Final Project: Twitter 2020 Democratic Canidate Sentiment

R Markdown and LaTeX

Forest Krueger and Victoria Owens

12/09/19

Contents

1. Introduction 1
2. Setup 1
3. Data 2
    Gathering 2
    Data Preparation 2
    Dataset Separation 2
4. Sentiment Analysis 3
5. Results 4
6. Discussion 16
7. References 17

1. Introduction

In recent years, Twitter has become an object of interest in public opinion measurement. Prior studies have successfully used Twitter data in lieu of survey data to predict presidential job approval (1), consumer sentiment (1), and the Irish General Election results (2) to name a few. However, Twitter's usefulness as a predictor of public opinion still remains an object of debate as it performs inconsistently.

In this project, we intend to compare sentiment analysis on tweets with favorability polls on leading Democratic candidates Joe Biden, Bernie Sanders, Kamala Harris, Elizabeth Warren, and Pete Buttigieg in the early primary states of Iowa, Nevada, South Carolina, and New Hampshire. If tweets align with favorability polls, it may be feasible for candidates to use them as a source of information to understand more quickly and cheaply the state of public opinion than waiting for poll results.

Our plan is to collect the tweets over several days in late November 2019, analyse the tweets to obtain a sentiment score for each candidate within each state, and compare the distributions of positive-to-negative tweets with their opinion poll counterparts. We hypothesize a similarity in distributions between the two.

2. Setup

Here we load the necessary libraries for preparation.

In addition, the working directory will need to be set. This will likely have to be manually changed.

Finally, we use create_token() to prepare to collect the data from Twitter.

3. Data

Gathering

The data was collected with the code below. This is included only to demonstrate how the data was gathered; the dataset can be loaded in the code chunk following after.

The initial data included 500,000 tweets covering in a 3-day span between November 23 and November 25 that mentioned candidates' twitter handles.

```
all_tweets <- search_tweets("@BernieSanders  OR @ewarren OR @KamalaHarris OR @PeteButtigieg OR @JoeBiden
save(all_tweets,file="all_tweets.Rda")
```

The full dataset as it was collected from Twitter may be loaded below. This data can be downloaded here for the code to be downloaded. It's too large for Github.: https://drive.google.com/open?id=1vuRT8jAY0_Pds-mleHNS8nWsalRa2Z1R

Data Preparation

For our purposes, we are only interested in three variables in the dataframe:

text: text content of the tweet location: Location of user according to the user's bio (not geotagging) created_at: The time and date that the tweet was created

Below, we limit the data to include only these columns. In addition, the text column is partially cleaned in preparation for text analysis by removing unwanted characters.

```
all_tweets_cut <-
  all_tweets %>%
  mutate(txt_clean = str_replace_all(text, "[^[:alnum:]]", " ")) %>%
  select(txt_clean, location, created_at)
```

We also take a quick look at the data to ensure it's what we expect before continuing forward.

```
head(all_tweets_cut, 5)
```

```
## # A tibble: 5 x 3
##   txt_clean                              location        created_at
##   <chr>                                  <chr>           <dttm>
## 1 " HugeHodor  JoeBiden I want hold th~ Beverly Hills ~ 2019-11-25 06:45:00
## 2 " LovelyFunerals  JoeBiden BOT"       Beverly Hills ~ 2019-11-25 06:38:37
## 3 " PattyCullinane  JoeBiden WTF plane~ Beverly Hills ~ 2019-11-25 06:37:11
## 4 " zachsjacobson  JoeBiden Whoever ca~ Beverly Hills ~ 2019-11-25 06:43:11
## 5 The federal minimum wage at the begi~ Brasília, Bras~ 2019-11-25 06:29:54
```

Dataset Separation

We want to examine information at the state- and candidate-level. For this purpose, we will create 20 separate datasets with each combination of candidates and states.

First, we create four datasets, one for each state of interest. Each dataset contains all of the records in which the location mentions a user's full state name ("Iowa") or abbreviation ("IA") which is chosen after someone names a city.

```
IA_filter <- c(" IA", "Iowa")
NH_filter <- c(" NH", "New Hampshire")
SC_filter <- c(" SC", "South Carolina")
NV_filter <- c(" NV", "Nevada")


tweets_IA <- filter (all_tweets_cut, str_detect(location, paste(IA_filter, collapse="|")))

source(file="loop_1.R")

head(tweets_IA, 5)
```

```
## # A tibble: 5 x 3
##   txt_clean                                  location    created_at
##   <chr>                                      <chr>       <dttm>
## 1 "People went to jail  people were beaten~ Iowa, USA   2019-11-25 06:41:45
## 2 Don t miss  ewarren in West Des Moines t~ Des Moines~ 2019-11-25 06:41:07
## 3 " vote20208  RamonaMassachi  ninaturner ~ Cedar Rapi~ 2019-11-25 05:53:26
## 4 " BernieSanders Climate change legislati~ Cedar Rapi~ 2019-11-25 06:40:37
## 5 " TenYearChallenge  Federal Minimum Wage~ Iowa, USA   2019-11-25 06:37:12
```

Then, we create the full 20 datasets by separating each state dataset by candidate. A record corresponds to a candidate if the candidate's Twitter handle is mentioned in the tweet ("text" column). Tweets which mention more than one candidate will be included in each dataset.

```
#Twitter handles
BS_filter <- c("BernieSanders")
EW_filter <- c("ewarren")
KH_filter <- c("KamalaHarris")
PB_filter <- c("PeteButtigieg")
JB_filter <- c("JoeBiden")
#Iowa
tweets_IA_BS <- filter(tweets_IA, str_detect(txt_clean, paste(BS_filter)))

source(file="loop_2.R")
```

4. Sentiment Analysis

Before sentiment analysis can be performed, the text must be formatted properly. First, words to which no sentiment value will be assigned (common nouns in the tweets like "bernie" or common words like "and," as identified by the Harvard dictionary) will be removed from the text.

```
#Problem words specific to our data
remove <- c("t.co","joebiden","berniesanders","https","ewarren","petebuttigieg","bernie","joe","kamalaha
#General stop words
data("stop_words")
```

Now the data is formatted properly to begin the sentiment analysis.

