Trump Tweet Project

Fengshou Liang

20/10/2020

As the US election approaches, we see President Donald Trump has been tweeting more often. In this project, I will be analyzing Trump’s tweet from 2016-01-01 to 2020-10-04. I will be looking at his tweeting pattern in terms of time and date, and hopefully I will come up with some interesting Trump’s tweeting pattern. This project is inspired by [David Robinson’s blog](http://varianceexplained.org/r/trump-tweets/) and [Nathan Taback’s report](http://utstat.toronto.edu/~nathan/teaching/sta4002/Class4/trumptweets-students.html). The data is collected from [here](http://www.trumptwitterarchive.com/).

# Data

The data is collected from Trump Tweet Archive created by Brendan Brown and I copied and pasted the data from 2016-01-01 to 2020-10-04 to txt file.

library(data.table)  
trump <- fread("C:/Users/sheri/Desktop/Data/trump.txt")

# Exploring the data

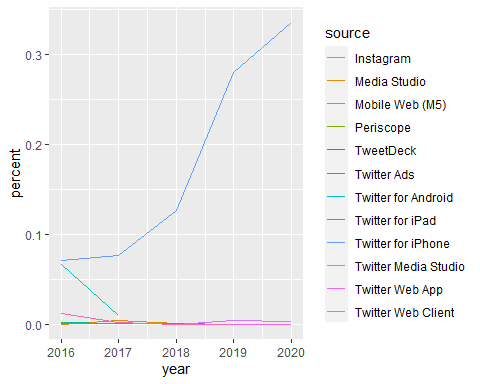
We notice that Trump used a variety of devices to access twitter. Apart from that, we find the tweet content are sometimes websites, which we are not interested in. So let’s first take a look at the devices Trump has used.

library(dplyr)  
trump %>% count(source) %>% arrange(desc(n))

## source n  
## 1: Twitter for iPhone 24469  
## 2: Twitter for Android 2115  
## 3: Twitter Web Client 396  
## 4: Twitter Media Studio 212  
## 5: Media Studio 156  
## 6: Twitter Ads 96  
## 7: Twitter for iPad 60  
## 8: Twitter Web App 7  
## 9: Instagram 2  
## 10: TweetDeck 2  
## 11: Mobile Web (M5) 1  
## 12: Periscope 1

In fact, the following graph shows a more detailed version of devices used in a time series.

library(ggplot2)  
library(lubridate)  
tweet <- trump %>%   
 mutate(created\_at =as.POSIXct(created\_at, format = "%m-%d-%Y %H:%M:%S"))  
tweet %>% count(source, year = year(created\_at)) %>%  
 mutate(percent = n / sum(n)) %>%  
 ggplot(aes(year, percent, color = source)) +  
 geom\_line()



Here we notice that while the usage of most of devices stay unchanged, that of iPhone and Android changes dramatically. The change of iphone usage may be explained by Trump using Twitter more frequently with his phone but why is there a decrease in the use of Android? Now I will set a constraint on the timeline and investigate what happened from 2016 to 2017.

# Trump’s tweeting pattern

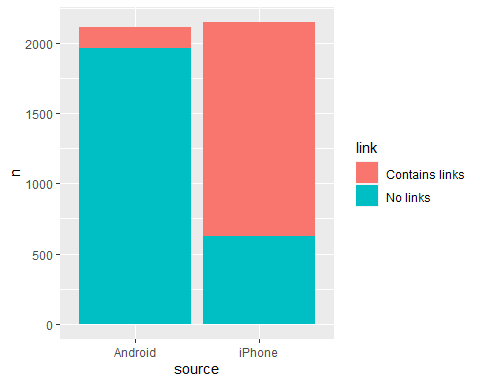
In fact, according to articles back in 2017, Trump has been reported posting slightly negative content with his Android device while positive tweets with iPhone. After such an exposure, Trump stopped using Android devices. Now let’s take a closer look and find the difference between them.

