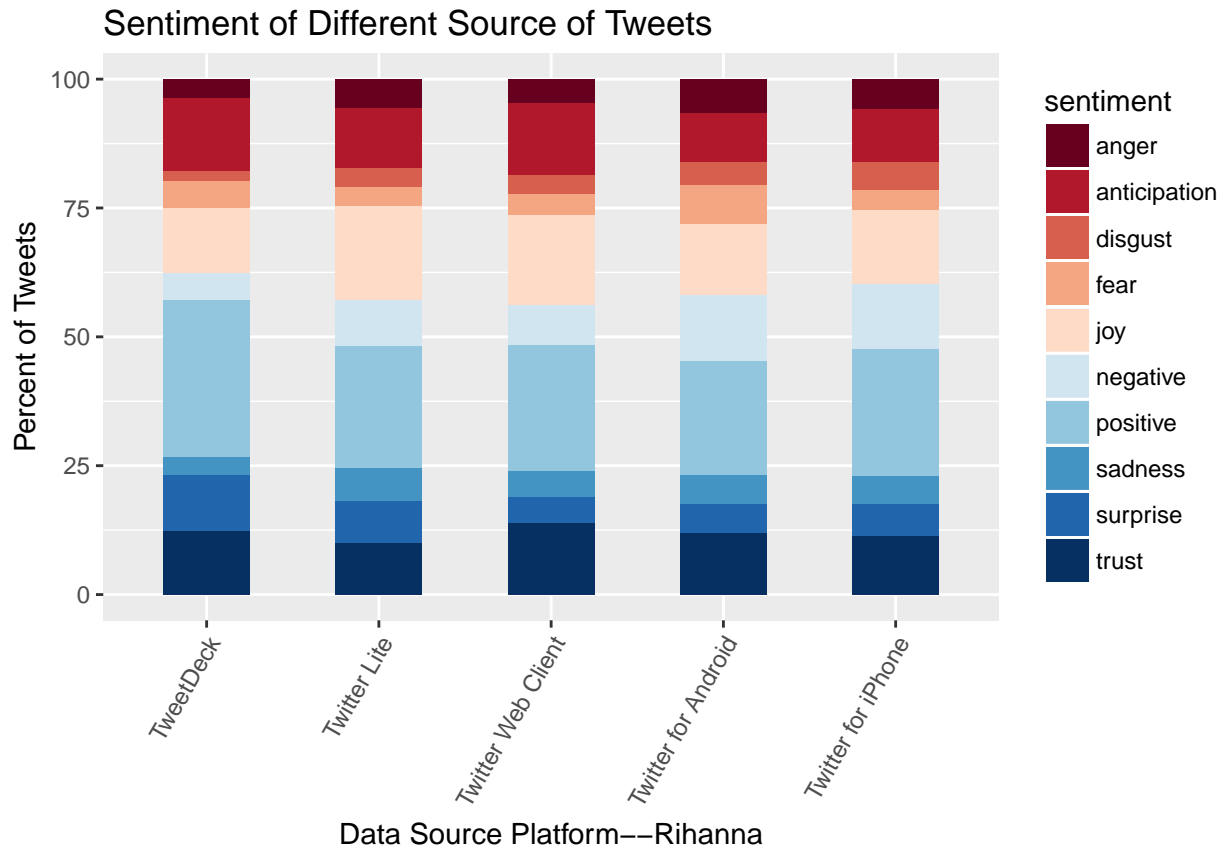
After clean the data, plot the Sentiment of Topic #Rihanna in total. We can see that most of the words are positive in the topic of #Rihanna. The least is disgust.

The Plot of Sentiment words--Total(Rihanna)