The scores of each word in each tweet will be added together to achieve a tweet-level score using the Harvard dictionary (sentimentGI) that is either positive (>0), negative (<0), or neutral (0).

```
#Sentiment analysis for Sanders in Iowa
BS_IA_sent <- analyzeSentiment(BS_IA_filter$text)
BS_IA_sent$Sentiment<-ifelse(BS_IA_sent$SentimentGI>0,"Pos",ifelse(BS_IA_sent$SentimentGI<0,"Neg","Neut
#Sentiment analysis for the rest of the states and candidates
source(file="loop_4.R")
```

5. Results

Below we include counts of tweets for each candidate/state by positive, negative, and neutral sentiment.

```
print("Sanders Iowa Sentiment:")
```

```
## [1] "Sanders Iowa Sentiment:"
```

```
print(table(BS_IA_sent$Sentiment))
```

```
##
##  Neg Neut  Pos
##   66   65  145
```

```
source(file="loop_5.R")
```

```
## [1] "Warren Iowa Sentiment:"
##
##  Neg Neut  Pos
##  151   93  225
## [1] "Harris Iowa Sentiment:"
##
##  Neg Neut  Pos
##  101  146  384
## [1] "Buttigieg Iowa Sentiment:"
##
##  Neg Neut  Pos
##   32   44  157
## [1] "Biden Iowa Sentiment:"
##
##  Neg Neut  Pos
##  128  159  223
## [1] "Sanders New Hampshire Sentiment:"
##
##  Neg Neut  Pos
##   38   86  202
## [1] "Warren New Hampshire Sentiment:"
##
##  Neg Neut  Pos
##  143  139  239
## [1] "Harris New Hampshire Sentiment:"
##
##  Neg Neut  Pos
##   16   21   38
```

```
## [1] "Buttigieg South Carolina Sentiment:"
##
##  Neg Neut  Pos
##   29   37  128
## [1] "Biden South Carolina Sentiment:"
##
##  Neg Neut  Pos
##  334  277  276
## [1] "Sanders Nevada Sentiment:"
##
##  Neg Neut  Pos
##  104  138  253
## [1] "Warren Nevada Sentiment:"
##
##  Neg Neut  Pos
##  151   74  167
## [1] "Harris Nevada Sentiment:"
##
##  Neg Neut  Pos
##   80   88  192
## [1] "Buttigieg Nevada Sentiment:"
##
##  Neg Neut  Pos
##   48   66  191
## [1] "Biden Nevada Sentiment:"
##
##  Neg Neut  Pos
##  213  200  211
```

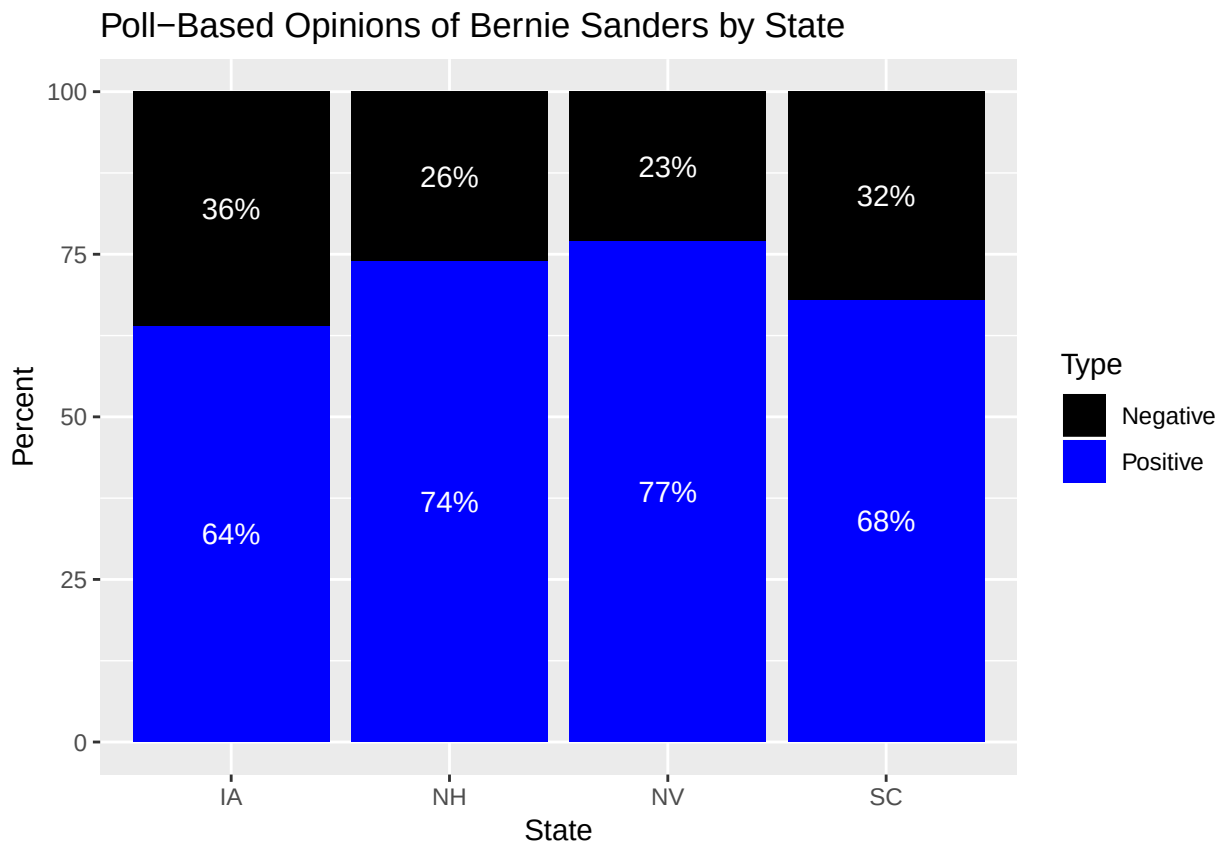
To visualise these results more easily, we plot two separate bar charts for each candidate: one for Twitter, and one for polls, with a bar for each state, divided into positive and negative favorability.

Below, we prepare the polls data for graphing (8, 9, 10, 11).

```
pp <- read.csv("polls_prepped.csv")
pp <-
  pp %>%
  mutate(Percent = round(Percent*100))
```

