Explore Twitter Data

Lyn Nguyen

2022-12-07

Contents

This script uses data from 3 annotators.

ELT

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

```
# 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"
```

```
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"
```

```
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.

```r
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.

```r
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:"
```

```r
dup_master
```

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

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

```r
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.

```r
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

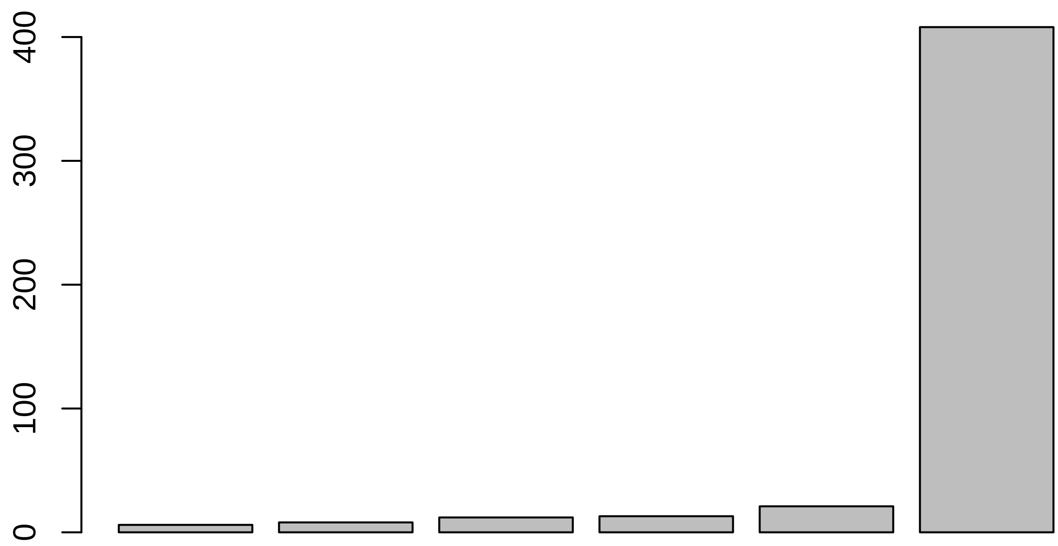
What is the share of replies to the 5 news sources? How do ('msnbc', 'cnn', 'npr', 'nytimes') compare to 'cnn'? - FoxNews make up 87% of our data points. When it comes to the student loan forgiveness discussion, the Department of Education has the least engagement from Twitter users, at only 1%.

```
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

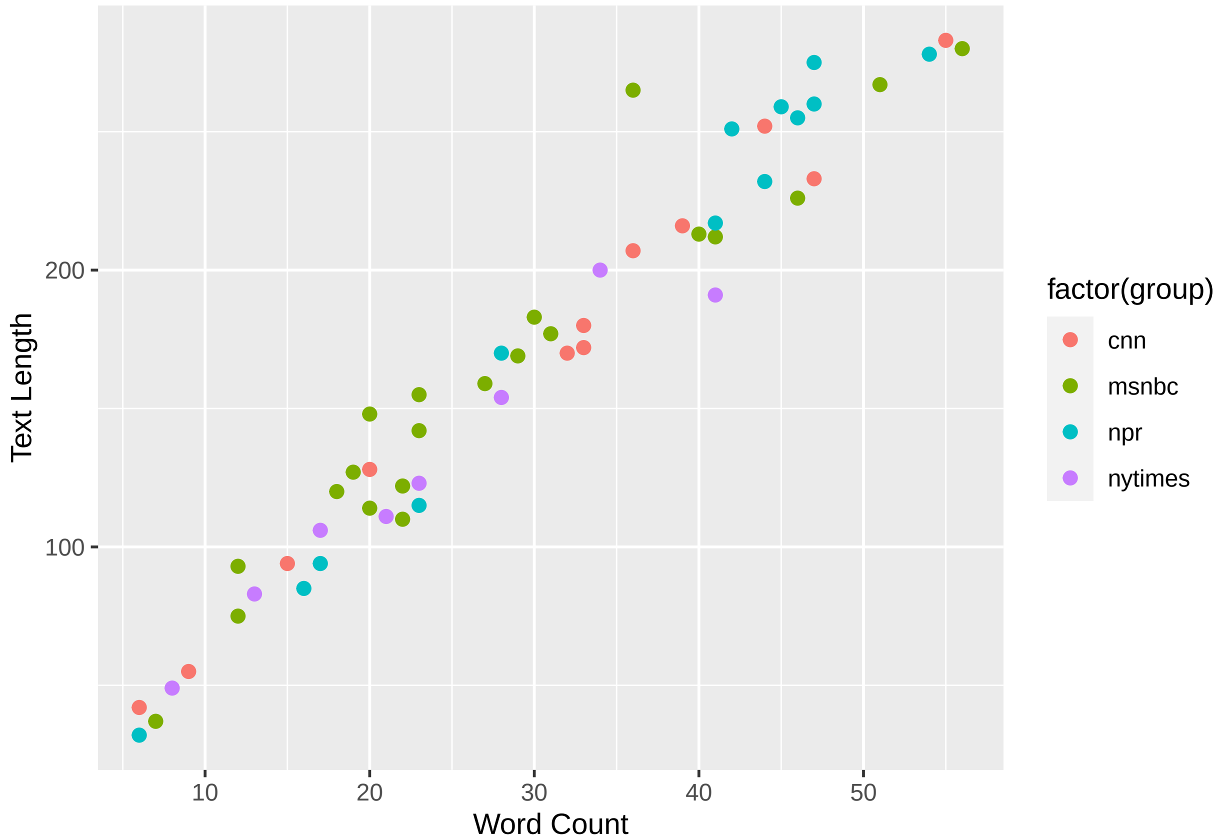
`text`

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.

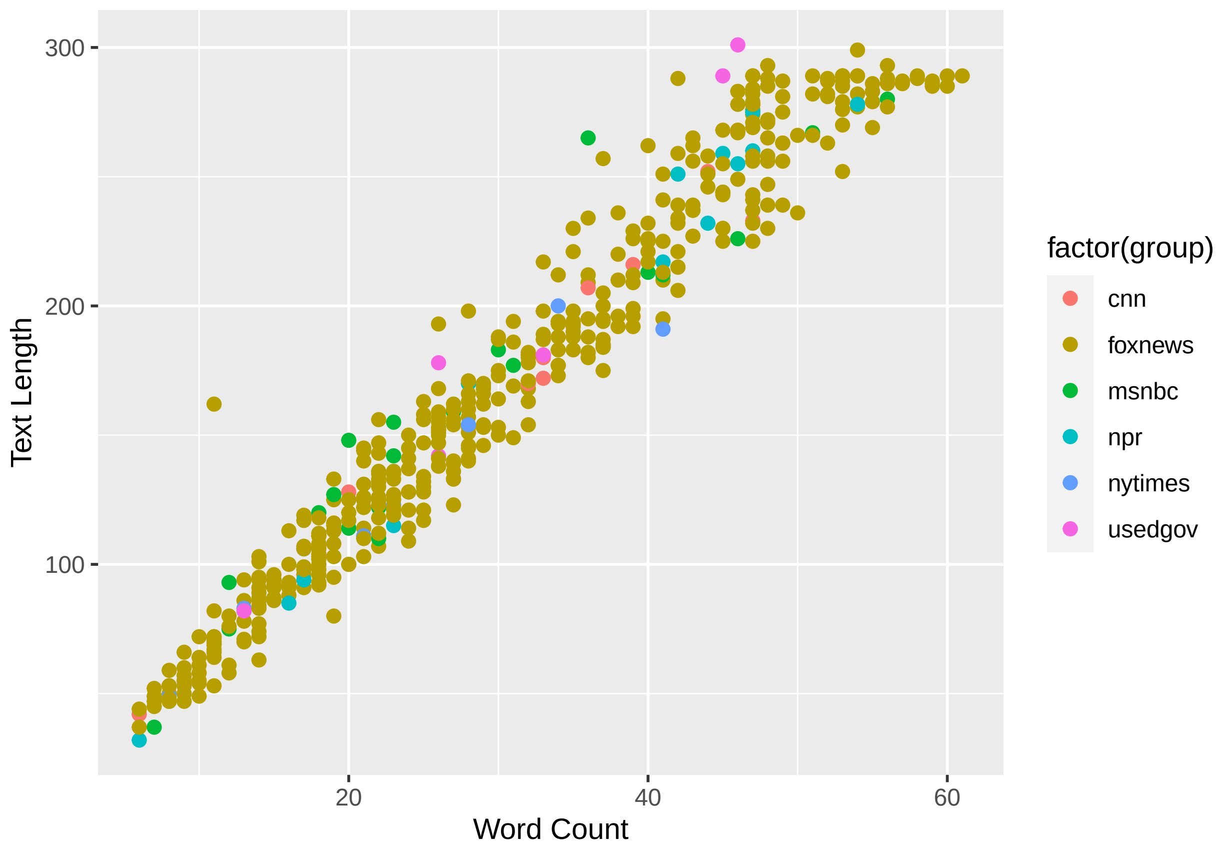
```
final_tweets <- final_tweets %>% mutate(text_length = nchar(text),
                                      text_word_count = str_count(text, '\\w+'))