### Sentiment of top 5 data source

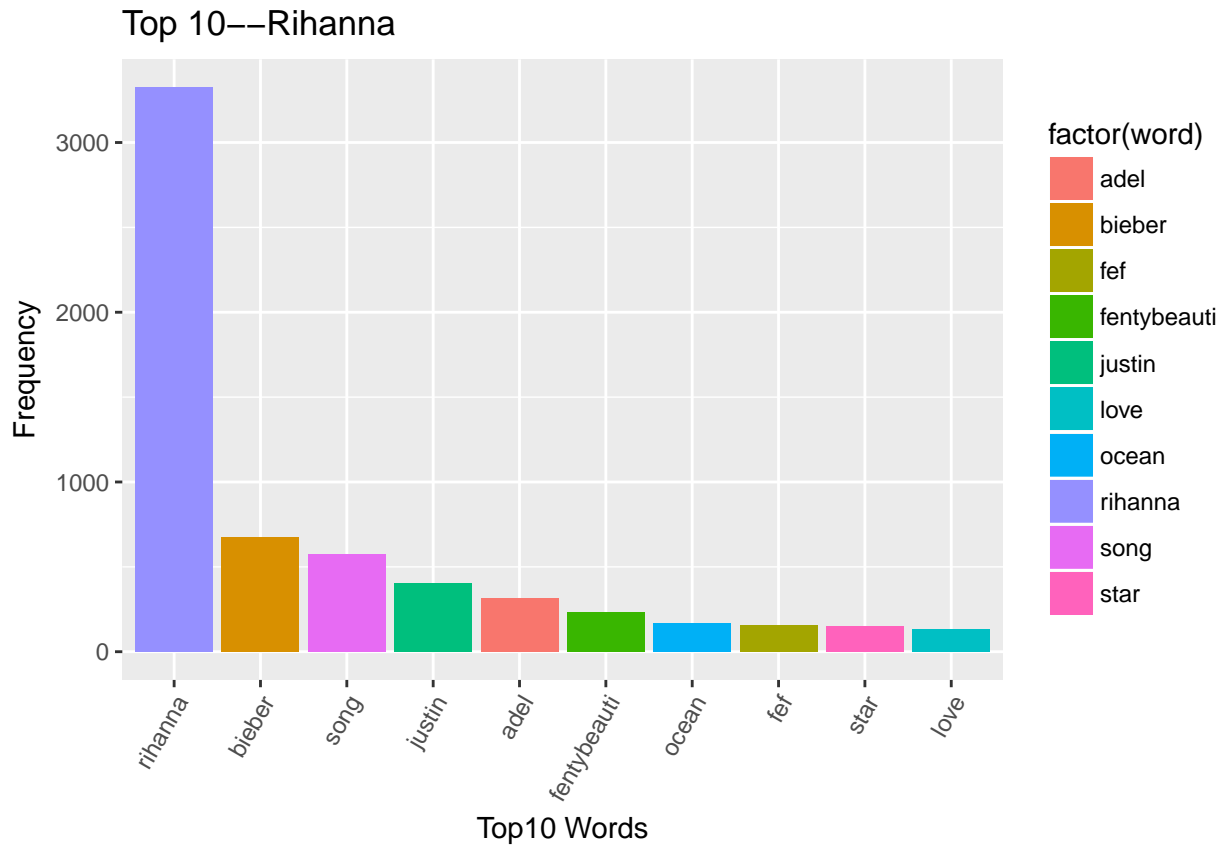
Calculate and plot the Sentiment of Topic #Rihanna from different data source platform. By comparing the sentiment from different data platforms, we can see that positive and joy words have biggest part in the bar.



### Wordclouds Analysis--Top 10 words in the Topic(#Rihanna)

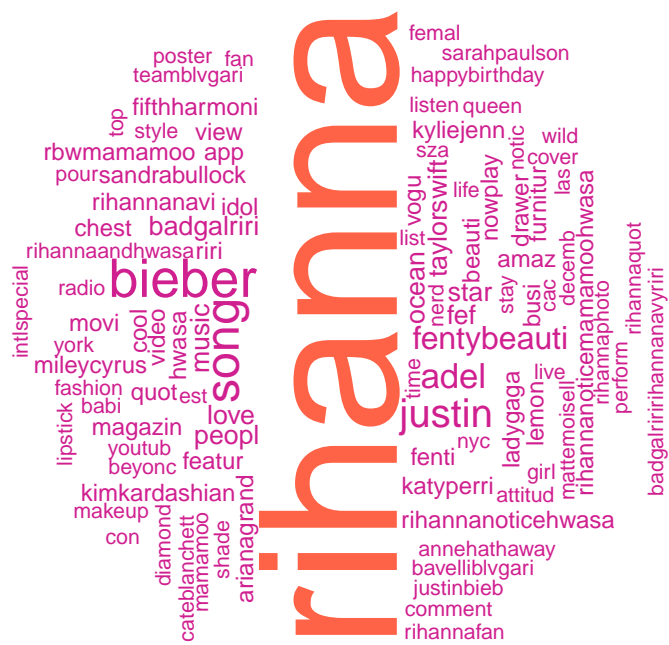
Plot the word clouds of #Rihanna. Top 1 word is rihanna, which is obviously. But it is surprise that the second word is bieber. We can also see that fentybeauty—the cosmetics brand established by Rihanna herself is also in top 10.

```
## chr [1:39621] "videos" "100m" "views" "#rihanna" "#taylorswift" ...
```



## The Word clouds

Here is the word clouds of the topic #Rihanna.



