BIS581\_Sentiment Analysis\_Poojitha

Loading Libraries

#load libraries  
##Student first name: Poojitha  
library(tidytext)  
library(ggplot2)  
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(tidyr)  
library(ggraph)  
library(igraph)

##   
## Attaching package: 'igraph'

## The following object is masked from 'package:tidyr':  
##   
## crossing

## The following objects are masked from 'package:dplyr':  
##   
## as\_data\_frame, groups, union

## The following objects are masked from 'package:stats':  
##   
## decompose, spectrum

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

Load our data

#get data  
##Student last name: Velthuri  
tweetdata <- readRDS("got2.rds")

Q1. how many English tweets are there? (5)

##Student full name: Poojitha Velthuri  
  
english\_tweets <- tweetdata %>%  
 filter(lang == "en")  
  
num\_english\_tweets <- nrow(english\_tweets)  
print(paste("Number of English tweets:", num\_english\_tweets))

## [1] "Number of English tweets: 35760"

Q2. what is the ratio of retweets to total tweets?

Calculated the retweet ratio assuming that RT in msg column as retweet

##Student first name: Poojitha  
library(dplyr)  
  
tweetdata <- tweetdata %>%  
 mutate(is\_retweet = grepl("^RT", msg))  
  
tweet\_counts <- tweetdata %>%  
 summarize(  
 total\_tweets = n(),  
 retweets = sum(is\_retweet)  
 )  
  
tweet\_counts <- tweet\_counts %>%  
 mutate(retweets\_to\_totaltweets = retweets / total\_tweets)  
  
  
#print(tweet\_counts)  
  
cat(sprintf(  
 "The number of retweets in the dataset is %d, which is %.2f%% of the total tweets.\n",  
 tweet\_counts$retweets,  
 tweet\_counts$retweets\_to\_totaltweets \* 100  
))

## The number of retweets in the dataset is 40825, which is 81.65% of the total tweets.

Q3. what is the most popular “source” used to post?

##Student last name: Velthuri  
source\_cleaned <- tweetdata %>%  
 mutate(source\_clean = gsub(".\*<a href=[^>]+>([^<]+)</a>.\*", "\\1", source))  
  
source\_counts <- source\_cleaned %>%  
 count(source\_clean, sort = TRUE) %>%  
 slice\_head(n = 5)  
  
print(source\_counts)

## source\_clean n  
## 1 Twitter for iPhone 25725  
## 2 Twitter for Android 18114  
## 3 Twitter Web Client 3030  
## 4 Twitter Web App 1549  
## 5 Twitter for iPad 460

most\_popular\_source <- source\_counts %>%  
 slice\_max(n, n = 1) %>%  
 pull(source\_clean)  
  
max\_count <- source\_counts %>%  
 slice\_max(n, n = 1) %>%  
 pull(n)  
  
cat("The most popular source used is ", most\_popular\_source, "with a count of", max\_count, "\n")

## The most popular source used is Twitter for iPhone with a count of 25725

Q4. Create a bar graph of most frequent Positive and Negative words (use BING)

##Student full name: Poojitha Velthuri  
library(tidytext)  
library(ggplot2)  
library(dplyr)  
  
tweetdata\_rows <- tweetdata %>% mutate(h\_number = row\_number())  
tweetdata\_Tidy <- tweetdata\_rows %>% unnest\_tokens(word, msg)  
  
  
tweetdata\_Tidy <- tweetdata\_Tidy %>% anti\_join(stop\_words)

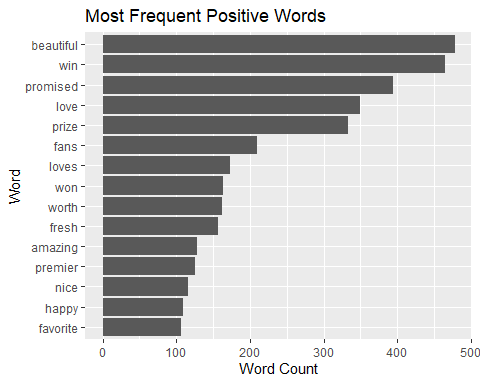
## Joining with `by = join\_by(word)`

sentiments <- get\_sentiments("bing")  
tweetdata\_sentiments <- tweetdata\_Tidy %>%  
 inner\_join(sentiments) %>%  
 count(word, sentiment, sort = TRUE)