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 Len
```




```
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

`retweets` & `experiment_group` *Which tweet has more* ***retweets****? Does it happen more often on liberal or conservative outlet?* One tweet has 85 retweets, one have 5 retweets, three have 3 retweets but the majority 447 (96%) do not have any retweets.

```
retweet_count <- data.frame(table(final_tweets$retweets)) %>% arrange(desc(Freq))
colnames(retweet_count) <- c("retweets", "freq")
retweet_count
```

```
##   retweets freq
## 1        0  447
## 2        1   16
## 3        2    3
## 4        5    1
## 5       85    1
```

Which outlet has posts with more than 1 retweet? Foxnews and NPR are the two sources where replies have over 1 retweet, with Foxnews holding the highest, 85 retweets.

```
many_retweets <- final_tweets %>% filter(retweets > 1) %>% select(experiment_group, retweets, screen_nam
many_retweets
```

```
##   experiment_group retweets     screen_name
## 1          foxnews       85   RastelliSteve
## 2          foxnews        2      bamakeyman
## 3              npr        5      mkurzawasc
## 4              npr        2      Dollerhide
## 5          foxnews        2 VT_Jeff_RE_Life
```

tweet_likes & experiment_group *Which tweet has more* ***likes****? Does it happen more often on liberal or conservative outlet?* One post has 5446 likes, but the majority (308 out of 478) have 0 likes.

```
tweet_like_count <- data.frame(table(final_tweets$tweet_likes)) %>% arrange(desc(Freq))
colnames(tweet_like_count) <- c("likes_count", "freq")
```