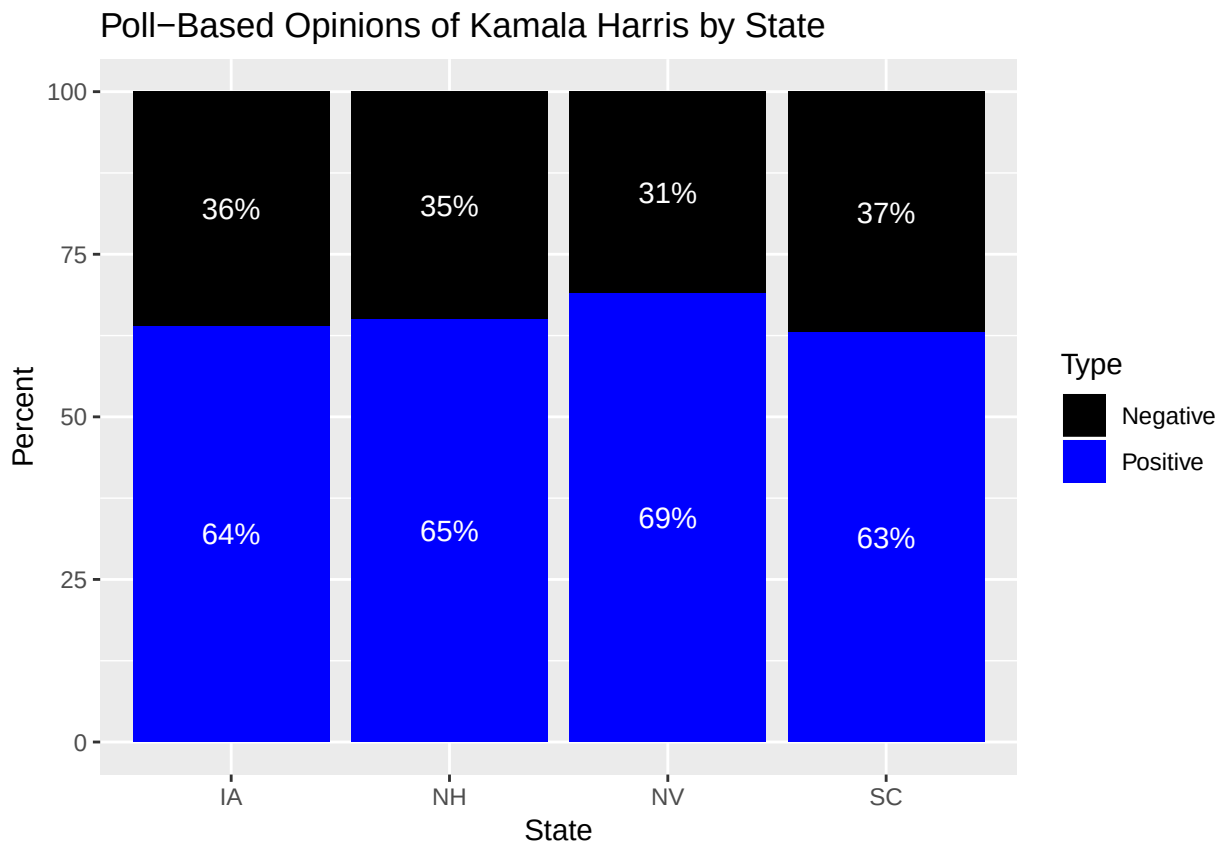
Now we graph the polls data for each candidate by state.

```
#Bernie Sanders
BS <-
  pp %>%
  filter(Candidate == "BS")
#text positioning
BS <- ddply(BS, .(State), transform, pos = cumsum(Percent) - (0.5 * Percent))
#create graph
plot_BS <- ggplot() +
  geom_bar(aes(y = Percent, x = State, fill = Type), data = BS, stat="identity") +
  scale_fill_manual(values=c("black", "blue")) +
  geom_text(data=BS, aes(x = State, y = pos, label = paste0(Percent,"%")), size=4, color="white") +
  ggtitle("Poll-Based Opinions of Bernie Sanders by State")
plot_BS
```



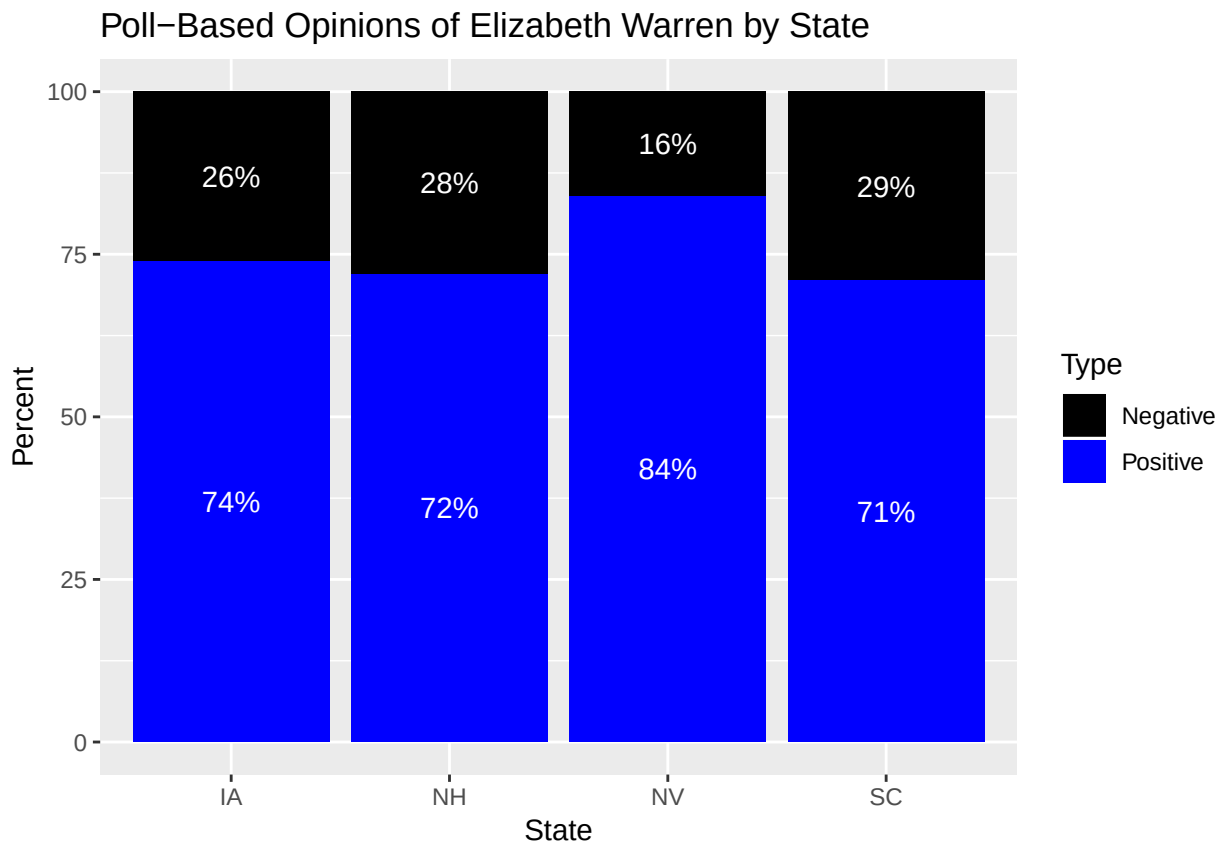

```
#Kamala Harris
KH <-
  pp %>%
  filter(Candidate == "KH")

#text positioning
KH <- ddply(KH, .(State), transform, pos = cumsum(Percent) - (0.5 * Percent))
#create graph
plot_KH <- ggplot() +
  geom_bar(aes(y = Percent, x = State, fill = Type), data = KH,stat="identity") +
  scale_fill_manual(values=c("black", "blue")) +
  geom_text(data=KH, aes(x = State, y = pos, label = paste0(Percent,"%")), size=4, color="white") +
  ggtitle("Poll-Based Opinions of Kamala Harris by State")
plot_KH
```