# Analysis of Taylor Swift

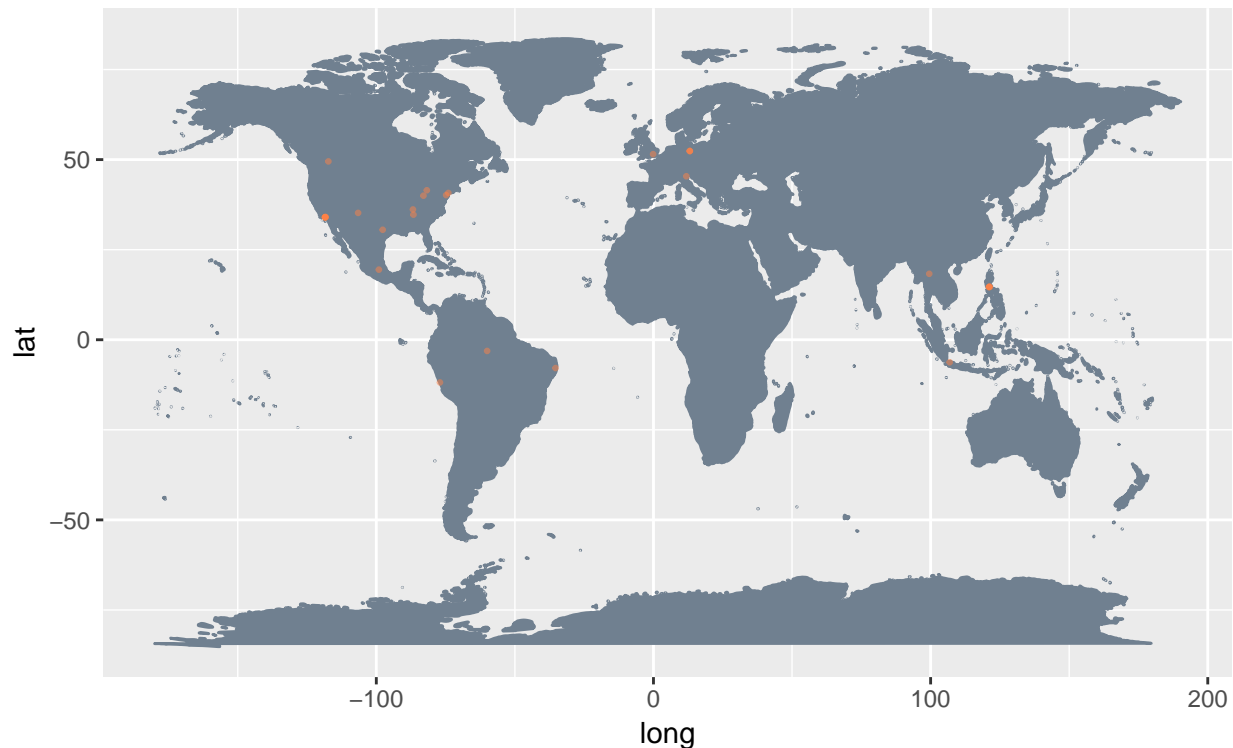
## Search Topic

First search 5000 results of the topic #Taylor Swift from Twitter. It looks like most of the majority part of points I collected are in the United States. There is no points in the Africa.

```
aT0 <-read.csv("TS.csv",header = T)
aT3 <- -1*is.na(aT0$longitude)+1
sum(aT3)
```

```
## [1] 32
```

```
aT4 <- which(aT3==1)
locationaT1 <-data.frame(aT0$latitude[aT4],aT0$longitude[aT4])
colnames(locationaT1) <- c("latitude","longitude")
locationaT1$longitude<-as.numeric(locationaT1$longitude)
locationaT1$latitude <-as.numeric(locationaT1$latitude)
map.world <- map_data(map = "world")
ggplot() + geom_polygon(data = map.world, aes(x=long, y = lat, group = group),color="slategray",fill="slategray") +
  coord_fixed(1.3)+geom_point(data= locationaT1,
                              aes(x = longitude, y = latitude),
                              colour = "sienna1",
                              alpha = 0.5,
                              size = 0.5)
```



