Fundamentals of Computing and Data Display

Term paper template

Author

2019-12-07

Contents

TO DO: - write results, and discussion sections - create presentation

Introduction

Twitter and Google Trends are becoming increasingly popular tools for social science researchers. They represent an easily accessible source of large amounts of data, which has become advantageous as survey response rates decline and costs rise. Researchers interested in predicting election results have begun looking to Twitter data to replace or supplement traditional election polls, with mixed results (Gayo-Avello 2013).

Recently, there have been studies using sentiment analysis of Twitter data to predict election outcomes in India (Salunkhe and Deshmukh 2017), to predict state-level polling results in the U.S. (Beauchamp 2017), and to predict the winners of three presidential elections in Latin America (Gaurav et al. 2013). Beauchamp (2017) found that Twitter data may be useful in making state-level campaign strategy decisions. Additionally, a recent study used Google Trends and Twitter data to predict the 2016 U.S. election outcomes with only 1% error (Kassraie, Modirshanechi, and Aghajan 2017).

We are interested in whether Twitter data and Google trends data could be used to supplement polling results, by providing real-time information to candidates while they wait for polling data to come in. For example, candidates may be interested in understanding how public opinion has shifted immediately after a debate in order to run a more agile campaign. This project is an exploratory analysis of Twitter and Google Trends data to see if pre- and post-debate polling data during the 2020 U.S. Democratic primary election aligns with these real-time sources.

Data

This section describes the data sources and the data gathering process.

Twitter

Below is our token to access the Twitter API and start collecting tweets. Note that these codes are fake, for security purposes.

```
create_token(
  app = "fcdd-course",
  consumer_key = "XXXXXXXXXXXXXXXXXXX",
  consumer_secret = "XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX",
  access_token = "XXXXXXXXXXXXXXXXXXX-XXXXXXXXXXXXXXXXXXX",
  access_secret = "XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX"
)
```

Quering Tweets

Using the search_tweets function in the rtweets package, we will be quering tweets with the keywords specified below. We limited our tweets pool to November 20th and November 21st. The reasonning behind this is to get a feel post and pre debate.

```
dem <- search_tweets("#democrats OR #candidates OR #election2020 OR #BIDEN OR #sanders OR #warren OR #ha
                     OR #buttigieg OR #steyer OR #yang OR #booker OR #klobuchar OR #gabbard
                     OR @KamalaHarris OR @JoeBiden OR @BernieSanders OR @ewarren OR @PeteButtigieg
                     OR @TomSteyer OR @AndrewYang OR @BookerCory OR @amyklobuchar OR @TulsiGabbard ",  s
```

Data Cleaning

Even after we specify the keywords, we know that we would still get irrelevants tweets mainly with the Ukraine issue at the time of pulling. We removed all tweets that had the word "Ukraine" in it.

```
myvars <- c("text", "location", "created_at")
df1 <- dem[myvars]
df2 <- dem2[myvars]
```

```
df1 <- df1 %>%
      mutate( ukraine = (str_detect(df1$text, regex("Ukraine", ignore_case = TRUE)))) %>%
      filter(ukraine=="FALSE") %>%
      select(-ukraine)

df2 <- df2 %>%
      mutate( ukraine = (str_detect(df2$text, regex("Ukraine", ignore_case = TRUE)))) %>%
      filter(ukraine=="FALSE") %>%
      select(-ukraine)
```

```
tweets <- rbind(df1, df2)
```

Next, we seperate the tweets by candidates in order to do the analysis at the candidate level. It's important to note that one tweets can be addressed to multiple candidates. In that case, the tweet will be found in the each of those candidate dataset. The code below exemplifies this step of the cleaning process.

```
tweets$BS <-(str_detect(tweets$text, regex("#Sanders|@BernieSanders", ignore_case = TRUE)))
tweets$KH <-(str_detect(tweets$text, regex("#harris |@KamalaHarris",  ignore_case = TRUE)))
tweets$JB <-(str_detect(tweets$text, regex("#biden |@JoeBiden", ignore_case = TRUE)))
tweets$EW <-(str_detect(tweets$text, regex("#warren |@ewarren", ignore_case = TRUE)))
tweets$PB <-(str_detect(tweets$text, regex("#buttigieg|@PeteButtigieg", ignore_case = TRUE)))
tweets$TS <-(str_detect(tweets$text, regex("#steyer | @TomSteyer", ignore_case = TRUE)))
tweets$AY <-(str_detect(tweets$text, regex("#yang| @AndrewYang", ignore_case = TRUE)))
```

```
tweets$BC <-(str_detect(tweets$text, regex("#booker | @BookerCory", ignore_case = TRUE)))
tweets$AK <-(str_detect(tweets$text, regex("#klobuchar | @amyklobuchar", ignore_case = TRUE)))
tweets$TG <-(str_detect(tweets$text, regex("#gabbard |@TulsiGabbard", ignore_case = TRUE)))
```