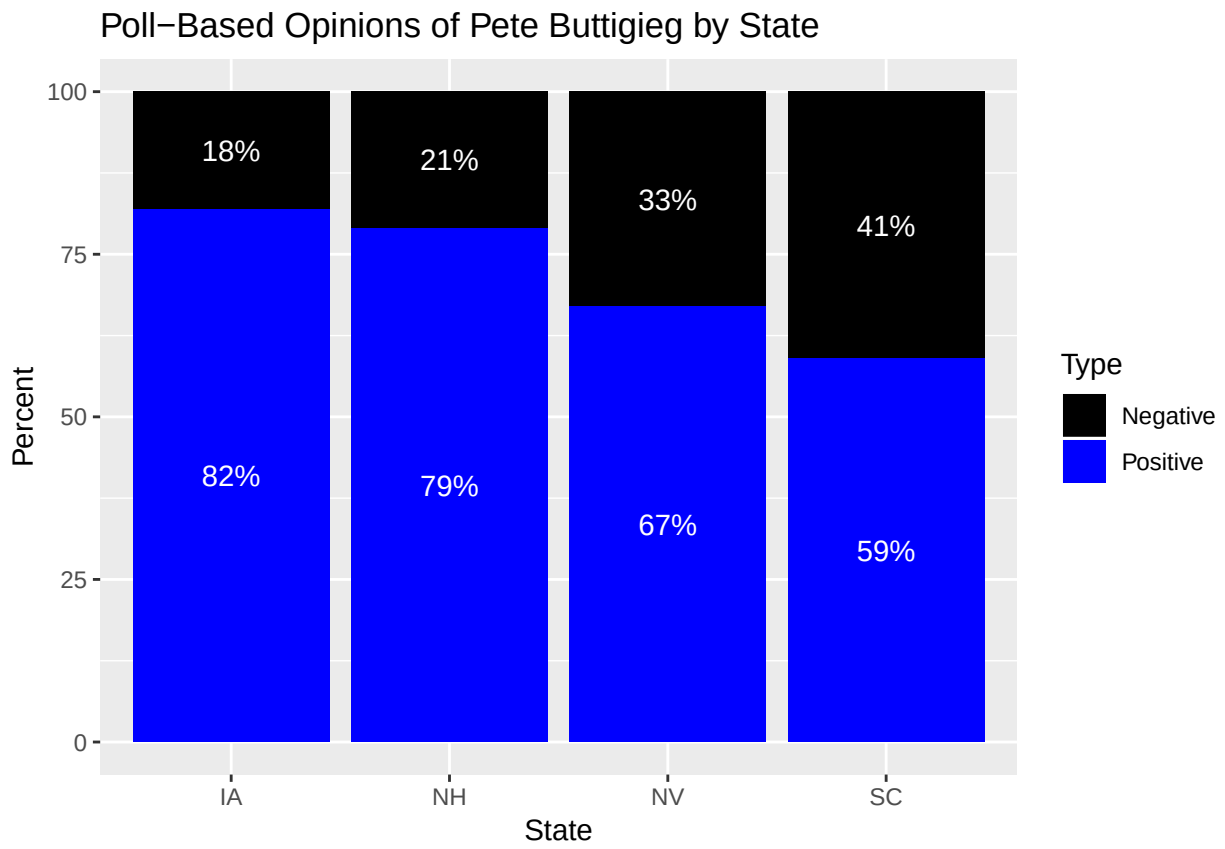

```
#Elizabeth Warren
EW <-
  pp %>%
  filter(Candidate == "EW")

#text positioning
EW <- ddply(EW, .(State), transform, pos = cumsum(Percent) - (0.5 * Percent))
#create graph
plot_EW <- ggplot() +
  geom_bar(aes(y = Percent, x = State, fill = Type), data = EW, stat="identity") +
  scale_fill_manual(values=c("black", "blue")) +
  geom_text(data=EW, aes(x = State, y = pos, label = paste0(Percent,"%")), size=4, color="white") +
  ggtitle("Poll-Based Opinions of Elizabeth Warren by State")
plot_EW
```



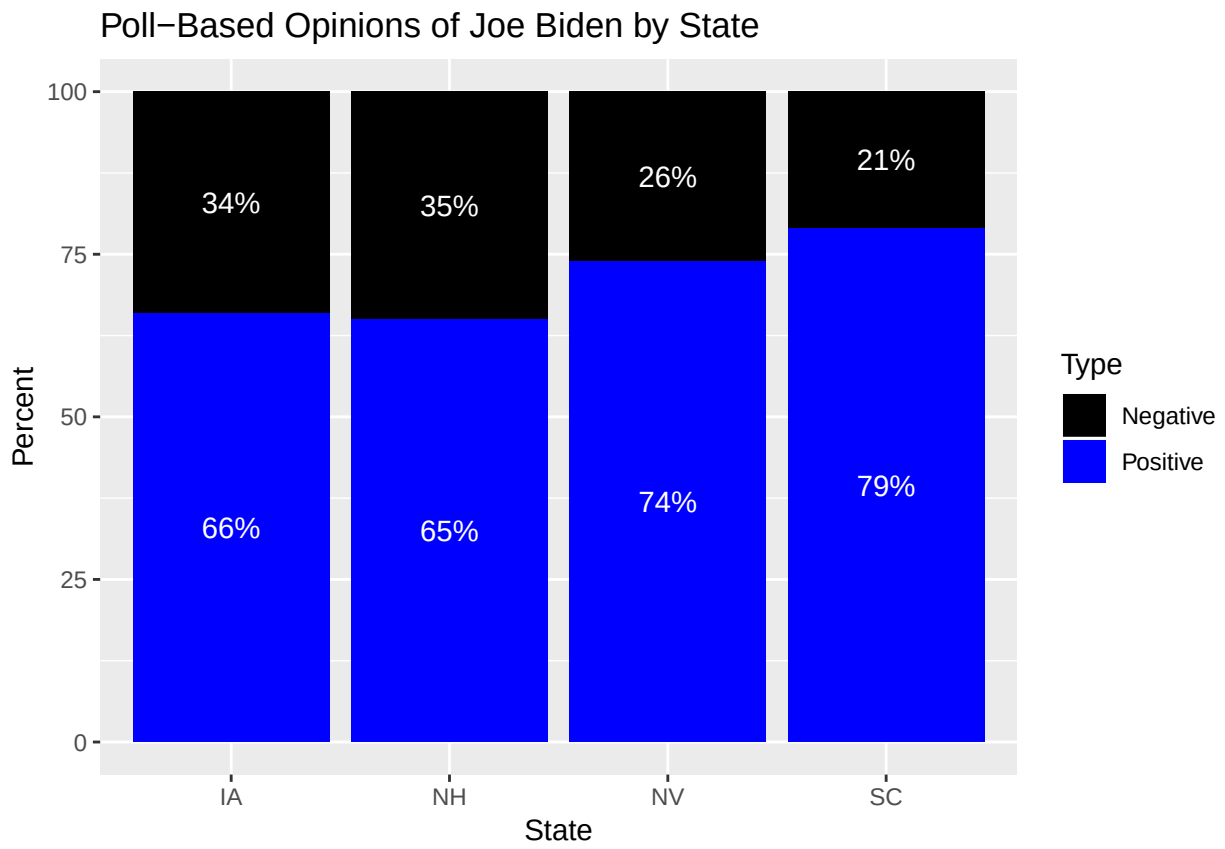

```
#Pete Buttigieg
PB <-
  pp %>%
  filter(Candidate == "PB")

#text positioning
PB <- ddply(PB, .(State), transform, pos = cumsum(Percent) - (0.5 * Percent))
#create graph
plot_PB <- ggplot() +
  geom_bar(aes(y = Percent, x = State, fill = Type), data = PB, stat="identity") +
  scale_fill_manual(values=c("black", "blue")) +
  geom_text(data=PB, aes(x = State, y = pos, label = paste0(Percent,"%")), size=4, color="white") +
  ggtitle("Poll-Based Opinions of Pete Buttigieg by State")
plot_PB
```




```
#Joe Biden
JB <-
  pp %>%
  filter(Candidate == "JB")

#text positioning
JB <- ddply(JB, .(State), transform, pos = cumsum(Percent) - (0.5 * Percent))
#create graph
plot_JB <- ggplot() +
  geom_bar(aes(y = Percent, x = State, fill = Type), data = JB, stat="identity") +
  scale_fill_manual(values=c("black", "blue")) +
  geom_text(data=JB, aes(x = State, y = pos, label = paste0(Percent,"%")), size=4, color="white") +
  ggtitle("Poll-Based Opinions of Joe Biden by State")
plot_JB
```