Since Trump’s last tweet with Android was on Mar.25, 2017, we are limiting his tweet to those before this date and categorize all other devices into “others” and focus on only iPhone and Android.

library(stringr)  
tweet\_new <-   
 tweet %>%   
 mutate(created\_at =as.POSIXct(created\_at, format = "%m-%d-%Y %H:%M:%S")) %>%  
 filter(created\_at <= "2017-03-26") %>%  
 mutate(source = ifelse(str\_detect(source, "iPhone"), "iPhone", ifelse(str\_detect(source, "Android"), "Android", "Others"))) %>%  
 filter(source %in% c("iPhone", "Android"))

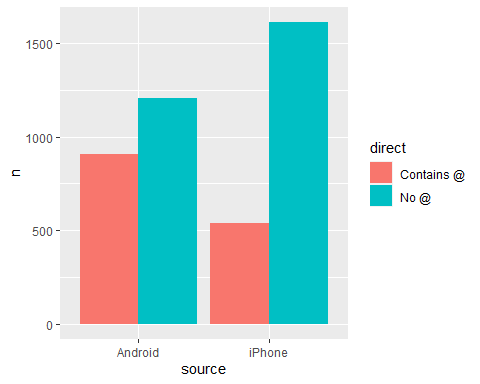
Some of Trump’s text contains an external link, so let’s take a look on how they distributed across two devices.

tweet\_link <-   
 tweet\_new %>%   
 count(source, link = ifelse(str\_detect(text, "http"), "Contains links", "No links"))  
  
ggplot(tweet\_link, aes(x=source, y=n, fill=link))+  
 geom\_bar(stat = "identity")



Similarly,the way he uses @ when referring to other people are also different but not as much as the one above.However, it appears that Trump @ people more frequently when he used Android phone.

tweet\_new %>%   
 count(source, direct = ifelse(str\_detect(text, "@"), "Contains @", "No @")) %>%  
 ggplot(aes(x=source, y=n, fill = direct)) +  
 geom\_bar(stat = "identity", position = "dodge")

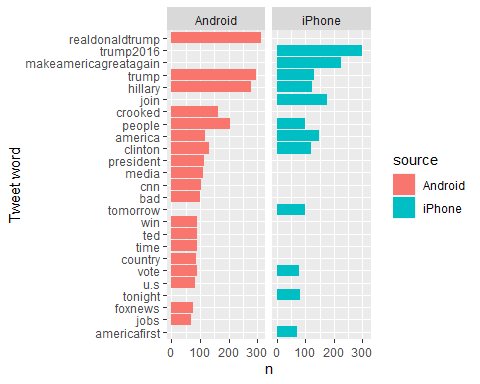


# Tweet words and sentiments

library(tidytext)  
tweet\_comparison <-   
tweet\_new %>%   
 mutate(text = str\_replace\_all(text, "https://t.co/[A-Za-z0-9]", " ")) %>%  
 unnest\_tokens(word, text) %>%   
 #With tokenization transform sentences into words  
 filter(!word %in% stop\_words$word & !word %in% c("rt","cruz", "amp", "â")) %>%  
 #some words appears to be random so they need to be removed  
 count(source, word) %>%  
 arrange(source, desc(n))

Now we have found the words Trump used most frequently in different devices. We can plot them side by side for a visual comparison.

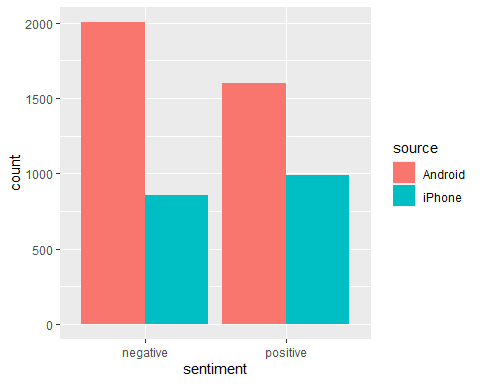
tweet\_comparison %>%   
 top\_n(30) %>%   
 ggplot(aes(x=reorder(word, n), y=n, fill=source)) +  
 geom\_bar(stat = "identity") +  
 xlab("Tweet word") +  
 coord\_flip() +  
 facet\_grid(. ~source)

 We can see Trump used his Android devices to tweet words including “clinton”, “cnn” and “bad” but not with his iPhone. Thus, I suspect that he had other agents to tweet for him on a separate device.

library(tidytext)  
#sentiments is a dataframe containing a list of words with their sentiments (positive or negative).  
head(sentiments)

## # A tibble: 6 x 2  
## word sentiment  
## <chr> <chr>   
## 1 2-faces negative   
## 2 abnormal negative   
## 3 abolish negative   
## 4 abominable negative   
## 5 abominably negative   
## 6 abominate negative

tweet\_comparison %>%  
 inner\_join(sentiments, by = "word") %>%  
 group\_by(sentiment, source) %>%  
 summarise(count = sum(n))%>%  
 ggplot(aes(sentiment, count, fill=source)) +  
 geom\_bar(stat = "identity", position = "dodge")



According to diagram above, when it comes to words associated with negative sentiments, Trump used his Android phone more often.