Here we reformat the created_at column as a date & time variable in order to seperate the tweets in two groups : post and pre debate.

```
tweets <- tweets %>%
  mutate( date= as.POSIXct(created_at, tryFormats = c("%Y-%m-%d %H:%M:%OS")))

df1 <- df1 %>%
    mutate( date= as.POSIXct(created_at, tryFormats = c("%Y-%m-%d %H:%M:%OS")))

df2 <- df2 %>%
    mutate( date= as.POSIXct(created_at, tryFormats = c("%Y-%m-%d %H:%M:%OS")))
```

All the tweets received before 9 p.m. on November 20th are accounted for in the pre-debate dataset and all tweets from 11 p.m. on November 20th to the next day are in the post-debate dataset. Note that tweets during the debate(9 - 11 p.m are ignored)

```
pre_debate <- tweets %>%
              filter(date(date) == "2019-11-20" & hour(date) < 21 )

post_debate <- tweets %>%
              filter(date(date) == "2019-11-20" & hour(date) >= 23 | date(date) == "2019-11-21"  )


pre_debate <- pre_debate %>% select(-date, -created_at)
post_debate <- post_debate %>% select(-date, -created_at)
```

Sentiment Analysis

Sentiment analysis of the tweets will performed for only 5 candidates.

First we need to prep the tweets by removing all non-words such as emojis and use the sentiment analysis package for analysis. Althought the sentiment analysis packae uses 5 dictionnairies, we will only look at the results from the GI dictionnary. For the pre-debate tweets sentimeent we used all the tweets from the dataset; however for the post-debate analysis, we sampled from the dataset due to volume and processing error. The code below exemplifies this process.

```
bs <- pre_debate %>%
      filter (BS == "TRUE")

usableText=str_replace_all(bs$text,"[^[:graph:]]", " ")

usableText <- tolower(bs$text)

usableText<- iconv(usableText, "UTF-8","ASCII",  sub="byte")

sentiments_bs = analyzeSentiment(as.character(usableText))
```

```
pre_sent_pos <- data.frame("Candidate" = c("Biden", "Sanders", "Warren", "Harris", "Buttigieg"),
                        "Positive" =c((sum(sentiments_jb$PositivityGI))/27477,
                                      (sum(sentiments_bs$PositivityGI))/31603,
                                     (sum(sentiments_kh$PositivityGI))/16362,
                                      (sum(sentiments_pb$PositivityGI))/27242))
```

Google Trends

This section describes gathering the Google Trends data using the package gtrends. First, we register our Google API key. Then, we pull data for each candidate from the 2 days preceding and two days following the debate (November 18 through 22). We have to pull the data in two separate blocks, because gtrends only allows us to use five search terms at a time. We limit the location of searches to the US. Please note that the key here is not a real API key, for security purposes.

```
register_google(key = "XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX")
res1 <- gtrends(c("Joe Biden", "Bernie Sanders", "Elizabeth Warren", "Kamala Harris", "Pete Buttigieg")
                time = "2019-11-18 2019-11-22", low_search_volume = T)

res2 <- gtrends(c("Tom Steyer", "Andrew Yang", "Cory Booker", "Amy Klobuchar", "Tulsi Gabbard"), geo =
                time = "2019-11-18 2019-11-22", low_search_volume = T)
```

Geocoding

Next, we compile and clean the Google Trends location data and prepare it for geocoding.

```
interest_by_location1 <- as_tibble(res1$interest_by_dma)
interest_by_location2 <- as_tibble(res2$interest_by_dma)
interest_by_location <- rbind(interest_by_location1, interest_by_location2)

locations_df <- as.data.frame(interest_by_location)
locations_df$location <- as.character(locations_df$location)
```

Then, we geocode the Google trends data.

```
gc_locations <- as_tibble(mutate_geocode(locations_df, location))
```