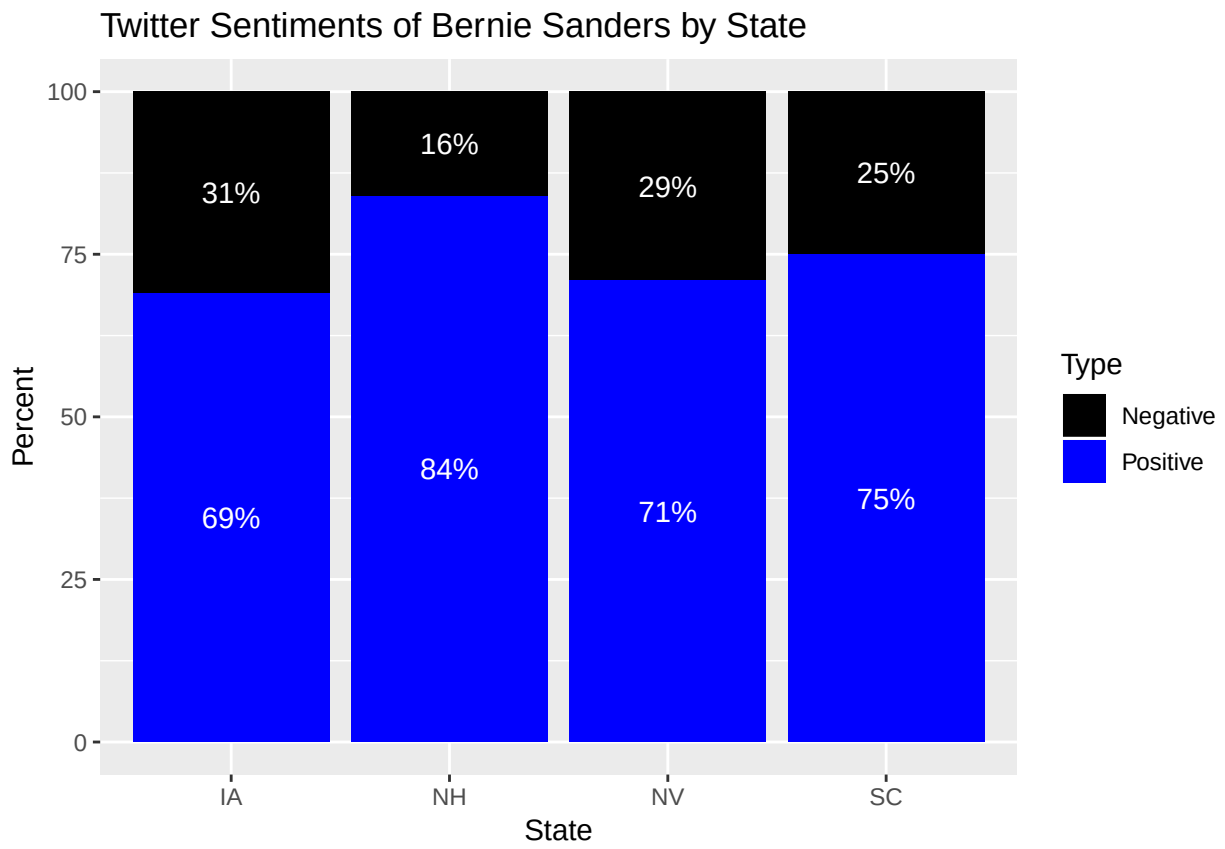
Similarly, below we load the data for graphing the Twitter sentiments.

```
sp <- read.csv("sentiments_prepped.csv")
sp <-
  sp %>%
  mutate(Percent = round(Percent*100))
```

Now we graph the Twitter sentiments for each candidate by state.

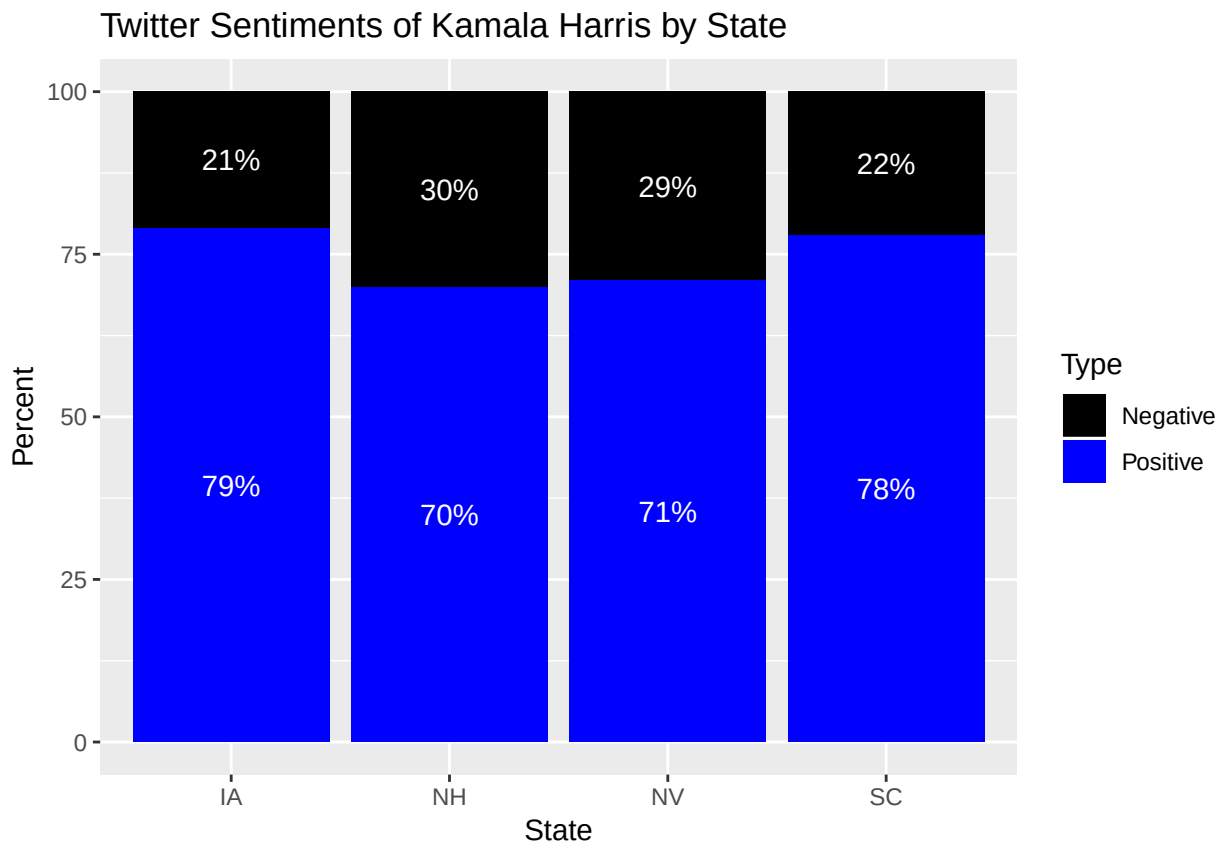
```
#Bernie Sanders
BS <-
  sp %>%
  filter(Candidate == "BS")

#text positioning
BS <- ddply(BS, .(State), transform, pos = cumsum(Percent) - (0.5 * Percent))
#create graph
plot_BS <- ggplot() +
  geom_bar(aes(y = Percent, x = State, fill = Type), data = BS, stat="identity") +
  scale_fill_manual(values=c("black", "blue")) +
  geom_text(data=BS, aes(x = State, y = pos, label = paste0(Percent,"%")), size=4, color="white") +
  ggtitle("Twitter Sentiments of Bernie Sanders by State")
plot_BS
```



