nlr30

Sentiment analysis in Tweets

The observations are the set of tweets by then US president to-be in 2016 election year. The tweets were sent from different electric devices, and the some devices show clear tendency at what time of a day the tweets were sent. We will investigate who is using which device, and if there are tendency of sentiment that may highlight the sender's state of mind.

This analysis is part of practice performed during the online lecture in Harvard X for Data Science in 2020. This is a recap of

https://courses.edx.org/courses/course-v1:HarvardX+PH125.6x+1T2020/courseware/82aee45f9f0b4511a7e86bde6b151d08/8f0e14d73cae4e6ea291d7fc66dea2aa/?child=first.

The data were taken from the following site.

http://www.trumptwitterarchive.com.

```
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------------------------
```

```
## v ggplot2 3.3.0     v purrr   0.3.4
## v tibble  3.0.1     v dplyr   0.8.5
## v tidyr   1.1.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0
```

```
## -- Conflicts ------------------------------------------------------------------------------------
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ggplot2)
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

```
## The following objects are masked from 'package:dplyr':
##
##     intersect, setdiff, union
```

```
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
library(dplyr)
library(scales)
```

```
##
## Attaching package: 'scales'
```

```
## The following object is masked from 'package:purrr':
##
##     discard
```

```
## The following object is masked from 'package:readr':
##
##     col_factor
```

```
library(tidytext)
library(rmarkdown)
```

Retrieve the raw data from <trumptwitterarchive.com>.

```
url <- 'http://www.trumptwitterarchive.com/data/realdonaldtrump/%s.json'
```

Use data from 2015 to 2016.

```
tw1 <- map(2015:2016, ~sprintf(url, .x))  %>% map_df(jsonlite::fromJSON, simplifyDataFrame = T)
```

Remove retweets.

```
tw2 <-tw1 %>%
  filter(!is_retweet & !str_detect(text, '^"')) %>%
  mutate(created_at = parse_date_time(created_at, orders="a b! d! H!:M!:S! z!* Y!", tz="EST")) %>%
  select(source, id_str, text, created_at)
```

```
## Date in ISO8601 format; converting timezone from UTC to "EST".
```

Find tweets that are sent only from Android or iPhone. Other devices/platform are ignored.

```
tw2 <- tw2 %>% arrange(created_at) %>%
  extract(source, 'source', 'Twitter for (.*)')
```

Include these tweets only that are posted during the election campaign.

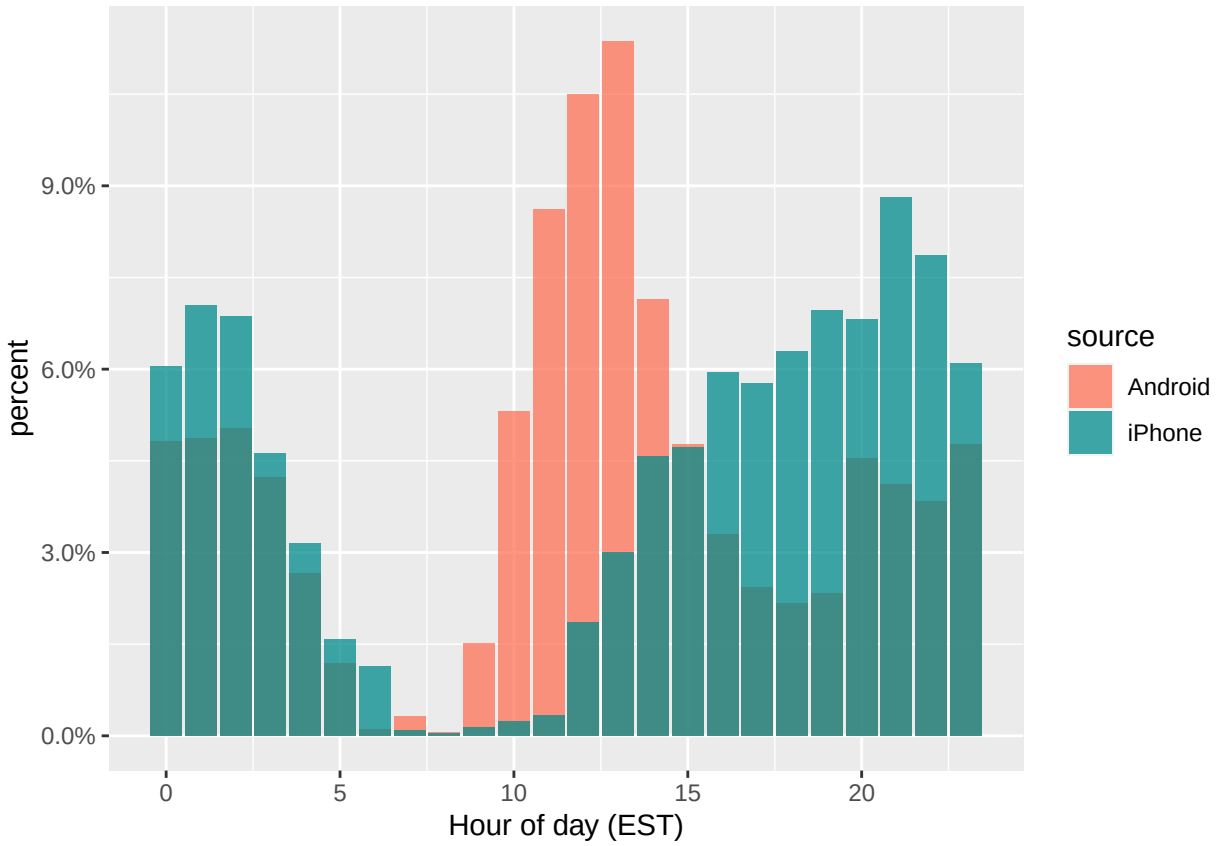
```
tw2 <- tw2 %>%  filter(source %in% c('Android', 'iPhone') & created_at >= ymd('2015-06-17') & created_a
```

Group the data by hours of a day to see if there is any trend in which devices for posting at each hour. Calculate how much fractions of whole tweets are posted that hour of a day.

