Explore Twitter Data

Lyn Nguyen

2022-12-07

Contents

ELT

This script uses output from analysis-of-public-opinion/scraper.py. Ultimately, we keep data pulled on Dec

```r
# created_at to date and day of week
test = head(tweets1)
dow <- substr(test$created_at, 1, 3)
month_day <- substr(test$created_at, 5, 10)
time<- substr(test$created_at, 12, 19)
yr <- substr(test$created_at, 26, 30)

ymd <- as.Date(paste0(month_day, yr), format = "%b %d %h:%m:%s %Y")

# as.Date(test$created_at, format = "%a %b %d %h:%m:%s +0000 %Y")
tweets1 <- tweets1 %>% mutate(dow = substr(created_at, 1, 3)
                              , month_day = substr(created_at, 5, 10)
                              , time = substr(created_at, 12, 19)
                              , yr = substr(created_at, 26, 30),
                              , ymd = as.Date(paste0(month_day, yr), format = "%b %d %Y"))
tweets2 <- tweets2 %>% mutate(dow = substr(created_at, 1, 3)
                              , month_day = substr(created_at, 5, 10)
                              , time = substr(created_at, 12, 19)
                              , yr = substr(created_at, 26, 30),
                              , ymd = as.Date(paste0(month_day, yr), format = "%b %d %Y"))
tweets3 <- tweets3 %>% mutate(dow = substr(created_at, 1, 3)
                              , month_day = substr(created_at, 5, 10)
                              , time = substr(created_at, 12, 19)
                              , yr = substr(created_at, 26, 30),
                              , ymd = as.Date(paste0(month_day, yr), format = "%b %d %Y")
                              , tweet_id_char = as.character(as.numeric(tweet_id)))
tweets4 <- tweets4 %>% mutate(dow = substr(created_at, 1, 3)
                              , month_day = substr(created_at, 5, 10)
                              , time = substr(created_at, 12, 19)
                              , yr = substr(created_at, 26, 30),
                              , ymd = as.Date(paste0(month_day, yr), format = "%b %d %Y")
                              , tweet_id_char = as.character(as.numeric(tweet_id)))
summary(tweets1$ymd)
```

```
##         Min.      1st Qu.       Median         Mean      3rd Qu.         Max.
## "2022-11-26" "2022-12-01" "2022-12-01" "2022-12-01" "2022-12-03" "2022-12-03"
```

```r
summary(tweets2$ymd)
```

```
##         Min.      1st Qu.       Median         Mean      3rd Qu.         Max.
## "2022-11-26" "2022-12-01" "2022-12-01" "2022-12-01" "2022-12-03" "2022-12-03"
```

```r
summary(tweets3$ymd)
```

```
##         Min.      1st Qu.       Median         Mean      3rd Qu.         Max.
## "2022-11-27" "2022-12-01" "2022-12-02" "2022-12-02" "2022-12-03" "2022-12-05"
```

```r
summary(tweets4$ymd)
```

```
##         Min.      1st Qu.       Median         Mean      3rd Qu.         Max.
## "2022-11-28" "2022-12-01" "2022-12-01" "2022-12-01" "2022-12-03" "2022-12-03"
```

`tweets1.csv` has data from 11/26/2022 but only cnn as liberal source. `tweet2.csv`: 11/26- 12/3 but only cnn as liberal source `tweet3.csv`: 11/28- 12/3 liberal sources has cnn, npr, msnbc, nytimes, `tweet4.csv`: 11/28- 12/3 but only cnn as liberal source

`tweets1` and `tweets2` have 814 fields total, but only 468 unique.

```
master <- rbind(tweets3, tweets4) %>% select(-experiment_id) %>% distinct()
```

`master` has 471 points, but `length(unique(master$tweet_id))` has 468 points. Where is the 3 difference? Since tweet4 hit the api after tweet3, some has updated values. For example tweet_id "1598304394931412992" has 0 like in tweet3 but 1 like in tweet 4. If there is duplicate in tweet_id, we will keep the one with the higher index.

```
master <- master %>% mutate(tweet_id_char = as.character(as.numeric(tweet_id)))
master_tweet_id <- master$tweet_id_char
dup_master <- master_tweet_id[duplicated(master_tweet_id) == T]

print("The duplicated tweet_ids are:")
```

```
## [1] "The duplicated tweet_ids are:"
```

```
dup_master
```

```
## [1] "1598304394931412992" "1598277959223083008" "1598411537667874816"
```

3 tweets are duplicated because they have updated "likes" count.

```
dup_val1 <- master[master$tweet_id == 1598304394931412992, ][2,]
dup_val2 <- master[master$tweet_id == 1598277959223083008, ][2,]
dup_val3 <- master[master$tweet_id == 1598411537667874816, ][2,]

m <- master %>% filter(!tweet_id %in% dup_master)
master <- rbind(m, dup_val1, dup_val2, dup_val3) %>% arrange(tweet_id) # in ascending tweet_id order

# write.csv(master, "prelim_data/tweets_master_dec5dec6.csv")
```

Get `text` and `tweet_id` only.