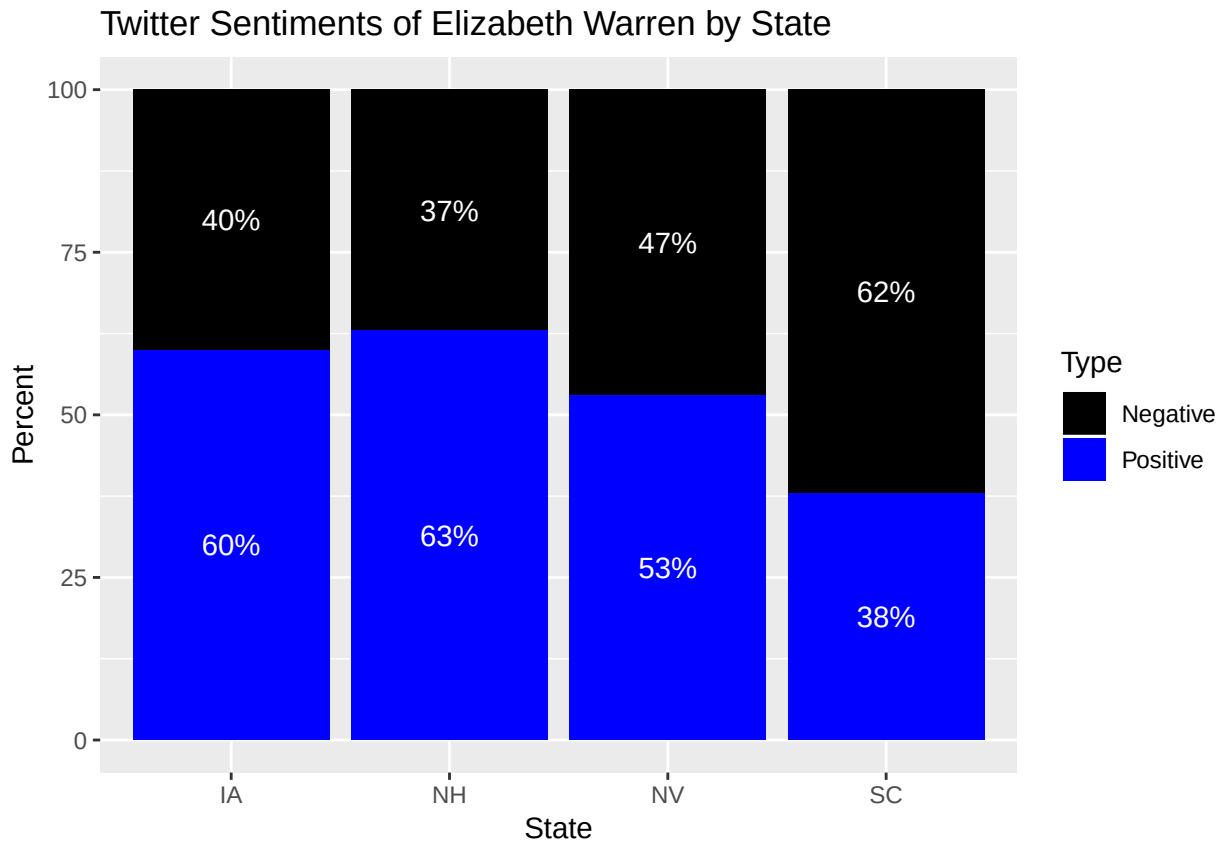

```
#Kamala Harris
KH <-
  sp %>%
  filter(Candidate == "KH")

#text positioning
KH <- ddply(KH, .(State), transform, pos = cumsum(Percent) - (0.5 * Percent))
#create graph
plot_KH <- ggplot() +
  geom_bar(aes(y = Percent, x = State, fill = Type), data = KH, stat="identity") +
  scale_fill_manual(values=c("black", "blue")) +
  geom_text(data=KH, aes(x = State, y = pos, label = paste0(Percent,"%")), size=4, color="white") +
  ggtitle("Twitter Sentiments of Kamala Harris by State")
plot_KH
```



