Predicting Election Trends by Web Data

Deji Suolang, Kaan Cem Ketenci

13 December 2019

Contents

Introduction

This project intends to provide ways to use web-based organic data to supplement and provide preliminary estimates for more costly and time-consuming nationally representative surveys. We explore the Democratic Party primaries for the 2020 US presidential elections and focus on the CNN debate of democratic candidates held on 15th of October 2019. The aim is to use real-time web data to predict changes in voter support as indicated by benchmark surveys. In this project, three main sources of web data are used: Twitter API, Oddschecker webscraping data, and Google Trends API. Predictions from each of these web sources are then combined in a final linear regression model estimating changes of voting intensions in reference surveys.

Why is election data important?

Transparency of election processes is crucial in any open democratic society. Nowadays, campaigns develop detailed databases about citizens to guide the election campaigns and develop strategies (Nickerson & Rogers). Campaign data analysts develop models using this information to produce individual-level predictions about people's likelihoods of performing certain political behaviors, of supporting candidates and issues, and of changing their support conditional on being targeted with specific campaign interventions (Nickerson & Rogers).

Why do we need real-time web data?

A barrier to further research on election is the availability of real-time data, making it difficult to assess change as it occurs or predict election results before it begins. An emerging response to this challenge is the use of social media data, a relatively new form of "big data" that often provides localized, real-time information about people and communities on a variety of topics (McCourt, MFIA). Politicians can increasingly take advantage of real-time web data that enables them to take action promptly. Next, as political campaigns and election outcomes involve complex financial interests, many more organizations want to have quick real-time insights in order to better understand the political developments. Furthermore, by analyzing web data and producing models to supplement traditional costly surveys, we could develop innovative, cost-saving methods to detect potential shifts in voting intensions early.

Why do we use a political TV debate as a central time point?

Political opinion polling prior to an election is of crucial importance for news reporters, journalists, financial investors aiming to predict election outcomes, election campaign organizers, voters and candidates. But election campaigning is a very dynamic process. Based on upcoming news, developments in the economy, changes in political stances of opposing candidates, speeches given and performances in debates, there can often be a rapid shift in voter intentions between candidates. Debates are among some of the distinct moments for such shifts since voter opinions may be affected in a matter of just 2-3 hours of time. Most US election

debates are watched by over 10 million viewers across the US (Wu, USA Today), hence the debates are seen as a remarkable opportunity for candidates to outperform their opponents.

Due to the importance of debates, a significant amount of time and money is spent for debate preparation and analysis, and it equally becomes crucial to statistically measure debate performance. As a result, a lot of the opinion polling efforts take place right after a debate to measure the impacts of the debate on each candidate and to measure the shifts in support between candidates. By focusing our political analysis on debate days using web data, we aim to provide quick measurements about how each candidate has performed. Debate days are particularly useful for Twitter and Google search analysis due to the extraordinarily high volume of political tweets and searches. For instance, there were an estimated 53.2 million social media interactions related with the October 2016 US presidential debate by 16.9 million distinct Facebook and Twitter users, according to a Nielson study. Lastly, the debate dates and times are scheduled in advance and occur during a very limited period of time (at most 3 hours), therefore making the data collection process simpler and easier to analyze.

Who is tweeting politics and what are the limitations?

Twitter is not a nationally representative platform at all. Those with limited internet access are mostly not covered, and younger people use Twitter much more frequently than older people (Pew Research Center). It should also be noted that those people who tweet about politics is not representative of general Twitter users either. According to a 2019 study by Pew Research Center, a very small share of US adults on Twitter produce a great majority of public tweets of political content, where an estimated 97% of all political tweets in the US were tweeted by only around 10% of the US Twitter users. People who tweet about politics tend to have much more strongly formed opinions and may represent more extreme views at higher rates. Particularly, those who are not sure about which candidate to support are less likely to tweet about it. Pew Research Center study also indicates that Twitter is more commonly used by liberal-inclining voters than conservative-inclining voters relative to the rates in the general public. Last but not least, those who are not eligible to vote (e.g. non-citizens or those below age 18) may also post tweets about the political debates and may therefore be included in the sampled data even though they should not be covered for estimates of eligible voters' voting intentions.

Data

This study uses three primary data sources: Twitter API, Oddschecker webscraping data, and Google Trends API. It can be viewed as an experiment using these three sources of web data to predict potential national voter support for presidential candidates.