Data Cleaning

Next, we categorize the Google Trends data into pre-debate and post-debate data, based on the date.

```
interest_over_time1 <- as_tibble(res1$interest_over_time)
interest_over_time2 <- as_tibble(res2$interest_over_time)
interest_over_time <- rbind(interest_over_time1, interest_over_time2)

interest_over_time_pre <-
  interest_over_time %>%
  filter(date < "2019-11-19") %>%
  mutate(Pre_post="Pre-debate")

interest_over_time_post <-
  interest_over_time %>%
```

```
  filter(date > "2019-11-20") %>%
  mutate(Pre_post="Post-debate")

interest_over_time_all <- rbind(interest_over_time_pre, interest_over_time_post)
```

Polls

Polling data was collected from RealClearPolitics, which aggregates weekly polls. We created a dataset using the RealClearPolitics average before and after the November 20 debate. Because the website changes often, and we only needed a small snapshot of the data, it was more efficient to clean the data in Excel and import to R than to use web scraping.

```
github_link <- "https://github.com/znpadgett/surv727_padgett_thiam/raw/master/Data/Project%20polling%20
temp_file <- tempfile(fileext = ".xlsx")
req <- GET(github_link,
           write_disk(path = temp_file))
polling_data <- readxl::read_excel(temp_file)
```

Results

This section presents the main results.

Data exploration

The results section may have a data exploration part, but in general the structure here depends on the specific project.

Twitter

Create a function to tabulate the count and proportion of tweets by candidates.

Create functions

```
type_var  <- unlist(map(pre_debate, class))

freq_tab <- function(x) {
# make table with count and frequency
tab <- cbind(Count = table(x, useNA = "ifany"),
Prop = round(prop.table(table(x, useNA = "ifany")),
2))
# get the categories as variable and rearrenge
tab <- as.data.frame(tab) %>%
tbl_df() %>%
mutate(Cat = row.names(tab)) %>%
select(Cat, Count, Prop)
}
```

```
props1 <- map(pre_debate[, type_var == "logical"], freq_tab)
props2 <- map(post_debate[, type_var == "logical"], freq_tab)


vars <- unlist(map(props1, nrow))

props_tab1 <- reduce(props1, rbind)
props_tab2 <- reduce(props2, rbind)
```

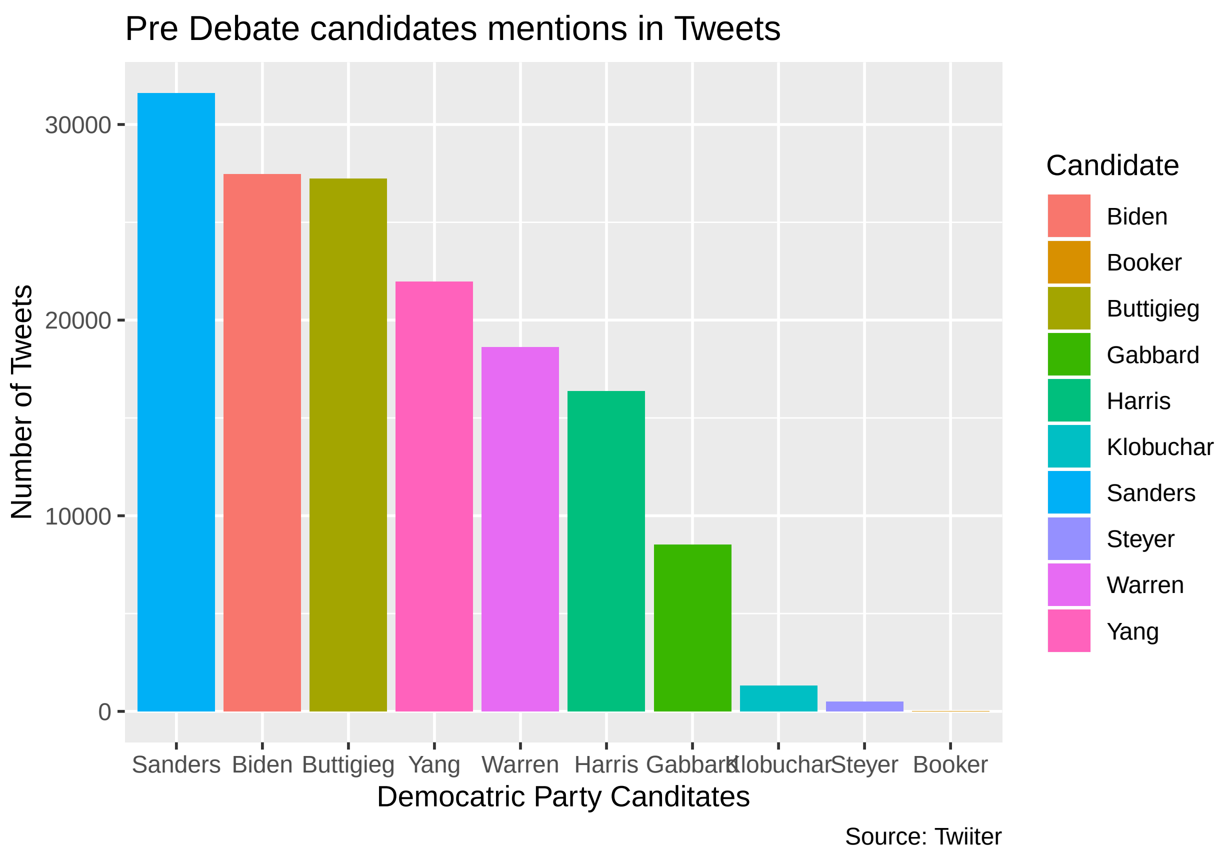
```
props_tab1 <- props_tab1 %>%
mutate(Variable = rep(names(vars), vars),
       Candidate = ifelse(Variable == "BS", "Sanders",
                   ifelse(Variable == "KH", "Harris",
                   ifelse(Variable == "JB", "Biden",
                   ifelse(Variable == "EW", "Warren",
                   ifelse(Variable == "PB", "Buttigieg",
                   ifelse(Variable == "TS", "Steyer",
                   ifelse(Variable == "AY", "Yang",
                   ifelse(Variable == "BC", "Booker",
                   ifelse(Variable == "AK", "Klobuchar",
                   ifelse(Variable == "TG", "Gabbard", NA)))))))))))

props_tab2 <- props_tab2 %>%
mutate(Variable = rep(names(vars), vars),
        Candidate = ifelse(Variable == "BS", "Sanders",
                   ifelse(Variable == "KH", "Harris",
                   ifelse(Variable == "JB", "Biden",
                   ifelse(Variable == "EW", "Warren",
                   ifelse(Variable == "PB", "Buttigieg",
                   ifelse(Variable == "TS", "Steyer",
                   ifelse(Variable == "AY", "Yang",
                   ifelse(Variable == "BC", "Booker",
                   ifelse(Variable == "AK", "Klobuchar",
                   ifelse(Variable == "TG", "Gabbard", NA)))))))))))
```