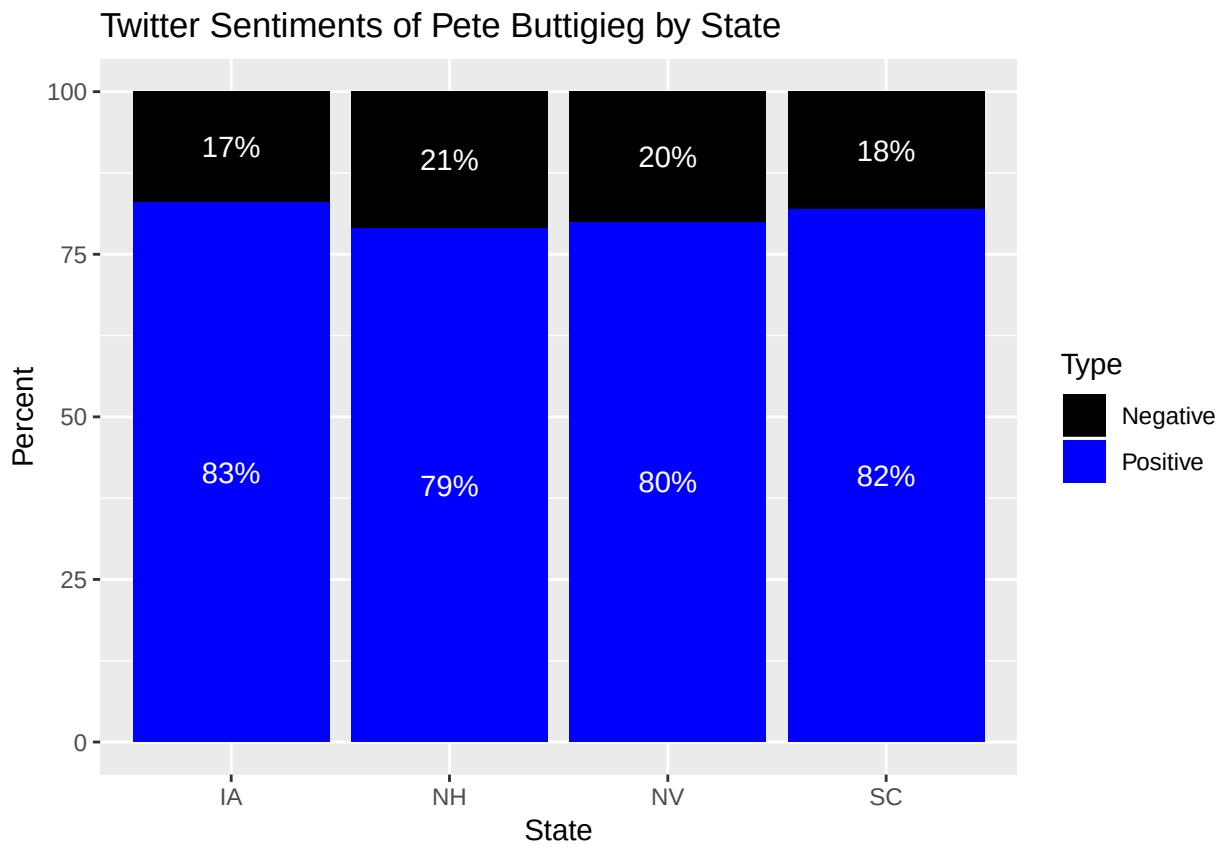

```
#Elizabeth Warren
EW <-
  sp %>%
  filter(Candidate == "EW")

#text positioning
EW <- ddply(EW, .(State), transform, pos = cumsum(Percent) - (0.5 * Percent))
#create graph
plot_EW <- ggplot() +
  geom_bar(aes(y = Percent, x = State, fill = Type), data = EW, stat="identity") +
  scale_fill_manual(values=c("black", "blue")) +
  geom_text(data=EW, aes(x = State, y = pos, label = paste0(Percent,"%")), size=4, color="white") +
  ggtitle("Twitter Sentiments of Elizabeth Warren by State")
plot_EW
```



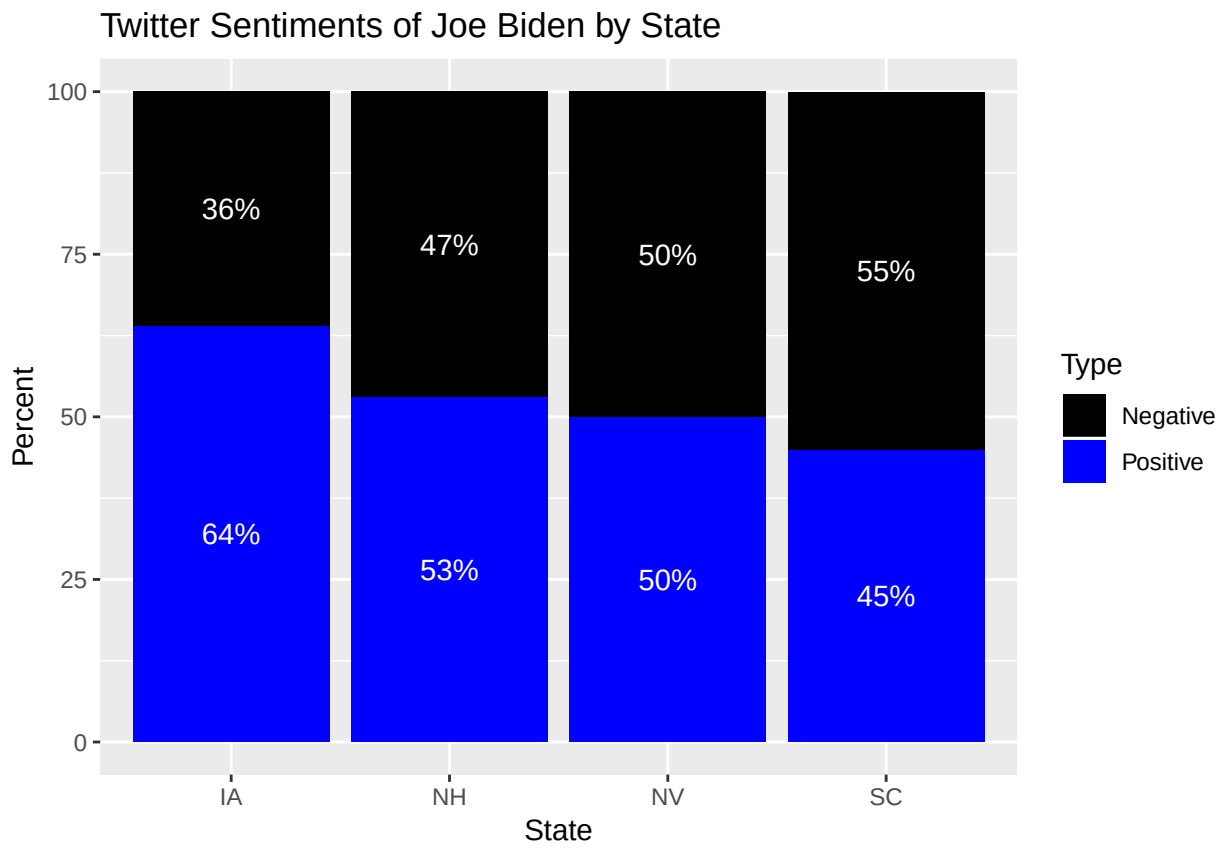

```
#Pete Buttigieg
PB <-
  sp %>%
  filter(Candidate == "PB")

#text positioning
PB <- ddply(PB, .(State), transform, pos = cumsum(Percent) - (0.5 * Percent))
#create graph
plot_PB <- ggplot() +
  geom_bar(aes(y = Percent, x = State, fill = Type), data = PB, stat="identity") +
  scale_fill_manual(values=c("black", "blue")) +
  geom_text(data=PB, aes(x = State, y = pos, label = paste0(Percent,"%")), size=4, color="white") +
  ggtitle("Twitter Sentiments of Pete Buttigieg by State")
plot_PB
```




```
#Joe Biden
JB <-
  sp %>%
  filter(Candidate == "JB")

#text positioning
JB <- ddply(JB, .(State), transform, pos = cumsum(Percent) - (0.5 * Percent))
#create graph
plot_JB <- ggplot() +
  geom_bar(aes(y = Percent, x = State, fill = Type), data = JB, stat="identity") +
  scale_fill_manual(values=c("black", "blue")) +
  geom_text(data=JB, aes(x = State, y = pos, label = paste0(Percent,"%")), size=4, color="white") +
  ggtitle("Twitter Sentiments of Joe Biden by State")
plot_JB
```