1. "Listening" Tweets:

Elections play crucial role in all democracies and social media is an important aspect in this process. Presently, political parties increasingly rely on social media platforms like Twitter and Facebook for political communication. The use of social media in political marketing campaigns has grown dramatically over the past few years. It is also expected to become even more critical to future political campaigns, as it creates two-way communication and engagement that stimulates and fosters candidates' relationships with their supporters. Furthermore, Twitter has become a platform where millions of people across the US actively share their real-time feelings and express their opinions about a variety of matters of interest to the general public. In this project, we are interested in political tweets and we particularly focus on tweets that discuss 12 democratic candidates who have qualified for democratic presidential primary debate that was held on October 15, 2019. Democratic presidential primary debate is an event that might be expected to affect the volume of tweets about these candidates and the sentiments expressed in them.

An important part of collecting tweets is to choose appropriate keywords and hashtags for tweet selection. First, we did an initial overview of the past tweets related to these 12 candidates. Besides their names and the campaign slogans, we found that people used hashtags such as "biden2020", "JulianForTheFuture", "AmyForAmerica" to express their supports to the candidates. We came up with our searching keywords

basket as follows: First names, last names, full names, the name with election year, campaign slogans, and popular hashtags, where we prepared approximately six keywords for each of the candidates.

We used the Twitter API to "listen" tweets and downloaded a corpus of tweets every day from October 10 to October 20. The collecting period included five days before the debate, the debate day, and five days after the debate.

2. Oddschecker Data:

The second source of online data used for this analysis is the Oddschecker data obtained by webscraping throughout the days of October 10th to October 20th. Oddschecker is a website that combines betting odds from a variety of major betting companies. The website contains many betting events from various categories. But our focus is the political betting event for the Democratic nominee for the 2020 US Presidential Elections, as can be seen at:

https://www.oddschecker.com/politics/us-politics/us-presidential-election-2020/democrat-candidate

Betting odds data obtained from Oddschecker is useful because it allows us to compute the approximate implied odds of winning the nomination for each candidate by taking the betting odds averages across all the available betting companies. The average odds obtained this way can be thought of as a reliable estimate of the most up-to-date nomination chances of each candidate based on the "public wisdom". Each betting company obtains their odds for the candidates by thousands of actual people who risk their money based on their perception of winning odds for specific candidates. Since this provides a risk and reward mechanism, there is direct incentive for people to bet based on their true understanding of odds. With the continuous participation of betting people, an approximate "market pricing" of candidates' odds are obtained.

The odds can then be converted into implied probabilities by simply using the formula for odds-to-probability conversion based on the definition of odds. A crucial advantage of Oddschecker is that it represents real-time odds of candidates. Whenever a new development happens, people would start to bet based on increased or decreased chances for particular candidates and the new implied odds would be reflected on the Oddschecker website within minutes. These odds represent both the current rankings of candidates and the betting people's predictions about the likely shifts in the future. Due to its nature of providing real-time data, Oddschecker can be used to produce early preliminary estimates of changes in voter support for candidates before any opinion polling data gets released several days after an event.

In this project, Oddschecker data is used to measure debate performances of each candidate by comparing odds prior to and after the October 15th debate. While no survey data regarding the debate gets published before 17th of October due to the time it takes to conduct and analyze surveys after the debate, the Oddschecker data is much quicker as there are evolving odds even during the course of the debate, and most of the shifts in odds get reflected in October 16th before any public survey outcome is known.

3. Google Trends API:

The final source of web data is obtained by Google API. We again focus on the same period of time covering five days before and after the debate day. By specifying the full names of each candidate, Google API provides the relative counts of Google searches for all the candidates. Relative increases or decreases in Google searches may be indicative of rising or falling popularity of candidates which may in turn be indicative of their voter support.

One particular limitation of using the Google API is that it cannot distinguish between "positive" or "negative" searches of candidates. For instance, if a candidate made a critical political gaffe during the debate, many people may search his or her name to make fun of or to criticize that candidate, but this would not be reflective of a rising voter support. Similarly, as discussed later in the analysis of Twitter counts, the heart attack by Bernie Sanders in early October may have likely resulted in tremendous increases of Google searches of him to learn more about the news regarding his health conditions, but this should once again not be considered as a boost in Bernie Sanders' support.