Madelaine will use this file in SageMaker. Need to keep row orders for annotation output.

```
tweet_text <- master %>% select("tweet_id", "text") %>% distinct() #468
# write.csv(tweet_text, "prelim_data/tweet_text_only.csv")
```

Clean up `users`.

```
# What user_id in user3 that's not in user4?
user4_id <- users4$user_id
user3_id <- users3$user_id
```

```
user3_id_only <- setdiff(user3_id, user4_id) #ids in user3 that's not in user4 - 118 total
user3_profiles_only <- users3 %>% filter(user_id %in% user3_id_only)

all_users <- rbind(user3_profiles_only, users4) %>% mutate(user_id_char = as.character(as.numeric(user_

# merge users and tweets ===
# rename overlap column names
tweet_cnames <- colnames(master)
colnames(master) <- c("experiment_group", "text", "tweet_id", "tweet_likes", "retweets","tweet_created_
                      "in_reply_to_screen_name", "screen_name","dow","month_day","time","yr","ymd","twe
author_cnames <- colnames(all_users)

final_tweets <- left_join(master %>% select(-c(screen_name)), all_users, by = "user_id")

write.csv(final_tweets, "data/master.csv")
```

There are 459 unique authors for these 468 tweets.

EDA

`final_tweets` have 25 columns and 468 observations (tweets).

```
glimpse(final_tweets)
```

```
## Rows: 468
## Columns: 25
## $ experiment_group        <chr> "msnbc", "msnbc", "msnbc", "msnbc", "msnbc", "~
## $ text                    <chr> "@MSNBC @MaddowBlog "Simpleton's defense"?  Yo~
## $ tweet_id                <dbl> 1.596988e+18, 1.596993e+18, 1.596997e+18, 1.59~
## $ tweet_likes             <int> 4, 0, 0, 2, 1, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0~
## $ retweets                <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0~
## $ tweet_created_at        <chr> "Sun Nov 27 22:01:59 +0000 2022", "Sun Nov 27 ~
## $ user_id                 <dbl> 1.518750e+18, 3.202809e+09, 1.409157e+08, 1.93~
## $ in_reply_to_status_id   <dbl> 1.596987e+18, 1.596987e+18, 1.596987e+18, 1.59~
## $ in_reply_to_user_id     <int> 2836421, 2836421, 2836421, 2836421, 2836421, 2~
## $ in_reply_to_screen_name <chr> "MSNBC", "MSNBC", "MSNBC", "MSNBC", "MSNBC", "~
## $ dow                     <chr> "Sun", "Sun", "Sun", "Sun", "Mon", "Mon", "Mon~
## $ month_day               <chr> "Nov 27", "Nov 27", "Nov 27", "Nov 27", "Nov 2~
## $ time                    <chr> "22:01:59", "22:22:27", "22:39:00", "23:13:38"~
## $ yr                      <chr> " 2022", " 2022", " 2022", " 2022", " 2022", "~
## $ ymd                     <date> 2022-11-27, 2022-11-27, 2022-11-27, 2022-11-2~
## $ tweet_id_char           <chr> "1596987727953924096", "1596992880002084864", ~
## $ created_at              <chr> "Tue Apr 26 00:33:21 +0000 2022", "Sat Apr 25 ~
## $ description             <chr> "No name", "People following me are president ~
## $ location                <chr> "", "Massachusetts, USA", "Washington, DC", "w~
## $ followers_count         <int> 8, 874, 375, 537, 5, 130, 28, 200, 15, 18, 91,~
## $ screen_name             <chr> "BigTex1022", "michael_favreau", "AlxHamiltn",~
## $ statuses_count          <int> 2333, 30060, 33016, 60763, 1102, 1636, 1637, 1~
## $ favourites_count        <int> 1941, 16373, 1061, 19861, 320, 586, 1414, 5190~
## $ verified                <chr> "False", "False", "False", "False", "False", "~
## $ user_id_char            <chr> "1518749825092788224", "3202808548", "14091571~
```