Where is the highest retweet reply? - A tweet addressing FoxNews from someone who is against student loan forgiveness.

```
print(final_tweets[which.max(final_tweets$tweet_likes),]$experiment_group)
```

```
## [1] "foxnews"
```

```
print(final_tweets[which.max(final_tweets$tweet_likes),]$text)
```

```
## [1] "@FoxNews Joe, you cannot spend money without Congress approval. Student loan is not a National
```

```
print(final_tweets[which.max(final_tweets$tweet_likes),]$tweet_likes)
```

```
## [1] 5446
```

On average, does conservative or liberal sources have more likes and retweets? (after discounting the post with 5446) - NPR has the most average likes and average retweets out of all 5 sources. Replies to Foxnews are 3rd from the bottom in average tweets, even though 87% of the replies in the population belongs to them. On average its replies stand 2nd to last, beating USEdGov, who has less than 1 like on average.

```
no_max_likes <- final_tweets %>% filter(tweet_likes != 5446)
no_max_likes <- no_max_likes %>% group_by(experiment_group) %>%
  summarize(avg_likes = mean(tweet_likes), agg_likes = sum(tweet_likes),
            avg_retweets = mean(retweets), agg_retweets = sum(retweets))
no_max_likes
```

```
## # A tibble: 6 x 5
##   experiment_group avg_likes agg_likes avg_retweets agg_retweets
##   <chr>                <dbl>     <int>        <dbl>        <int>
## 1 cnn                  2.25         27       0.0833            1
## 2 foxnews              1.23        502       0.0369           15
## 3 msnbc                1.48         31       0.143             3
## 4 npr                  8.23        107       0.615             8
## 5 nytimes              2.5          20       0                 0
## 6 usedgov              0.833         5       0                 0
```

- When combining the liberal sources, the liberal sources on average have 30% more likes and has 6 times more average retweets than the conservative foxnews.

```
no_max_likes <- final_tweets %>% filter(tweet_likes != 5446) %>%
  mutate(politics = ifelse(experiment_group %in% c('cnn', 'msnbc', 'npr', 'nytimes'), 'liberal',
                          ifelse(experiment_group == 'usedgov', 'controlled', 'conservative')))
no_max_likes <- no_max_likes %>% group_by(politics) %>%
  summarize(avg_likes = mean(tweet_likes), agg_likes = sum(tweet_likes),
            avg_retweets = mean(retweets), agg_retweets = sum(retweets))
no_max_likes
```

```
## # A tibble: 3 x 5
##   politics     avg_likes agg_likes avg_retweets agg_retweets
##   <chr>            <dbl>     <int>        <dbl>        <int>
## 1 conservative     1.23        502       0.0369           15
## 2 controlled       0.833         5       0                 0
## 3 liberal          3.43        185       0.222            12
```

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?

```
today <- as.Date("2022-12-08")

ft <- final_tweets %>% mutate(age_dow = substr(created_at, 1, 3)
                              , age_month_day = substr(created_at, 5, 10)
                              , age_time = substr(created_at, 12, 19)
                              , age_yr = substr(created_at, 26, 30),
                              , age_ymd = as.Date(paste0(age_month_day, age_yr), format = "%b %d %Y"),
                              , account_age = today - age_ymd)

today <- as.Date("2022-12-08")
ft <- ft %>% mutate(account_age = today - age_ymd)

print(paste("min age of acct (days): ", min(ft$account_age)))
```

```
## [1] "min age of acct (days):  5"
```

```
print(paste("max age of acct (days): ", max(ft$account_age)))
```

```
## [1] "max age of acct (days):  5257"
```

```
print(paste("mean age of acct (days): ", mean(ft$account_age)))
```

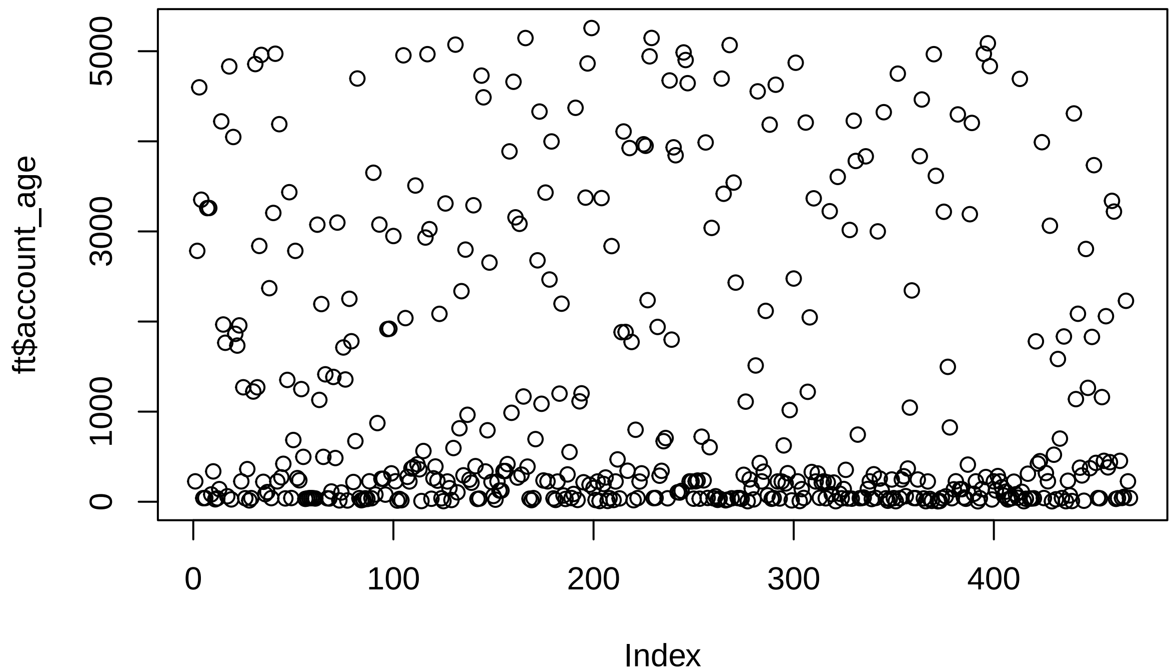
```
## [1] "mean age of acct (days):  1184.82905982906"
```

```
print(paste("median age of acct (days): ", median(ft$account_age)))
```

```
## [1] "median age of acct (days):  268.5"
```