## Joining with `by = join\_by(word)`

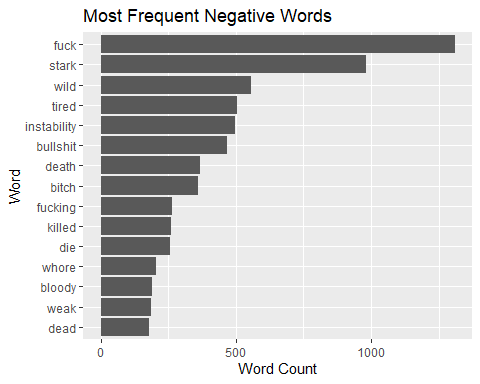
tweetdata\_positive <- tweetdata\_sentiments %>% filter(sentiment == "positive")  
tweetdata\_negative <- tweetdata\_sentiments %>% filter(sentiment == "negative")  
  
  
ggplot(tweetdata\_positive %>% top\_n(15) %>% mutate(word = reorder(word, n)),   
 aes(x = word, y = n)) +  
 geom\_bar(stat = "identity") +  
 coord\_flip() +  
 labs(x = "Word", y = "Word Count", title = "Most Frequent Positive Words")

## Selecting by n



ggplot(tweetdata\_negative %>% top\_n(15) %>% mutate(word = reorder(word, n)),   
 aes(x = word, y = n)) +  
 geom\_bar(stat = "identity") +  
 coord\_flip() +  
 labs(x = "Word", y = "Word Count", title = "Most Frequent Negative Words")

## Selecting by n

 Q5. create a wordcloud of most frequent positive and negative words

##Student last name: Velthuri  
  
library(wordcloud)

## Loading required package: RColorBrewer

library(RColorBrewer)   
  
tweetdata\_rows <- tweetdata %>% mutate(h\_number = row\_number())  
tweetdata\_Tidy <- tweetdata\_rows %>% unnest\_tokens(word, msg)  
  
  
tweetdata\_Tidy <- tweetdata\_Tidy %>% anti\_join(stop\_words)

## Joining with `by = join\_by(word)`

sentiments <- get\_sentiments("bing")  
tweetdata\_sentiments <- tweetdata\_Tidy %>%  
 inner\_join(sentiments) %>%  
 count(word, sentiment, sort = TRUE)

## Joining with `by = join\_by(word)`

tweetdata\_positive <- tweetdata\_sentiments %>% filter(sentiment == "positive")  
tweetdata\_negative <- tweetdata\_sentiments %>% filter(sentiment == "negative")  
  
  
wordcloud(  
 words = tweetdata\_positive$word,  
 freq = tweetdata\_positive$n,  
 min.freq = 3, # Adjust as needed  
 scale = c(4, 0.5),  
 random.order = FALSE,  
 colors = brewer.pal(8, "Blues"),   
 rot.per = 0.35,   
 main = "Word Cloud of Most Frequent Positive Words"  
)



wordcloud(  
 words = tweetdata\_negative$word,  
 freq = tweetdata\_negative$n,  
 min.freq = 3,  
 scale = c(4, 0.5),  
 random.order = FALSE,  
 colors = brewer.pal(8, "Reds"),   
 rot.per = 0.35,   
 main = "Word Cloud of Most Frequent Negative Words"  
)

 Q6. create a network diagram (adjust the filter so that individual words are legible. ie: don’t have a tangled mess of a plot)

##Student full name: Poojitha Velthuri  
# Load necessary libraries  
library(tidytext)  
library(dplyr)  
library(igraph)   
library(ggraph)   
library(tidyr)   
library(ggplot2)

## bigram n  
## 1 https tco 34020  
## 2 gameofthrones https 15169  
## 3 u 0001f602 5065  
## 4 u 0627 4498  
## 5 0627 u 4417