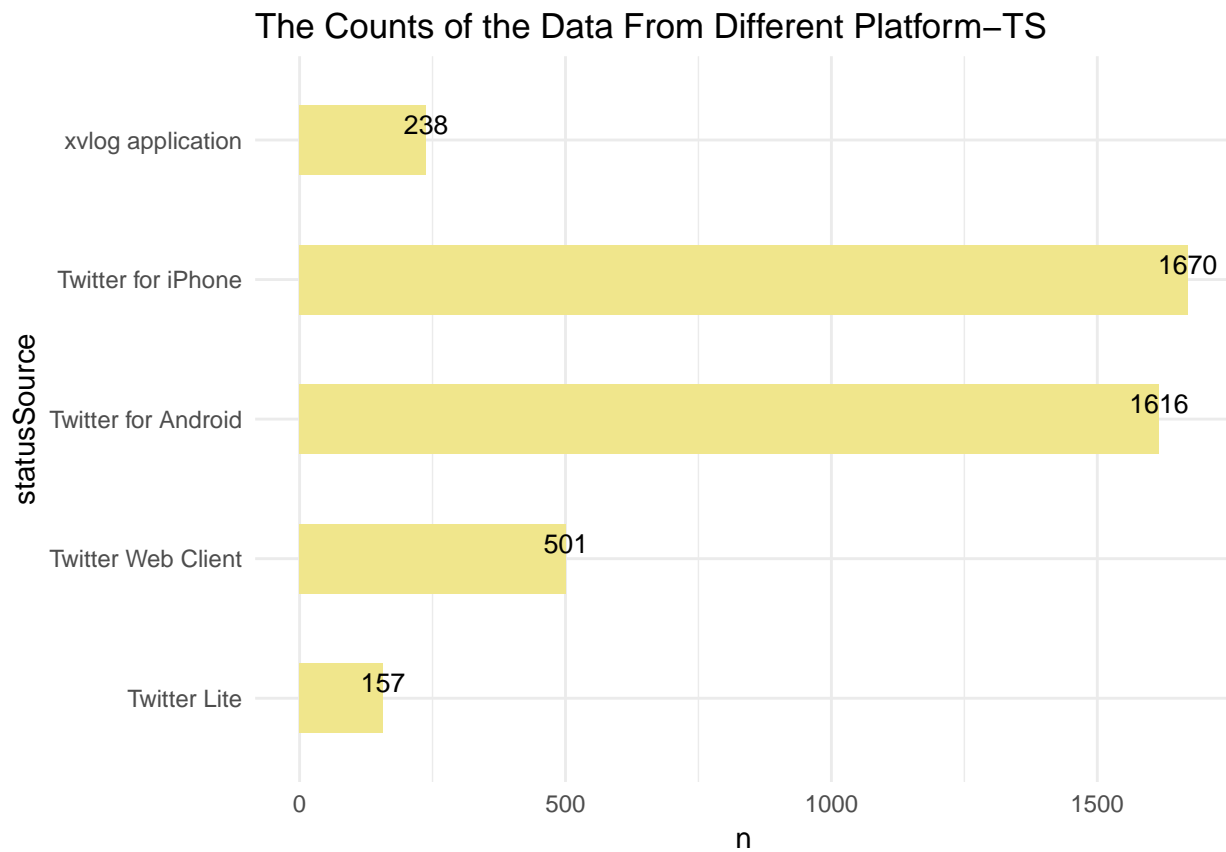
## Calculate the count of tweets by platform.