The youngest `account_age` is 5 days, and the oldest account is 14 years (5257 days)

```
plot(ft$account_age)
```



While most accounts are under 3 years old, there are a handful of accounts in the 4000-5000 days range. Let's look at the text of the accounts with more than 5000 days in age. 6 accounts are over 5000 days old. Majority of them are critical to student loan forgiveness.

One text https://twitter.com/jack__jackson/status/1598689928946323458 @ both NPR and FoxNews. However, the text is an original text (`in_reply_to_status_id` is N/A). Maybe it's okay to keep the `experiment_group` as NPR since mentioning them first prioritize them over Foxnews?

```
ft %>% filter(account_age > 5000) %>% select(experiment_group, text)
```

```
##   experiment_group
## 1          foxnews
## 2              npr
## 3          foxnews
## 4              npr
## 5          foxnews
## 6          foxnews
##
## 1
## 2                       @NPR How about, in the meantime, Congress just rewrites the law that makes s
## 3
## 4 @npr @foxnews @dnc @gop Our Constitution requires that all ins and outs of the Treasury originate a
## 5
## 6                              @FoxNews Yeah... no... I'm not paying anything extra for that barista
```

description

How many have profile descriptions? More than half of the tweeters don't have an account profile description. Are the share of those with and without description proportional based on who they reply to?

```
(final_tweets %>% mutate(has_profile_desc = ifelse(nchar(description) == 0, 0, 1)) )%>% group_by(has_pro
```

```
## # A tibble: 2 x 2
##   has_profile_desc agg_profile_desc
##              <dbl>            <int>
## 1                0              248
## 2                1              220
```

location

How many have profile location display? Is one location more dense?

Most of the tweets belong to tweeter with no locations (66%).

```
(final_tweets %>% mutate(has_profile_loc= ifelse(nchar(location) == 0, 0, 1)) )%>% group_by(has_profile_
```

```
## # A tibble: 2 x 2
##   has_profile_loc agg_profile_loc
##             <dbl>           <int>
## 1               0             312
## 2               1             156
```

ymd & dow

Which day of the week do people discuss student loan forgiveness the most often? Recall that data is from Sunday 11/27 - Tuesday 12/6Thursdays and Saturdays get the most tweets.

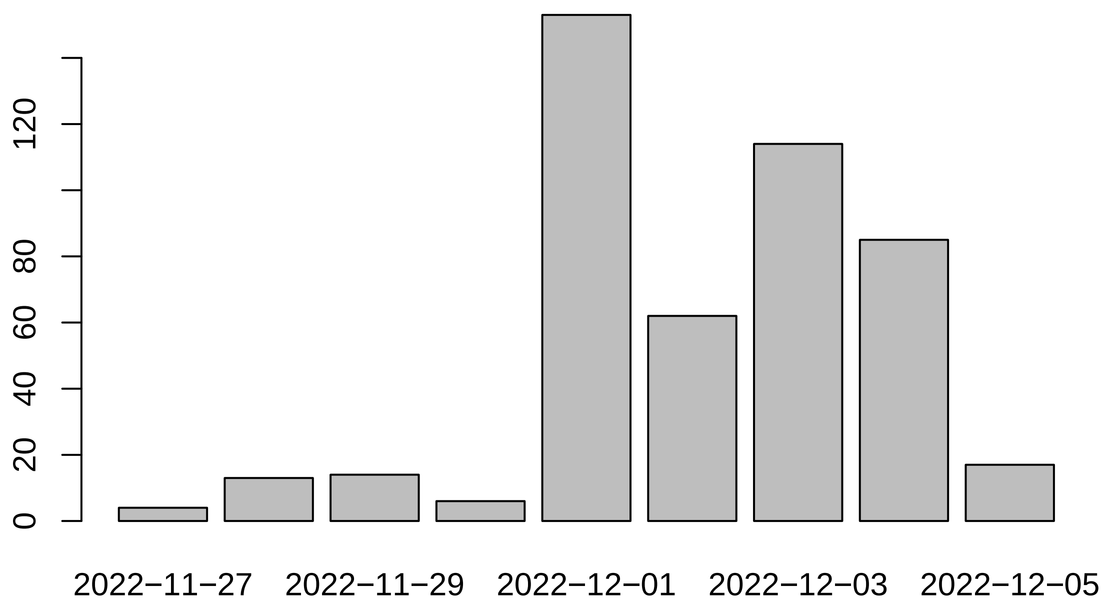
```
final_tweets %>% group_by(dow) %>% summarize(tweet_count = n())
```

```
## # A tibble: 7 x 2
##   dow   tweet_count
##   <chr>       <int>
## 1 Fri            62
## 2 Mon            30
## 3 Sat           114
## 4 Sun            89
## 5 Thu           153
## 6 Tue            14
## 7 Wed             6
```