A visual of the proportion of tweets pre and post debate

Graphing

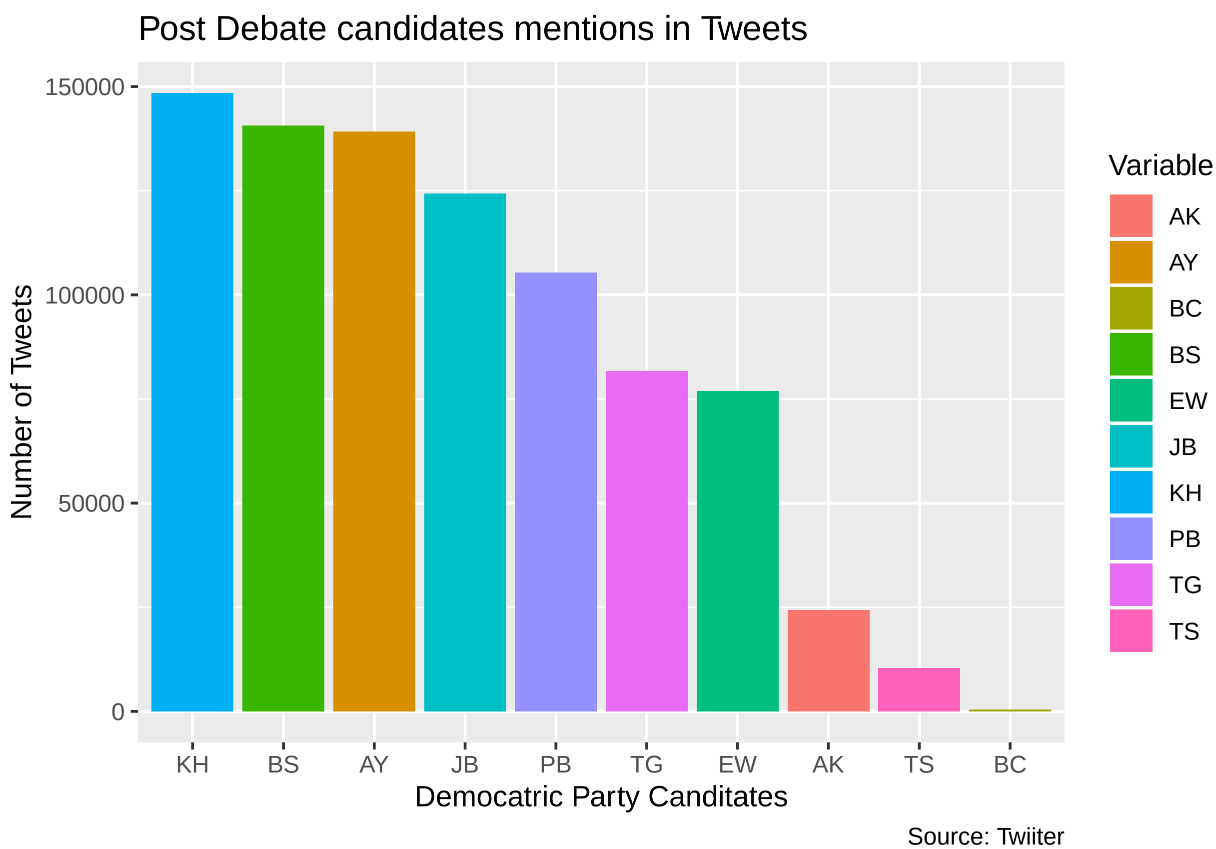
Pre-debate number of tweets

```
props_tab1 %>%
  filter(Cat == "TRUE") %>%
  ggplot() +
  geom_col(mapping = aes(x = reorder(Candidate, -Count), y=Count, fill = Candidate)) +
    labs(x = "Democatric Party Canditates",
       y = "Number of Tweets",
       title = "Pre Debate candidates mentions in Tweets",
       caption = "Source: Twiiter")
```