Method

1. Twitter API:

We have streamed tweets from October 10th to October 20th, and saved Twitter data for each day into separate JSON files. However, the JSON files we obtained were more than 10 GB for most days and we faced an error stating that the JSON file is too large to read them into R. After trying numerous approaches, we finally solved this problem with the help of an online file splitting tool: https://pinetools.com/split-filesand. In the splitting processes, we choose the number of splitted files to be approximately original file size divided by 100 MB. This guarantees that each splitted piece is around 100 MB. This is a sufficiently small size to be read into R. Out of all the splitted pieces, we sample them by a proportionate systematic sampling approach. For example, if there are 100 splitted pieces for a 10GB file, take the splitted pieces indicated by .000, .010, .020, .030, .040, .050, .060, .070, .080 and .090; if there are 200 splitted pieces for a 20GB file, download 20 sampled pieces chosen in a similar way. Systematic sampling ensures that the obtained sample of tweets are spread across the full day and do not just come from a limited time of the day. After dowloading the sampled files, we replace the "." in the file name with _ to ensure the computer does not treat the section after the "." as a file extension name. We also change the last digits in the file name so that 000, 012, 024, 036, 048, 060, 072, 084, 096 and 108 becomes 1, 2, 3, 4, 5, 6, 7, 8, 9, 10. This allows consistency in the R code. At the end of the file name, we add ".csv" to convert the undefined file into a csv file. The code of the above procedures are demonstrated in a separate file "Read Large JSON file.Rmd".

The csv files are read into R by readLines function. As the original dataset contains a long string of information related with each tweet, it is necessary to extract only the important pieces of information in a systematic way. The gsub() function is used to select the right components of the tweets and a dataset containing four columns of ID, tweet_text, tweet_location and tweet_time is obtained. Next, data cleaning includes removing punctuation, numbers, html-links and unnecessary spaces in tweet content, converting tweet time from Greenwich Time to Eastern Standard Time, and converting the location information into standard state names. For the purposes of comparing the tweets for different candidates, grep() function is used to categorize the collected tweets into different candidates based on their related keywords. We exclude tweets that do not seem to concern any of the 12 candidates, and obtain 12 groups of tweets where each group corresponds to one specific candidate.

2. Oddschecker webscraping:

To read Oddschecker data into R, we use the gambleR and betScrapeR packages. With the available oddschecker function in these libraries, we obtain betting odds data every day from over 20 companies for all the potential and actual democratic candidates including each of the 12 candidates participating in the October 15th debate. As rapid changes in some odds is expected during and shortly after the debate, we collect Oddschecker data more frequently during the 24 hours following the start of the debate. In addition to the daily data we import from Oddschecker, we record data at 8 pm on October 15th which is the time the debate begins, and further at 2 am on October 16th which is three hours after the end of the debate, and at 7 am and 7 pm on October 16th which corresponds to the morning and the evening following the debate night. By analyzing the shift in odds during this period, we are able to estimate the "winners" and "losers" of the debate and make conclusions about which candidates are more likely to observe an increase or decrease in their voter support in the upcoming opinion polls.

One important point to note is that the available data is in odds format which requires care to convert into final probabilities. First, after calculating the average odds for each candidate across all the betting companies, the definition of odds is followed to obtain unscaled probabilities by the formula: probability = $1/(\text{odds} + 1)$. For example, if the average odds for a candidate appears to be 4.0, this means that anyone who bets 1 dollar on this candidate will win 4 dollars more if this candidate actually wins the democratic nomination, and will lose that 1 dollar if this candidate does not win. This corresponds to a probability of $1/(4.0 + 1) = 20\%$. After the probabilities for all candidates are computed, the next step is to normalize these probability estimates so that the sum of the probabilities is exactly 1. This is necessary because the sum of the probabilities obtained this way usually slightly exceeds 1 because of the profit margins of betting companies included in the betting odds. For example, if a betting company has a 20% profit margin in its

betting odds, the available rewards would be shrunk by 20%. i.e. if the betting market valuations imply an odds of 5.0 for a candidate, the betting company would display an available odds of 4.0 so that 4 dollars is earned by the winners of the bet and the 1 dollar difference between 5.0 and 4.0 remains as the profit margin of the company. In this example, a displayed betting odds of 4.0 would actually correspond to a "scaled probability" of $1/(5.0 + 1) = 16.7\%$. In this project, this task of scaling is achieved by dividing each probability by the sum of the 12 probabilities to get normalized final probabilities adding up to 1.