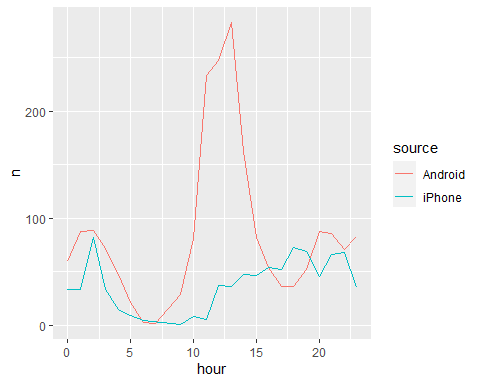
# incomplete  
#get number of Android negative word/total negative word per day   
tweet\_new %>%  
 mutate(text = str\_replace\_all(text, "https://t.co/[A-Za-z0-9]", " ")) %>%  
 unnest\_tokens(word, text) %>%   
 filter(!word %in% stop\_words$word & !word %in% c("rt","cruz", "amp", "â")) %>%  
 inner\_join(sentiments, by = "word") %>%  
 filter(sentiment == "negative") %>%  
 group\_by(date(created\_at)) %>%  
 mutate(sth = n())

## # A tibble: 2,865 x 10  
## # Groups: date(created\_at) [377]  
## source created\_at retweet\_count favorite\_count is\_retweet id\_str  
## <chr> <dttm> <int> <int> <lgl> <dbl>  
## 1 Andro~ 2017-03-25 14:37:52 22518 104321 FALSE 3.87e-252  
## 2 Andro~ 2017-03-25 14:37:52 22518 104321 FALSE 3.87e-252  
## 3 iPhone 2017-03-24 12:23:00 13364 61991 FALSE 3.67e-252  
## 4 iPhone 2017-03-24 12:14:32 12566 68241 FALSE 3.67e-252  
## 5 iPhone 2017-03-24 12:14:32 12566 68241 FALSE 3.67e-252  
## 6 iPhone 2017-03-23 15:16:24 23258 117021 FALSE 3.52e-252  
## 7 iPhone 2017-03-23 15:16:24 23258 117021 FALSE 3.52e-252  
## 8 iPhone 2017-03-23 15:16:24 23258 117021 FALSE 3.52e-252  
## 9 iPhone 2017-03-23 12:18:33 21324 94648 FALSE 3.50e-252  
## 10 iPhone 2017-03-23 12:18:33 21324 94648 FALSE 3.50e-252  
## # ... with 2,855 more rows, and 4 more variables: word <chr>, sentiment <chr>,  
## # `date(created\_at)` <date>, sth <int>

# Tweet hour and sentiments

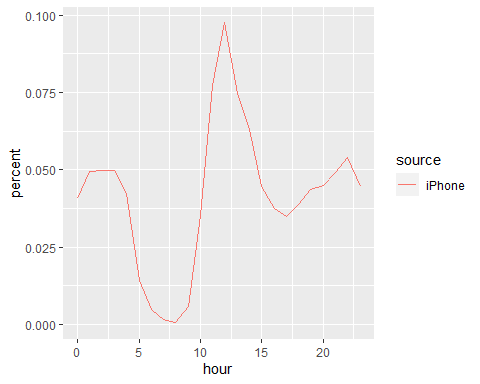
In fact, we can take a closer look at how he tweeted in during different hours. When it comes to tweeting negative words, Android phone usage outperformed that of iPhone. The peak of usage is around noon.

tweet\_new %>%  
 mutate(text = str\_replace\_all(text, "https://t.co/[A-Za-z0-9]", " ")) %>%  
 unnest\_tokens(word, text) %>%  
 filter(!word %in% stop\_words$word) %>%  
 inner\_join(sentiments, by = "word") %>%  
 filter(sentiment == "negative") %>%  
 count(source, hour = hour(created\_at)) %>%  
 ggplot(aes(hour, n, color = source)) +  
 geom\_line()