```
tw3 <- tw2 %>%
  mutate(hour = hour(with_tz(created_at,'EST'))) %>%
  count(source, hour)  %>%
  group_by(source)  %>%
  mutate(percent=n/sum(n)) # %>% ungroup() %>%
```

Plot the hourly trend.

```
tw3 %>% ggplot(aes(x=hour, y=percent, fill=source) )+
      geom_bar(aes(fill=source), stat="identity",position="identity") +
      scale_fill_manual(values = alpha(c("coral1","cyan4"), 0.75)) +
      xlab('Hour of day (EST)')+
      scale_y_continuous(labels=percent_format())
```



There is an obvious peak between 9 am and 5 pm posted from Android. Apparently there are two teams –

1. who use Androlid during morning
2. who use iPhone in the afternoon

1) is likely Trump himself, and 2) is staff. Now we will investigate if there is any difference in sentiment in the tweets posted by two teams above.

Remove twitter web site URL

```
tw4 <- tw2 %>% mutate(text=str_replace_all(text, 'http://t.co/[A-Za-z\\d]+|&amp;', '')) %>%
  unnest_tokens(word, text, token='tweets')  %>%
  filter(!word %in% stop_words$word & !str_detect(word, '^\\d+$')) %>%
  mutate(word = str_replace(word, "^'", ""))
```

```
## Using `to_lower = TRUE` with `token = 'tweets'` may not preserve URLs.
```

These are kind of words that most frequenly appear in tweets

```
tw4 %>% count(word) %>% arrange(desc(n)) %>% head
```

```
## # A tibble: 6 x 2
##   word                      n
##   <chr>                 <int>
## 1 #trump2016              414
## 2 hillary                 405
## 3 people                  302
## 4 #makeamericagreatagain  294
## 5 america                 254
```

```
## 6 clinton                   237
```

Now group by device, Android or iPhone remove those words where the samples are less than 32.

```
aoi <- tw4 %>% count(word, source) %>%
  spread(source, n, fill=0)  %>%
  filter(Android + iPhone >= 32) %>% arrange(desc(Android))
head(aoi)
```

```
## # A tibble: 6 x 3
##   word    Android iPhone
##   <chr>     <dbl>  <dbl>
## 1 hillary     289    116
## 2 people      194    108
## 3 crooked     156     49
## 4 clinton     136    101
## 5 poll        116    101
## 6 america     114    140
```

Selection of dictionaries that sentiment labels are set. We will use nrc sentiment.