Post-debate number of tweets

```
props_tab2 %>%
  filter(Cat == "TRUE") %>%
  ggplot() +
  geom_col(mapping = aes(x = reorder(Variable, -Count), y=Count, fill = Variable)) +
    labs(x = "Democatric Party Canditates",
       y = "Number of Tweets",
       title = "Post Debate candidates mentions in Tweets",
       caption = "Source: Twiiter")
```

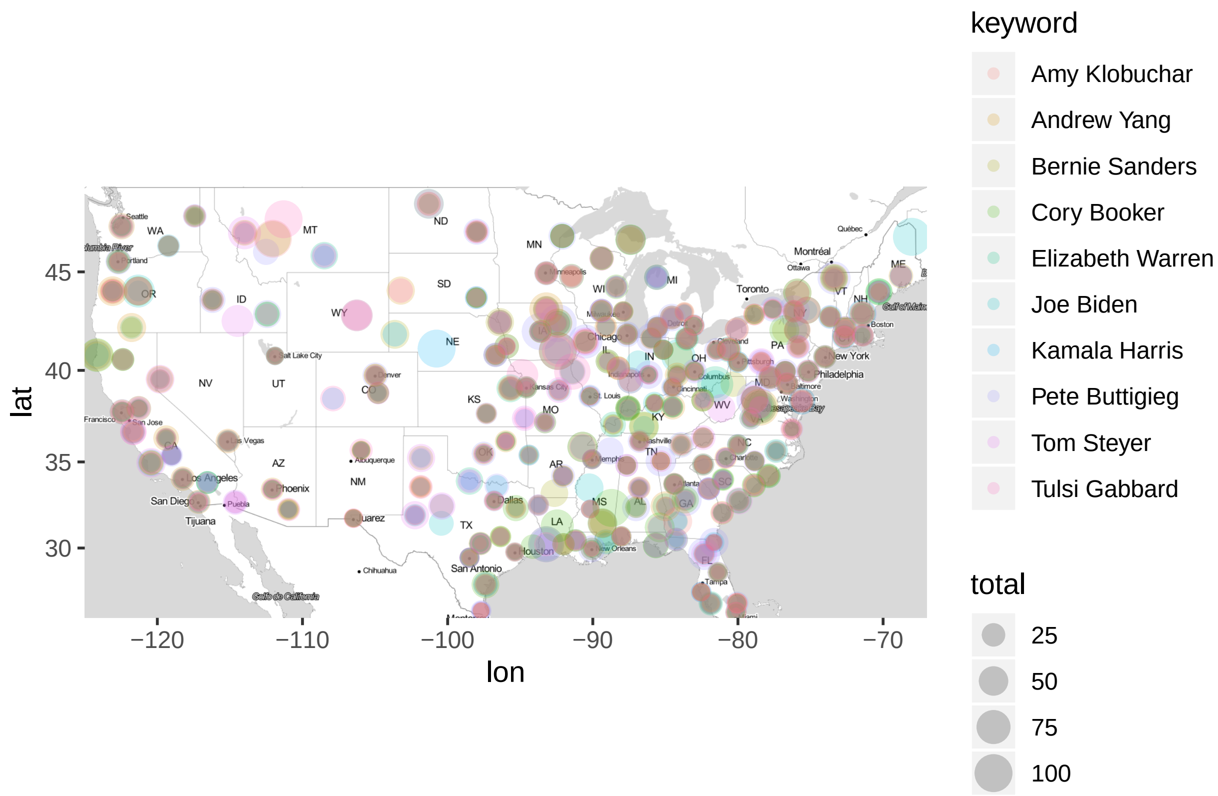


Google Trends

To explore the Google Trends data, we examined the location and density of searches for each candidate. We created a map showing our results. Because it was difficult to see the results for each candidate, we created a shiny app that allows users to toggle between candidates.