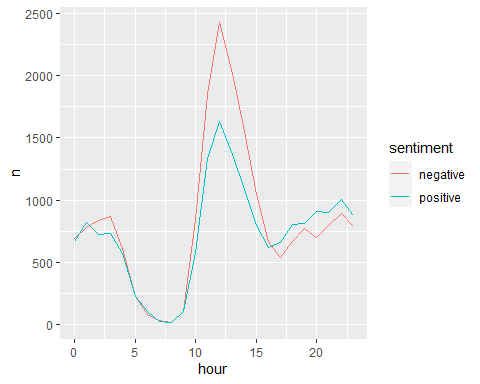


If we disregard the time limit adn look at how Trump used his iPhone since 2016 to present, around 10% of his negative words come from 11-13 o’clock.

trump %>%  
 mutate(created\_at =as.POSIXct(created\_at, format = "%m-%d-%Y %H:%M:%S")) %>%  
 mutate(source = ifelse(str\_detect(source, "iPhone"), "iPhone", "Others")) %>%  
 filter(source == "iPhone") %>%  
 count(source, hour = hour(created\_at)) %>%  
 mutate(percent = n / sum(n)) %>%  
 ggplot(aes(hour, percent, color = source)) +  
 geom\_line()

 I wanted to explore whether Trump experienced similar emotions at fixed period of time, but the result below shows that during lunch time, he tweets both positive and negative words and there is no apparent difference between the two.

trump %>%  
 mutate(created\_at =as.POSIXct(created\_at, format = "%m-%d-%Y %H:%M:%S")) %>%  
 mutate(source = ifelse(str\_detect(source, "iPhone"), "iPhone", "Others")) %>%  
 filter(source == "iPhone") %>%  
 mutate(text = str\_replace\_all(text, "https://t.co/[A-Za-z0-9]", " ")) %>%  
 unnest\_tokens(word, text) %>%  
 filter(!word %in% stop\_words$word) %>%  
 inner\_join(sentiments, by = "word") %>%  
 count(sentiment, hour = hour(created\_at)) %>%  
 ggplot(aes(hour, n, color = sentiment)) +  
 geom\_line()



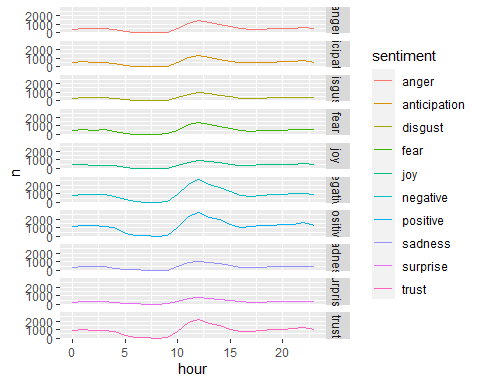
So I took a more detailed look on emotions but still the data doesn’t verify my hypothesis.

library(textdata)

## Warning: package 'textdata' was built under R version 4.0.3

sentiment\_detail <- get\_sentiments("nrc")  
trump %>%  
 mutate(created\_at =as.POSIXct(created\_at, format = "%m-%d-%Y %H:%M:%S")) %>%  
 mutate(source = ifelse(str\_detect(source, "iPhone"), "iPhone", "Others")) %>%  
 filter(source == "iPhone") %>%  
 mutate(text = str\_replace\_all(text, "https://t.co/[A-Za-z0-9]", " ")) %>%  
 unnest\_tokens(word, text) %>%  
 filter(!word %in% stop\_words$word) %>%  
 inner\_join(sentiment\_detail, by = "word") %>%  
 count(sentiment, hour = hour(created\_at)) %>%  
 ggplot(aes(hour, n, color = sentiment)) +  
 geom\_line() +  
 facet\_grid(sentiment~.)

## Warning: Removed 5 row(s) containing missing values (geom\_path).



However, I found an interesting pattern regarding Trump’s sleep schedule. Assume he is the only one with his Twitter account on his iPhone, Trump sleeps very late and gets up early and it seems that he only sleeps for around 5 hours.