In addition, we run a regression in an attempt to see whether the Twitter sentiments correlate with poll outcomes and could be used for prediction. Below, the data is prepared for this purpose, deriving sentiment and poll favorability ratios for comparison.

```
df_polls<-read.csv("polls.csv")
df_twitter<-read.csv("twitter.csv")
```

```
df_polls$p_ratio<-df_polls$Positive/df_polls$Negative
df_twitter$t_ratio<-df_twitter$Positive/df_twitter$Negative
drops <- c("Positive", "Negative")
df_polls<-df_polls[ , !(names(df_polls) %in% drops)]
df_twitter<-df_twitter[ , !(names(df_twitter) %in% drops)]
pred <- merge(x = df_polls, y = df_twitter, by=c("State","Candidate"))
```

Using regression and not supervised machine learning. Not nearly enough data. As you can see there is not good predictive power trying to directly predict the canidate sentiments.

```
regress <- lm(p_ratio ~ t_ratio + State + Candidate, data=pred)
summary(regress)
```

```
##
## Call:
## lm(formula = p_ratio ~ t_ratio + State + Candidate, data = pred)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -1.4021 -0.5055 -0.1528  0.2443  1.5938
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.80105    1.52746   1.179    0.263
## t_ratio       0.18286    0.38965   0.469    0.648
## StateNH      -0.01922    0.69532  -0.028    0.978
## StateNV       0.70110    0.74030   0.947    0.364
## StateSC      -0.21006    0.71039  -0.296    0.773
## Candidate EW  1.11303    1.10656   1.006    0.336
## Candidate JB  0.46827    1.11933   0.418    0.684
## Candidate KH -0.58945    0.78237  -0.753    0.467
## Candidate PB  0.20801    0.87569   0.238    0.817
##
## Residual standard error: 1.099 on 11 degrees of freedom
## Multiple R-squared:  0.3258, Adjusted R-squared:  -0.1646
## F-statistic: 0.6644 on 8 and 11 DF,  p-value: 0.7131
```

6. Discussion

Twitter results were not consistenly close to poll results for any candidate or in any state. In several cases (for example, Elizabeth Warren in South Carolina), Tweets were drastically more negative than poll results indicated. Visual inspection of the graphs is confirmed by the regression, which did not show that twitter results were useful for predicting poll results. Hence, we would not advocate attempting predictions using the methods above. However, there are many aspects of predictions using social media data that we did not explore here, such as attempting weighting the twitter data by user characteristics; accounting for retweets or likes; and measuring sentiment with different methods. Tracking trends over longer periods of time would potentially be more useful, as Cody, Reagan, Dodds, and Danforth did with success previously (Cody 2016). A model that compared a lag of time ahead would be ideal and our sentiments are not exactly at the same point in time. A different study examined the ratio between negitive sentiment and volume as well as the log transformation of this (Bermingham 2011). We did not have enough volume to use this method either due to the limited amount of Twitter users that shared their location on their profile in each of these states.

In fact, the limitatations of this study are numerous, and using social media data for this type of application is still quite experimental. For this project, limitations divide mainly into general problems with political Tweets in general or problems with comparison to polls.

Analysing political Tweets as done in this project is challenging for multiple reasons. One problem with our approach is that we did not limit to unique users; one user could be included in the dataset multiple times, expressing the same sentiment repeatedly. It is likely there are many organizational Twitter accounts in the data that are likely to often tweet for or against a specific candidate. In fact, it has been shown that the top 10% of users produce 80% of Twitter content (Wojcik 2019). In addition, attempting to assign a sentiment value to a political tweet is a complicated process if only using individual word values. Political comments are often sarcastic, which our approach is not sophisticated enough to account for.

When comparing to the polls, the demographics of Twitter users does not match that of the general population; they are known to be younger, more educated, wealthier, and more likely to identify as Democratic (Mislove 2011). Additionally, for many candidate-state combinations, sample sizes were lacking; the smallest size was only 75 tweets, while the largest only 889. Also, although the data was collected in a period close to when polls were conducted, there was not a perfect overlap. Tweets were collected from November 23, 2019 - November 25, 2019, while the most recently available polls were conducted at various points in each state between September 22 and November 9 (Company 2019) (McKinley 2019) (CNN 2019b) (CNN 2019a). Clearly, polls that reflect opinions at a different time period are not easily comparable.

However, it may be possible that Tweets are indicitive of the trajectory of public opinion, rather than its momentary state. Comparing the current polls to those immediately previous, Biden and Warren both fell, while Buttigieg rose and Sanders was constant. Interestingly, the Twitter sentiments compared to polls showed lower ratings for Biden and Warren, higher ratings for Buttigieg, and similar ratings for Sanders. This suggests there may be some correlation between the state of Twitter opinions and the shifting of public opinion. If true, this information would be useful for candidates to know whether they are likely to be falling or rising before the next poll results become available, particularly after debates. Data over a longer period of time to correspond to multiple poll results are needed to verify this hypothesis. This is the avenue of research we believe most promising from this point.

7. References

Bermingham, & Smeaton, A. 2011. "On Using Twitter to Monitor Political Sentiment and Predict Election Results." Journal Article. https://www.aclweb.org/anthology/W11-3702.pdf.

CNN. 2019a. "CNN 2020 Nv Primary Poll." Poll. http://cdn.cnn.com/cnn/2019/images/09/28/rel1_nv.pdf.

———. 2019b. "CNN 2020 Sc Primary Poll." Poll. http://cdn.cnn.com/cnn/2019/images/09/28/rel1_sc.pdf.

Cody, Reagan, E. M. 2016. "Public Opinion Polling with Twitter." Journal Article. https://arxiv.org/pdf/1608.02024.pdf.

Company, Selzer &. 2019. "Des Moines Register/Cnn/Mediacom Iowa Poll." Poll. https://cdn.cnn.com/cnn/2019/images/11/16/rel4a_ia.pdf.

McKinley, Azem, S. P. 2019. "CNN 2020 Nh Primary Poll." Poll. https://scholars.unh.edu/cgi/viewcontent.cgi?article=1567&context=survey_center_polls.

Mislove, Lehmann, A. 2011. "Understanding the Demographics of Twitter Users." Journal Article. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2816/3234.

Wojcik, HUGHES, S. 2019. "Sizing up Twitter Users." News Article. https://www.pewresearch.org/internet/2019/04/24/sizing-up-twitter-users/.