Here is the table of the count of the source. Different from #Rihanna, most of the tweets contain topic #Taylor Swift are sent from iPhone.

statusSource	n	percent_of_tweets
Twitter for iPhone	1670	0.334
Twitter for Android	1616	0.323
Twitter Web Client	501	0.100
xvlog application	238	0.048
Twitter Lite	157	0.031

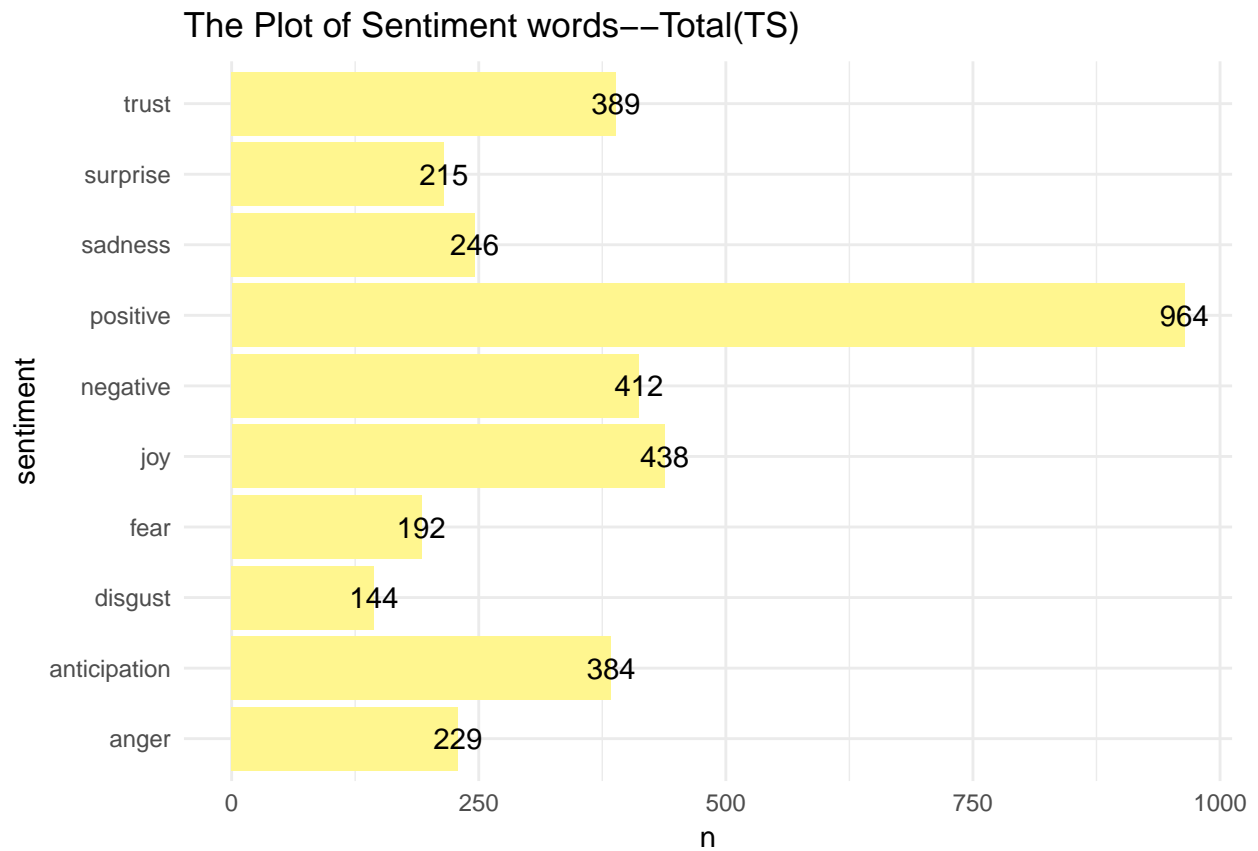
## Barplot of the count of the source

Here is the barplot of the count of the source.