bar plot of tweet count by day 5 of 9 days have fewer than 20 tweets. Thursday Dec. 1st makes up 33% of all tweets. On Dec 1st, Supreme Court announced they will expedite the process. - https://www.nytimes.com/2022/12/01/us/politics/supreme-court-student-loan-forgiveness.html - https://www.washingtonpost.com/politics/2022/12/01/supreme-court-review-student-loan-forgiveness/

```
ymd_data <- final_tweets %>% group_by(ymd) %>% summarize(tweet_count = n())

barplot(height = ymd_data$tweet_count, names = (ymd_data$ymd))
```



time

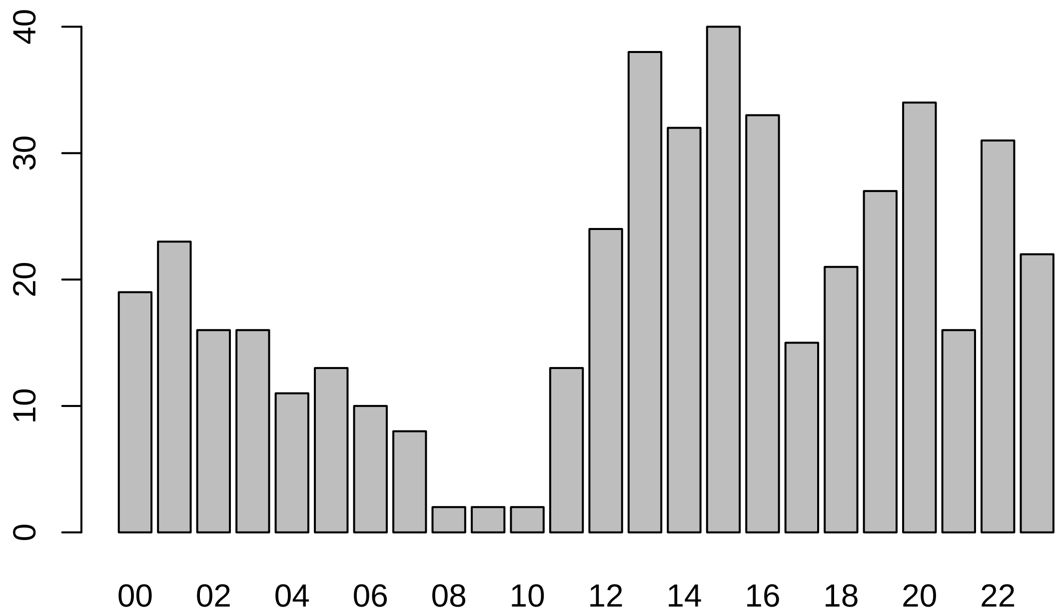
What time of day has the most discussion?

```
time <- final_tweets %>% mutate(hour = substr(time, 1, 2))
time_gr <- time %>% group_by(hour) %>% summarize(freq_by_hr = n())
time_gr
```

```
## # A tibble: 24 x 2
##    hour  freq_by_hr
##    <chr>      <int>
##  1 00            19
##  2 01            23
##  3 02            16
##  4 03            16
##  5 04            11
##  6 05            13
##  7 06            10
##  8 07             8
##  9 08             2
## 10 09             2
## # ... with 14 more rows
## # i Use `print(n = ...)` to see more rows
```

bar plot of tweet count by hour Tweets on this topic lulls between 8-10 am. The afternoon has the highest engagement, with a decrease before commuting time, and an rise right after.

```
barplot(height = time_gr$freq_by_hr, names = (time_gr$hour))
```



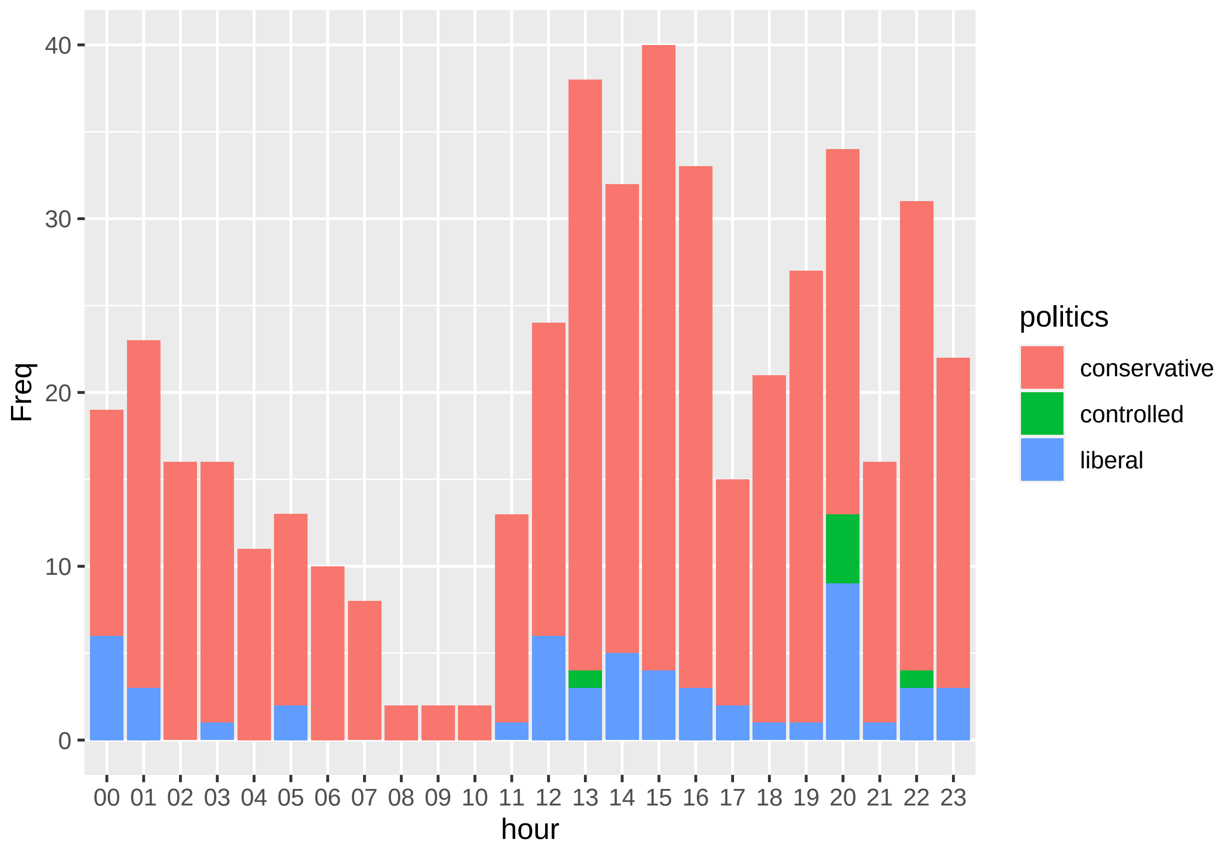
bar plot with multiple colors for conservative vs. liberal