experiment_group / in_reply_to_screen_name

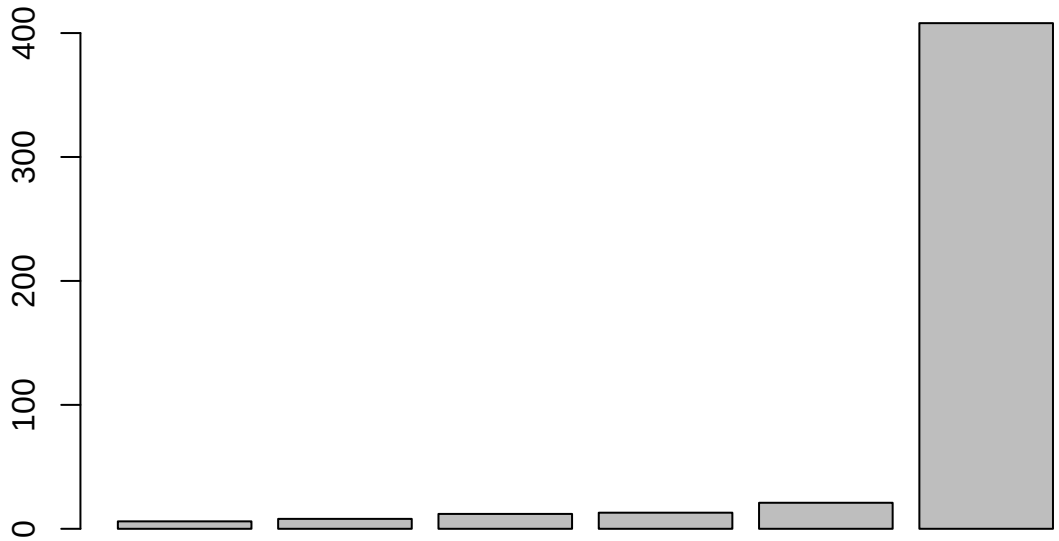
What is the share of replies to the 5 news sources? How do ('msnbc', 'cnn', 'npr', 'nytimes') compare to 'cnn'? - FoxNews make up 87% of our data points. When it comes to the student loan forgiveness discussion, the Department of Education has the least engagement from Twitter users, at only 1%.

```
liberal <- c('msnbc', 'cnn', 'npr', 'nytimes')
conservative <- c('foxnews')

source_count <- as.data.frame(table(final_tweets$in_reply_to_screen_name)) %>% mutate(Proportion = roun
source_count
```

```
##      Var1 Freq Proportion
## 1 usedgov    6       0.01
## 2 nytimes    8       0.02
## 3     CNN   12       0.03
## 4     NPR   13       0.03
## 5   MSNBC   21       0.04
## 6 FoxNews  408       0.87
```

```
barplot(source_count$Freq)
```