```
nrc <- get_sentiments('nrc') # %>% select(word, sentiment)
afi <- get_sentiments('afinn') # %>% select(word, sentiment)
bing <- get_sentiments("bing")
loug <- get_sentiments("loughran") %>% count(sentiment)
```

Assign the sentiments.

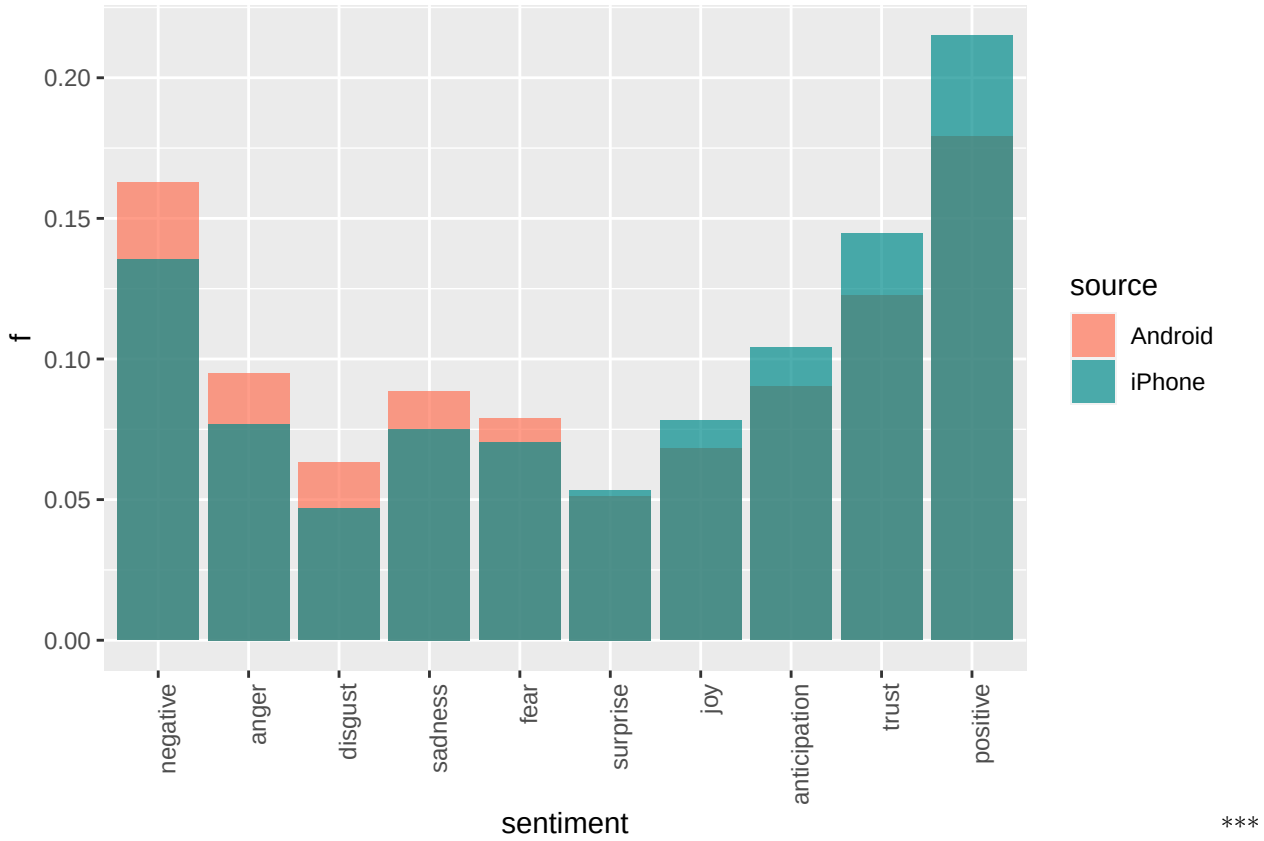
```
snt1 <- tw4 %>% left_join(nrc, by='word') %>%
  count(source, sentiment)%>%
  spread(source, n) %>%
  filter(!is.na(sentiment)) %>%
  mutate(Android = Android / sum(Android) ) %>%
  mutate(iPhone = iPhone / sum(iPhone) )  %>%
  mutate(sentiment = fct_reorder(sentiment, desc(Android - iPhone)))

snt1
```

```
## # A tibble: 10 x 3
##    sentiment    Android iPhone
##    <fct>          <dbl>  <dbl>
##  1 anger         0.0950 0.0769
##  2 anticipation  0.0902 0.104
##  3 disgust       0.0633 0.0469
##  4 fear          0.0788 0.0703
##  5 joy           0.0682 0.0781
##  6 negative      0.163  0.135
##  7 positive      0.179  0.215
##  8 sadness       0.0887 0.0752
##  9 surprise      0.0514 0.0534
## 10 trust         0.123  0.145
```

Plot sentiment in the order Android is more prone to

```
snt1 %>% mutate(sentiment = fct_reorder(sentiment, desc(Android - iPhone))) %>%
    gather("source", "f", 2:3) %>%
    ggplot(aes(sentiment, f, fill=source)) +
    geom_bar(aes(fill=source), stat='identity', position='identity') +
    scale_fill_manual(values=alpha(c('coral1', 'cyan4'), .7)) +
    theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



***

Conclusion: Sentiment that were sent from Android during morning hours are more negative than those from iPhone in the afternoon. The sentiments that are more often seen in the former are “negative”, “anger”, “disgust”, “sadness”, and “fear”.