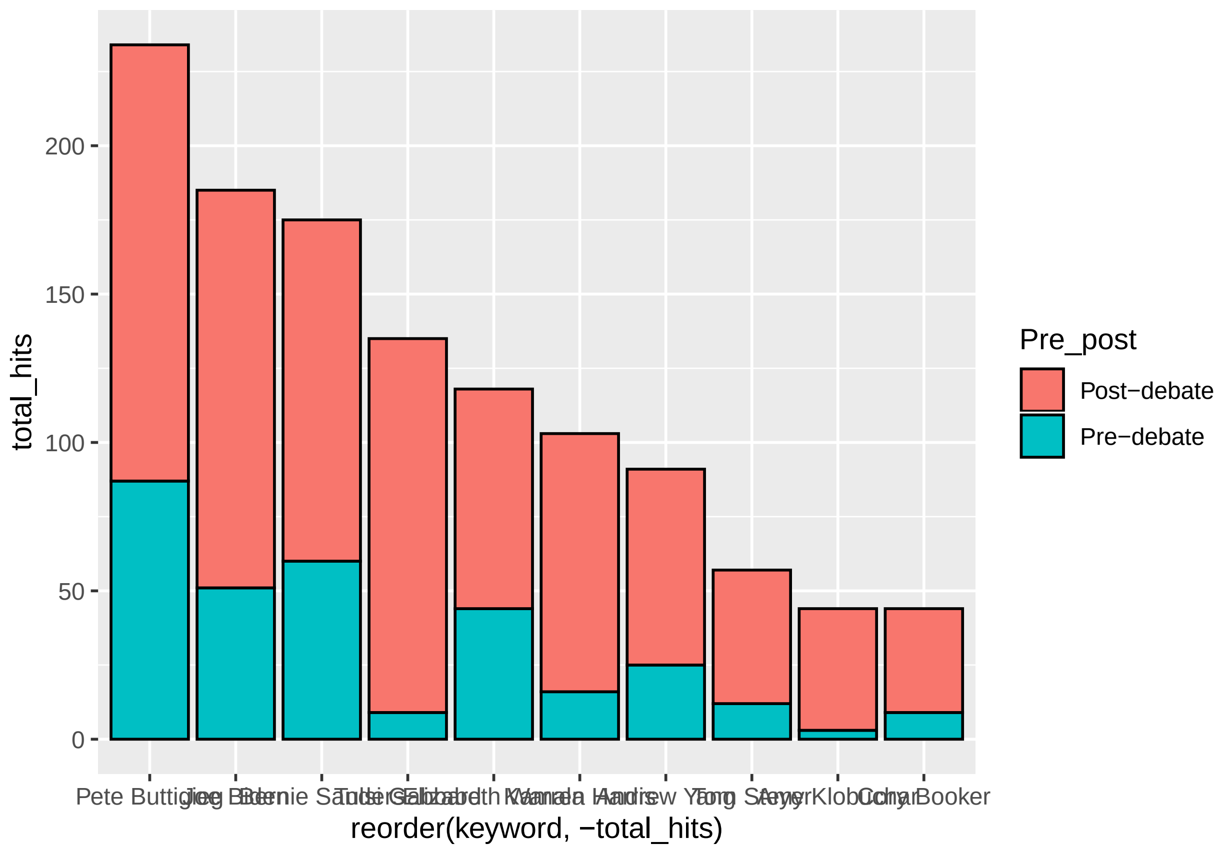
```
#get map
us <- c(left = -125, bottom = 25.75, right = -67, top = 49)
us_map <- get_stamenmap(us, zoom = 5, maptype = "toner-lite")
ggmap(us_map)
```

```
#clean data for mapping
trends_map <-
  gc_locations %>%
  group_by(keyword, location) %>%
  mutate(total=sum(hits))

#generate map
ggmap(us_map) +
  geom_point(data = trends_map, aes(x = lon, y = lat, size=total, color=keyword), alpha = 0.2)
```



We also explored the number of hits pre- and post- debate for each candidate.