Tweet

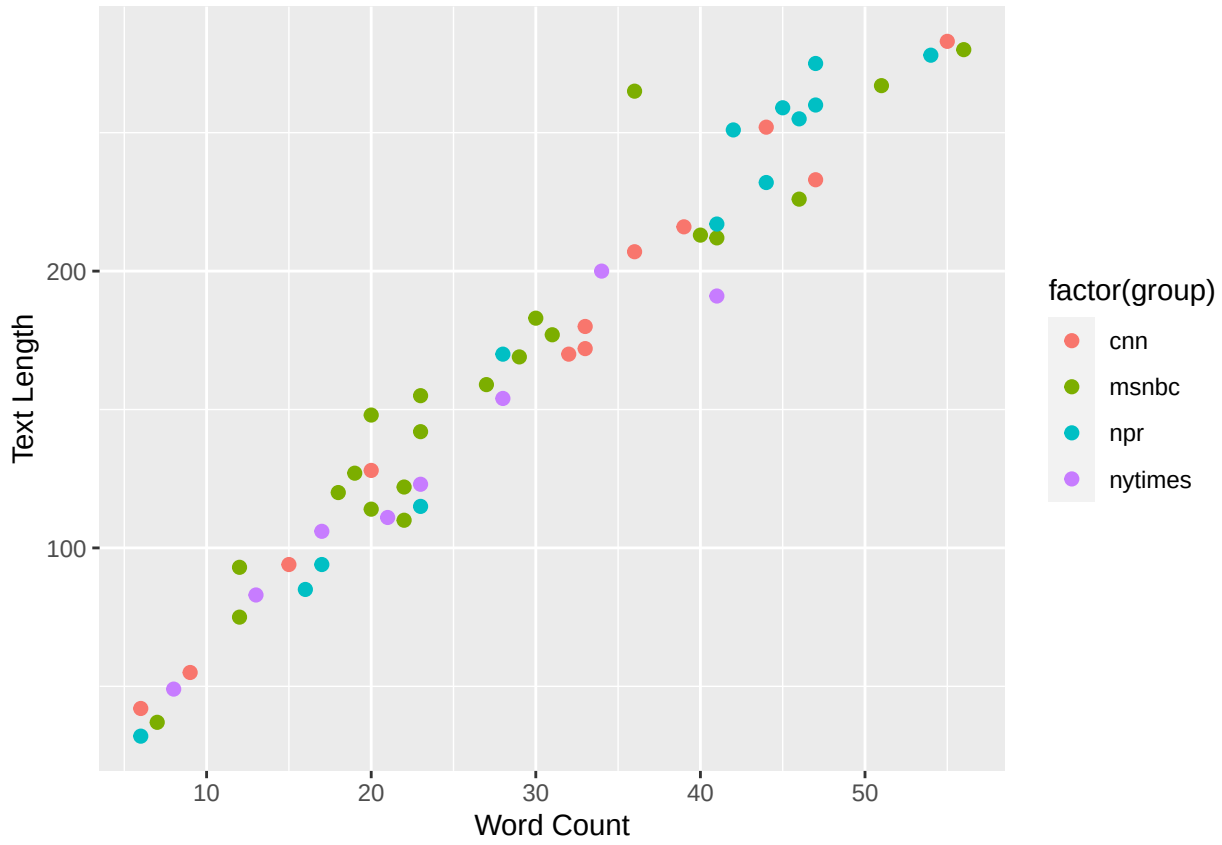
text

Is tweet length a distinguishable characteristic for the experiment groups? Within the liberal groups, most of NPR replies have over 40 words. NYTimes's reply lengths are scattered on the lower end.

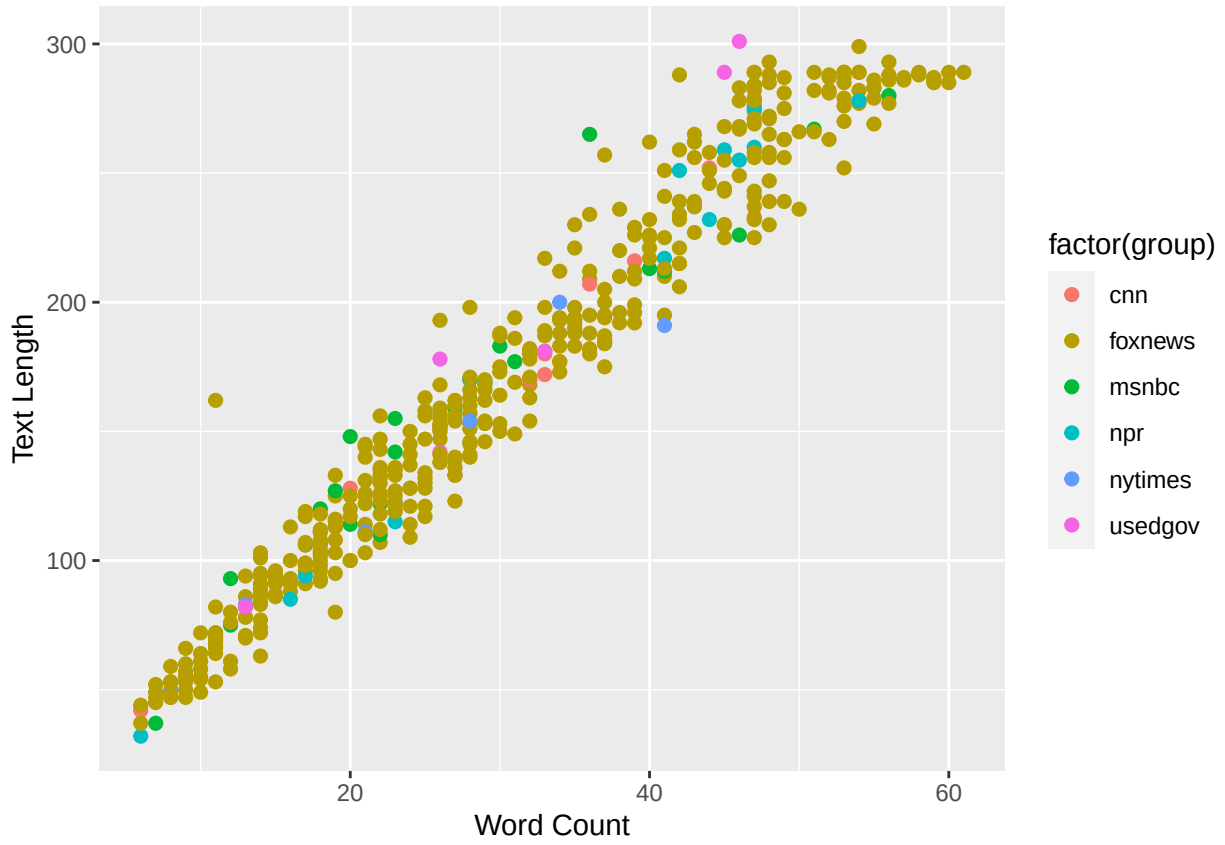
```
final_tweets <- final_tweets %>% mutate(text_length = nchar(text),
                                        text_word_count = str_count(text, '\\w+'))


l <- final_tweets %>% filter(experiment_group %in% liberal) %>%
  select(experiment_group, text_length, text_word_count)

colors <- c("#FDAE61", # Orange
            "#D9EF8B", # Light green
            "#66BD63") # Darker green
x <- l$text_word_count
y <- l$text_length
group <- l$experiment_group
# Scatter plot
ggplot(l, aes(x, y, color = factor(group))) + geom_point(size = 2) + xlab("Word Count") + ylab("Text Le
```




```
x <- final_tweets$text_word_count
y <- final_tweets$text_length
group <- final_tweets$experiment_group
ggplot(final_tweets, aes(x, y, color = factor(group))) + geom_point(size = 2) + xlab("Word Count") + yla
```




```
# how does nchar treat emojis? - no.
test = final_tweets[463,] %>% select("text", "text_word_count", "text_length")
```