8pm is a popular time for engagement within our control and liberal groups.

```
final_tweets <- final_tweets %>%
  mutate(hour = substr(time, 1, 2),
         politics = ifelse(experiment_group %in% c('cnn', 'msnbc', 'npr', 'nytimes'), 'liberal',
                           ifelse(experiment_group == 'usedgov', 'controlled', 'conservative')))
stacked_time <- final_tweets %>% group_by(politics, hour) %>% summarize(Freq = n())
```

```
## `summarise()` has grouped output by 'politics'. You can override using the
## `.groups` argument.
```

```
ggplot(stacked_time, aes(fill=politics, y=Freq, x=hour)) +
    geom_bar(position="stack", stat="identity")
```

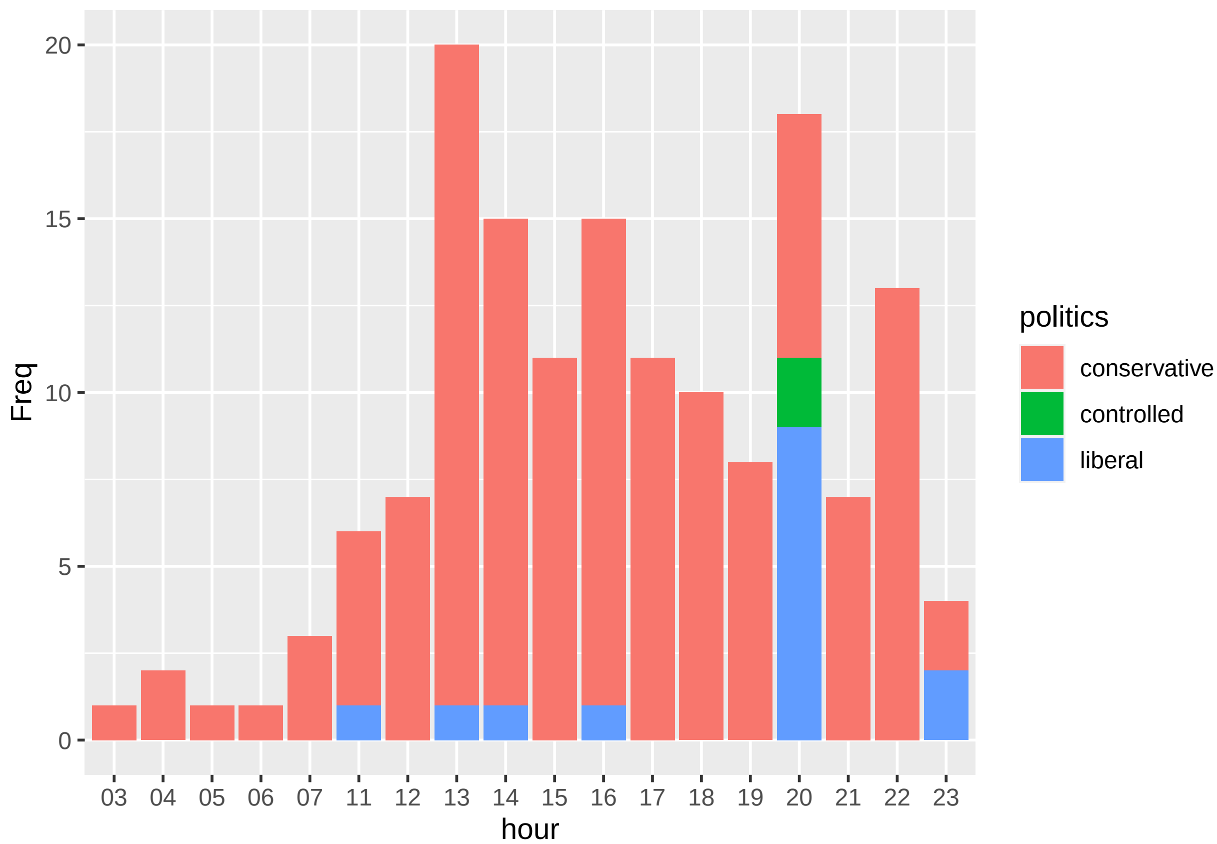


Is the 8pm mostly due to the Supreme Court announcement on Dec.1st? Yes, Dec 1st makes up over 50% of total tweets at 8pm.

```
dec1_stacked_time <- final_tweets %>% filter(ymd == "2022-12-01") %>%
  group_by(politics, hour) %>% summarize(Freq = n())
```

```
## `summarise()` has grouped output by 'politics'. You can override using the
## `.groups` argument.
```

```
ggplot(dec1_stacked_time, aes(fill=politics, y=Freq, x=hour)) +
    geom_bar(position="stack", stat="identity")
```