## Sentiment of total

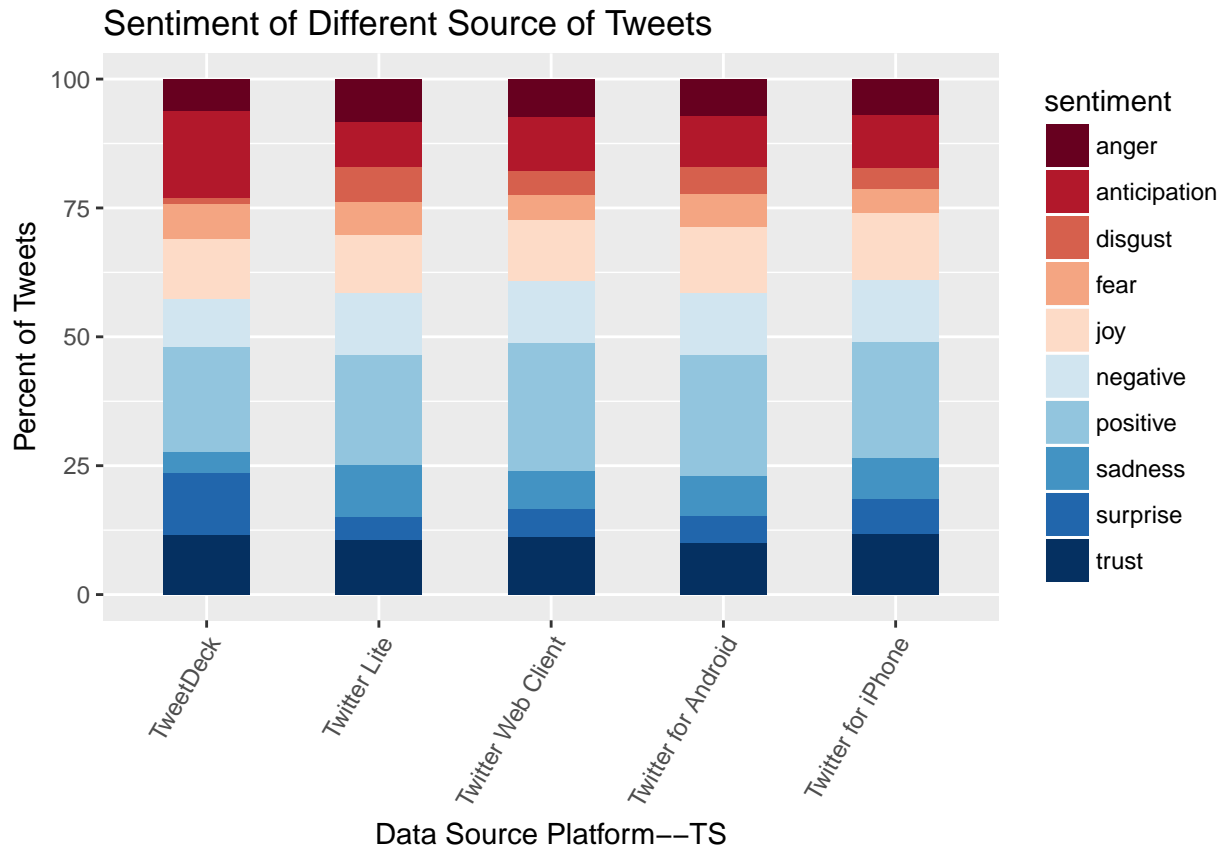
After clean the data, plot the Sentiment of Topic #Taylor Swift in total. The result below shows that positive words appear frequently in the topic of Taylor Swift.



### Sentiment of top 5 data source

Calculate and plot the Sentiment of Topic #Taylor Swift from different data source platform. The results is similar to the topic of #Rihanna.





## Wordclouds Analysis—Top 10 words in the Topic(#Taylor Swift)

The character of “Taylor Swift” appears a lot in the top 10 words.

```
str(aT5$word)
```

```
## chr [1:24497] "#demilovato" "#taylorswift" "#beyonc" "00e9" ...
```

```
aT5word <- as.character(aT5$word)
```

```
aT5text <- Corpus(VectorSource(aT5word))
```

```
aT5text <- aT5text %>%
```

```
  tm_map(removePunctuation)%>%
```

```
  tm_map(removeNumbers)%>%
```

```
  tm_map(stripWhitespace)%>%
```

```
  tm_map(tolower)%>%
```

```
  tm_map(removeWords, stopwords("english"))
```

```
aT5text <- tm_map(aT5text,stemDocument)
```

```
aT5count <- as.matrix(TermDocumentMatrix(aT5text))
```

```
aT5fre <- sort(rowSums(aT5count),decreasing = T)
```

```
aT5fd <- data.frame(word=names(aT5fre),freq=aT5fre)
```

```
ggplot(aT5fd[1:10,],aes(x=reorder(factor(word), -freq),y=freq))+geom_bar(stat="identity",aes(fill=factor(word)))
  theme(axis.text.x = element_text(angle = 60, hjust = 1))
```