Tweet Popularity

favourites_count

followers_count

verified

User

screen_name

1. Which author has multiple replies? Do they reply to the same source or not?

- 8 people replied twice, 2 of which to multiple news source twitters, but only 1 engage with conservative (FoxNews) and liberal (MSNBC).

```
author_multtweet <- c(data.frame(table(final_tweets$screen_name)) %>% filter(Freq > 1) %>% select(Var1))

author_overlap <- final_tweets %>% filter(screen_name %in% c("DahlmanCarl", "fabulosi_t", "jackSpa81774

author_overlap
```

```
##    in_reply_to_screen_name     screen_name statuses_count favourites_count
## 1                    MSNBC michael_favreau          30060            16373
## 2                    MSNBC RogerWPetersen1           1636              586
## 3                  FoxNews     thomaslew13           6530                0
## 4                  nytimes jackSpa81774793            243                5
## 5                  FoxNews     thomaslew13           6530                0
## 6                    MSNBC michael_favreau          30060            16373
## 7                  FoxNews     DahlmanCarl           2626             1336
## 8                  FoxNews     DahlmanCarl           2626             1336
## 9                      CNN jackSpa81774793            243                5
## 10                 FoxNews    PCopposition             59                0
## 11                 FoxNews    PCopposition             59                0
## 12                 usedgov      fabulosi_t           5125             5527
## 13                 usedgov      fabulosi_t           5125             5527
## 14                 FoxNews RogerWPetersen1           1636              586
## 15                 FoxNews   johnbutler410           1637              166
## 16                 FoxNews   johnbutler410           1637              166
##    followers_count tweet_likes retweets
## 1              874           0        0
## 2              130           0        0
## 3               12           0        0
## 4                2           0        0
## 5               12           0        0
## 6              874           0        0
## 7                2           1        0
## 8                2           0        0
## 9                2           0        0
## 10               0           1        0
## 11               0           1        0
## 12              72           1        0
## 13              72           0        0
## 14             130           1        0
## 15             184           7        0
## 16             184           0        0
```

created_at

Does age of account tell who they might engage with?

description

How many have profile descriptions?

location

How many have location display? Is one location more densed?

`ymd` & `dow`

Which day of the week do people discuss student loan forgiveness the most often?

`time`

What time of day has the most discussion?

User Popularity

`retweets` & `experiment_group` *Which tweet has more `retweets`? Does it happen more often on liberal or conservative outlet?*

`tweet_likes` & `experiment_group` *Which tweet has more `likes`? Does it happen more often on liberal or conservative outlet?*