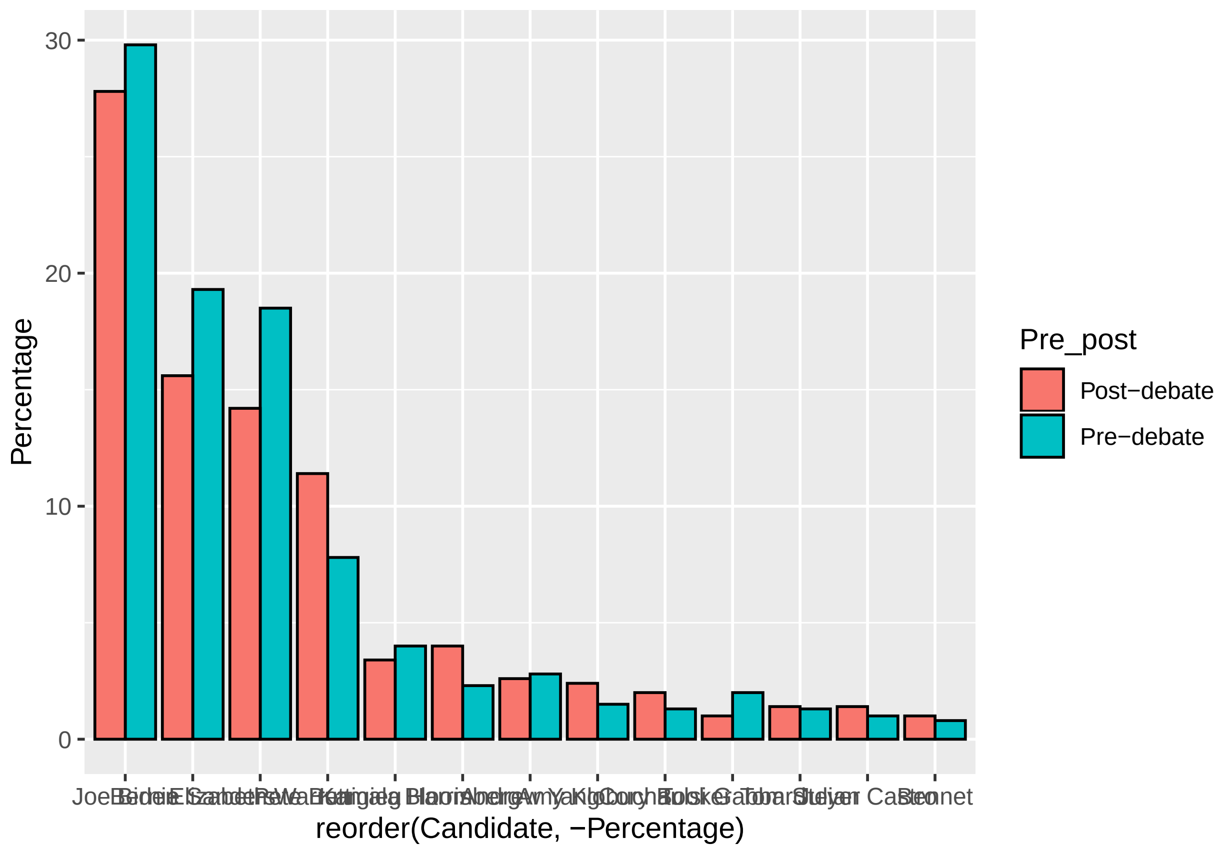
```
interest_over_time_all %>%
  group_by(keyword, Pre_post) %>%
  summarise(total_hits=sum(hits)) %>%
  ggplot() +
  geom_col(mapping = aes(x=reorder(keyword, -total_hits), y=total_hits, fill=Pre_post), color="black")
```