User Popularity

`favourites_count` & `followers_count` Which media sources has engagement from the most favorite tweeter?

Tweeters engaging with liberal news source had 5 times more profile favorites and 3.6 times more followers, on average. Although fewer tweets addressed the controlled sources, they have more followers on average than accounts engaging in both conservative and liberal media.

```
fav_counts_tweet <- data.frame(table(final_tweets$favourites_count)) %>% arrange(desc(Freq))

fc <- final_tweets %>% #filter(tweet_likes != 5446) %>%
  mutate(politics = ifelse(experiment_group %in% c('cnn', 'msnbc', 'npr', 'nytimes'), 'liberal',
                           ifelse(experiment_group == 'usedgov', 'controlled', 'conservative')))
fc <- fc %>% group_by(politics) %>%
  summarize(avg_fav = mean(favourites_count), agg_likes = sum(favourites_count),
            avg_followers = mean(followers_count), agg_retweets = sum(followers_count))
fc
```

```
## # A tibble: 3 x 5
##   politics     avg_fav agg_likes avg_followers agg_retweets
##   <chr>          <dbl>     <int>         <dbl>        <int>
## 1 conservative   4024.   1641955          86.1        35144
## 2 controlled    10502.     63013         360.          2157
## 3 liberal       20764.   1121263         309.         16681
```

`verified` **Are there any verified accounts? If so, where did they engage with?** None of the author is verified.

```
unique(final_tweets$verified)
```

```
## [1] "False"
```

EDA - with annotations

```r
annotated <- read.csv("data/master_annotated.csv")
glimpse(annotated)
```

```
## Rows: 468
## Columns: 33
## $ X                      <int> 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, ~
## $ experiment_id          <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14,~
## $ experiment_group       <chr> "msnbc", "msnbc", "msnbc", "msnbc", "msnbc", "~
## $ text                   <chr> "@MSNBC @MaddowBlog "Simpleton's defense"?  Yo~
## $ tweet_id               <dbl> 1.6e+18, 1.6e+18, 1.6e+18, 1.6e+18, 1.6e+18, 1~
## $ 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> 1.520000e+18, 3.202809e+09, 1.409157e+08, 1.93~
## $ in_reply_to_status_id  <dbl> 1.6e+18, 1.6e+18, 1.6e+18, 1.6e+18, 1.6e+18, 1~
## $ 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> "27-Nov", "27-Nov", "27-Nov", "27-Nov", "28-No~
## $ time                   <chr> "22:01:59", "22:22:27", "22:39:00", "23:13:38"~
## $ yr                     <int> 2022, 2022, 2022, 2022, 2022, 2022, 2022, 2022~
## $ ymd                    <chr> "11/27/22", "11/27/22", "11/27/22", "11/27/22"~
## $ tweet_id_char          <dbl> 1.6e+18, 1.6e+18, 1.6e+18, 1.6e+18, 1.6e+18, 1~
## $ 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               <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALS~
## $ user_id_char           <dbl> 1.520000e+18, 3.202809e+09, 1.409157e+08, 1.93~
## $ text_length            <int> 183, 114, 148, 226, 159, 93, 136, 213, 169, 16~
## $ text_word_count        <int> 30, 20, 20, 46, 27, 12, 22, 40, 29, 30, 31, 11~
## $ opinion_key            <int> 0, 1, 1, 2, 0, 1, 2, 0, 0, 0, 1, 2, 0, 1, 2, 3~
## $ opinion_label          <chr> "FOR student loan forgiveness", "NEUTRAL suppo~
## $ ego_involvement_key    <int> 0, 1, 1, 1, 1, 1, 1, 0, 0, 1, 0, 1, 0, 0, 1, 1~
## $ ego_involvement_label  <chr> "Very important", "Somewhat important", "Somew~
```

```r
# split up (profile) created_at, today = 12/8/22
annotated <- annotated %>% mutate(age_dow = substr(created_at, 1, 3)
                             , age_month_day = substr(created_at, 5, 10)
                             , age_time = substr(created_at, 12, 19)
                             , age_yr = substr(created_at, 26, 30),
                             , age_ymd = as.Date(paste0(age_month_day, age_yr), format = "%b %d %Y"),
                             , account_age = today - age_ymd)

annotated <- annotated %>% mutate(politics = ifelse(experiment_group %in% c('cnn', 'msnbc', 'npr', 'nyt
                          ifelse(experiment_group == 'usedgov', 'controlled', 'conservative')))
```

`opinion_key` & `opinion_label`

41% of tweets are NEUTRAL in support of student loan forgiveness.29% are AGAINST, and 26% are FOR. Only 4% of the tweets are undetermined in sentiment.

```
annotated %>% group_by(opinion_label) %>% summarize(count = n(), proportion = n()/nrow(annotated))
```

```
## # A tibble: 4 x 3
##   opinion_label                    count proportion
##   <chr>                            <int>      <dbl>
## 1 AGAINST student loan forgiveness   110     0.235
## 2 cannot judge support                22     0.0470
## 3 FOR student loan forgiveness       123     0.263
## 4 NEUTRAL support                    213     0.455
```

Surprisingly, engagement with liberal has higher sentiment against student loan forgiveness (43%) compared to those engaging with FoxNews (27%). The conservative groups replies are mostly in the NEUTRAL support group (43%). **This raise the question of do people tend to reply to sources that they oppose (conservatives replying to liberal sources) or is there more dissent among those engaging with the liberal sources?** Both liberal and conservative sources have ~25% supportive replies for the forgiveness program.

```
## `summarise()` has grouped output by 'politics'. You can override using the
## `.groups` argument.
## `summarise()` has grouped output by 'politics'. You can override using the
## `.groups` argument.
## `summarise()` has grouped output by 'politics'. You can override using the
## `.groups` argument.
```

```
## # A tibble: 10 x 4
## # Groups:   politics [3]
##    politics     opinion_label                    count proportion
##    <chr>        <chr>                            <int>      <dbl>
##  1 conservative AGAINST student loan forgiveness   102     0.25
##  2 conservative cannot judge support                19     0.0466
##  3 conservative FOR student loan forgiveness       103     0.252
##  4 conservative NEUTRAL support                    184     0.451
##  5 controlled   FOR student loan forgiveness         2     0.333
##  6 controlled   NEUTRAL support                      4     0.667
##  7 liberal      AGAINST student loan forgiveness     8     0.148
##  8 liberal      cannot judge support                 3     0.0556
##  9 liberal      FOR student loan forgiveness        18     0.333
## 10 liberal      NEUTRAL support                     25     0.463
```

Future expansion - grab the profile description of each author's "friend/following" and create a threshold label on political affiliation based on verified profiles of those they follow. NLP through the profile description will also let us know if they are more left or right leaning. Currently, we cannot determine if the author political stand based on which news outlet they engage with on twitter (example @BUnskinkable appears to be more right leaning based on who he follows but he addressed @MSNBC)

```
annotated %>% filter(screen_name == 'BUnskinkable') %>% select(experiment_group, text, tweet_id_char, s
```

```
##   experiment_group
## 1            msnbc
##
## 1 @MSNBC Good, tax payers shouldn’t be on the hook for the student loans. Thier decision not ours. G
##   tweet_id_char  screen_name
## 1       1.6e+18 BUnskinkable
```

Let’s flip and look at “FOR student loan forgiveness” on the conservative side. @Richard41020: “@FoxNews Since I don’t have any student loans I’d like for the government (Taxpayers), to payoff my mortgage. Where do I sign up?”

Although it’s categorized as FOR loan forgiveness, it is sarcasm. After exploring the profile, it is apparent that this author is conservative and does not support loan forgiveness.

```
annotated %>% filter(screen_name == 'Richard41020') %>% select(experiment_group, text, tweet_id_char, s
```

```
##   experiment_group
## 1          foxnews
##
## 1 @FoxNews Since I don’t have any student loans I’d like for the government (Taxpayers), to payoff m
##   tweet_id_char  screen_name
## 1       1.6e+18 Richard41020
```

Let’s choose another author. @TonyShockey6 is labeled as FOR forgiveness with .95 confidence, but it appears that his text does not support this annotation :(

```
annotated %>% filter(screen_name == 'TonyShockey6') %>% select(experiment_group, text, tweet_id_char, s
```

```
##   experiment_group
## 1          foxnews
##
## 1 @FoxNews FJB &amp; your student loans!!!\nYou signed up for your student loans, so flipping pay th
##   tweet_id_char  screen_name
## 1       1.6e+18 TonyShockey6
```

After further skimming, it appears that a lot of these FOR student loan forgiveness is categorized incorrectly based on the text provided by the annotators.

```
annotated %>% filter(politics == 'conservative', opinion_label == 'FOR student loan forgiveness ') %>%
```

```
## [1] text
## <0 rows> (or 0-length row.names)
```

ego_involvement_label

75% of the tweets are from authors who find that student loan forgiveness issue is at least somewhat important. Only 15% have low ego involvement.

```
annotated %>% group_by(ego_involvement_label) %>% summarize(count = n(), proportion = n()/nrow(annotate
```

```
## # A tibble: 4 x 3
##   ego_involvement_label   count proportion
##   <chr>                   <int>      <dbl>
## 1 cannot judge importance     6     0.0128
## 2 Not important at all       21     0.0449
## 3 Somewhat important        258     0.551
## 4 Very important            183     0.391
```

`opinion_annotation_confidence`

What is the average of the confidence? What is the average of confidence of each annotated categories? What is the confidence for each news source?

Our observations above that many of the text care incorrectly identified in opinion on student loan forgiveness program. We forgo answer the questions above and pivot to looking at how NLP and built in sentiment analysis fair against SageMaker's Mechanical Turks.

Excess

`ego_involvement_annotation_confidence`

Give summary of the stand

What is the stand of the "older twitter accounts"?

Stacked bar plot on views on student loans and how it matches up against `experiment_group` ggplot(stacked_time, aes(fill=politics, y=Freq, x=hour)) + geom_bar(position="stack", stat="identity")

Is there an reason why someone who is against student loan forgiveness would reply to fox vs. cnn or vice versa? - refer to max likes, why is the opposition not addressing someone?