Polling Data

For the polling data, we looked at the pre- and post-debate polling numbers for each candidate.

```r
polling_data %>%
  group_by(Candidate, Pre_post) %>%
  ggplot() +
    geom_col(mapping = aes(x=reorder(Candidate, -Percentage), y=Percentage, fill=Pre_post), color="blac
```



Analysis

We created a Shiny app that allows us to compare the polling, Google Trends, and Twitter sentiment data. First, we cleaned up and combined the three data sources for use in the Shiny app.

```
#clean data
gtrends <-
  interest_over_time_all %>%
  mutate(Candidate=keyword, Data="GTrends") %>%
  group_by(Candidate, Pre_post, Data) %>%
  summarise(Percentage=mean(hits))

poll <-
  polling_data %>%
  select(Candidate, Pre_post, Percentage) %>%
  mutate(Data="Polling") %>%
  group_by(Candidate, Pre_post, Data)

twitter_data <- data.frame("Candidate" = c("Joe Biden", "Bernie Sanders", "Elizabeth Warren",
                                           "Kamala Harris", "Pete Buttigieg", "Joe Biden",
                                           "Bernie Sanders", "Elizabeth Warren", "Kamala Harris",
                                           "Pete Buttigieg"),
                           "Pre_post" =c("Pre-debate","Pre-debate","Pre-debate","Pre-debate","Pre-debat
                   "Data" = c("Twitter", "Twitter","Twitter","Twitter","Twitter",
                              "Twitter","Twitter","Twitter","Twitter","Twitter"),
```

```
                   "Percentage"=c(pre_sent_pos$Positive[1],pre_sent_pos$Positive[2],
                               pre_sent_pos$Positive[3],pre_sent_pos$Positive[4],
                               pre_sent_pos$Positive[5],post_sent_pos$Positive[1],
                               post_sent_pos$Positive[2],post_sent_pos$Positive[3],
                               post_sent_pos$Positive[4],post_sent_pos$Positive[5]))

twitter <- as_tibble(twitter_data)
twitter <-
  twitter %>%
  group_by(Candidate, Pre_post, Data) %>%
  mutate(Percentage=Percentage*100)

all_data <- rbind(poll, gtrends, twitter)
```

Then, we created a shiny app, which can be viewed by accessing the following site: ________________________.

```
# Define UI
ui <- fluidPage(

  # Application title
  titlePanel("2020 Democratic Primary Candidate Data"),

  # Sidebar with a dropdown
  sidebarLayout(
    sidebarPanel(
      selectInput(inputId = "Candidate",
                  label = "Candidate",
                  choices = c("Joe Biden", "Pete Buttigieg", "Kamala Harris", "Bernie Sanders", "Elizabe
                  selected = "Joe Biden"),
      selectInput(inputId = "Data",
                  label = "Data Type",
                  choices = c("Polling", "GTrends", "Twitter"),
                  selected = "Polling")
    ),

    # Show plot
    mainPanel(
      plotOutput(outputId = "graph")
    )
  )
)

# Define server logic
server <- function(input, output) {

  output$graph <- renderPlot({
    all_data %>%
      filter(Candidate == input$Candidate, Data == input$Data) %>%
      ggplot() +
      geom_col(mapping = aes(x=Pre_post, y=Percentage, fill=input$Candidate)) +
      ylim(0,100)
  })
}
```

```r
# Run the application
shinyApp(ui = ui, server = server)
```

```r
# What happens here depends on the specific project
```

```r
# What happens here depends on the specific project
```

Discussion

This section summarizes the results and may briefly outline advantages and limitations of the work presented.

References

Beauchamp, N. 2017. "Predicting and Interpolating State-Level Polls Using Twitter Textual Data." *American Journal of Political Science* 61 (2).

Gaurav, M., A. Srivastava, A. Kumar, and S. Miller. 2013. "Leveraging Candidate Popularity on Twitter to Predict Election Outcome." *Proceedings of the 7th Workshop on Social Network Mining and Analysis*.

Gayo-Avello, D. 2013. "A Meta-Analysis of State-of-the-Art Electoral Prediction from Twitter Data." *Organization Studies* 31 (6): 211–48.

Kassraie, P., A. Modirshanechi, and H. Aghajan. 2017. "Election Vote Share Prediction Using a Sentiment-Based Fusion of Twitter Data with Google Trends and Online Polls." *Proceedings of the 6th International Conference on Data Science, Technology and Applications*.

Salunkhe, P., and S. Deshmukh. 2017. "Twitter Based Election Prediction and Analysis." *International Research Journal of Engineering and Technology* 4 (10).