3. Google Trends API:

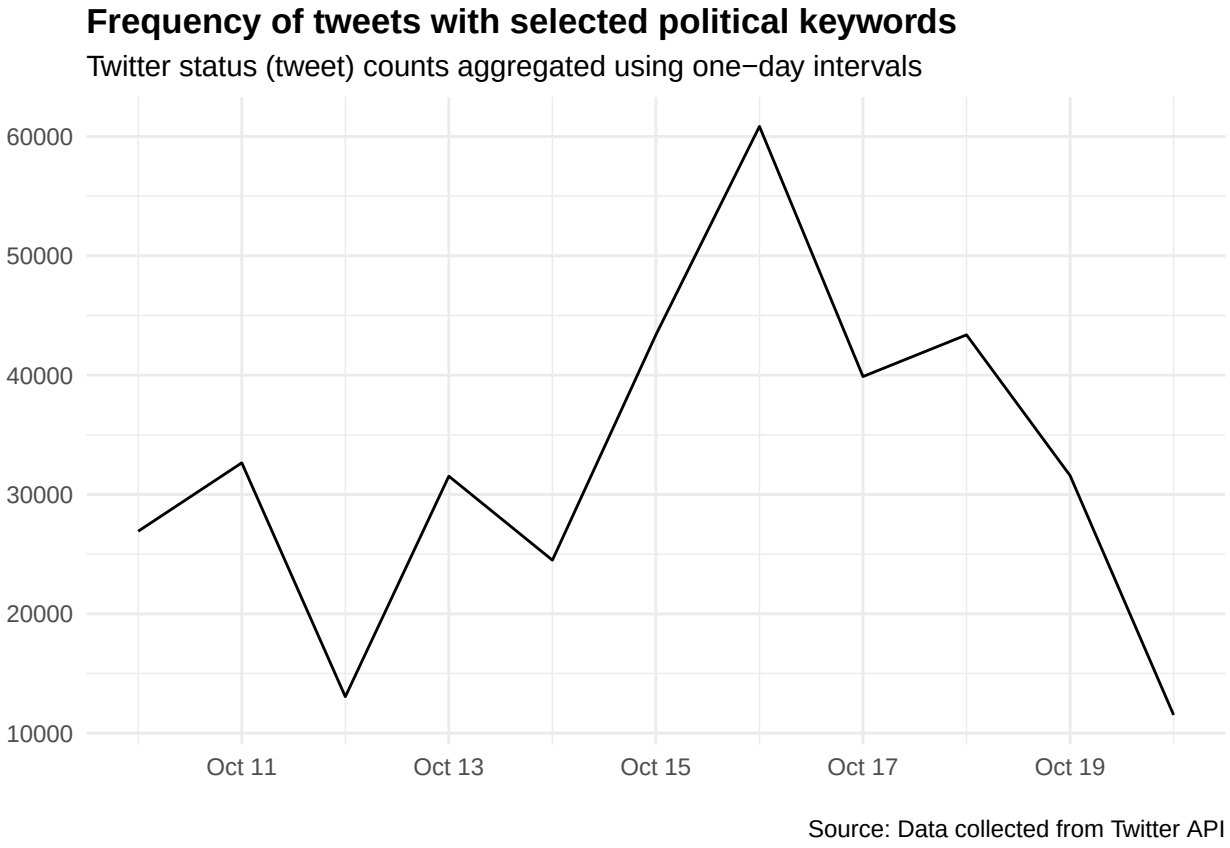
Lastly, results from the Google trends API is obtained by using the gtrendsR package in R. This package allows the time interval and location of searches to be specified as desired, and produces line charts of daily or hourly frequency of Google searches for the given keywords. One particular restriction of this package is that the plots do not provide the actual counts of Google searches. Instead, it only provides a scaled plot where the highest frequency of searches of a keyword at a given time is matched to 100 and all other counts are scaled as a proportion to this highest frequency of searches. Another restriction is that each plot displays separate trends for at most 5 different keywords. Considering that we have 12 candidates, this restriction prevents us from displaying all the candidates in a single plot. To overcome this issue and to ensure the same scaling factor is used for each of our Google trends plots for comparability, we test different groupings of candidates to first detect the candidate that reaches the highest frequency of Google searches in the given time period. We then include that candidate together with up to 4 other candidates for each plot and produce 3 separate plots that can jointly be viewed to compare trends for all the 12 candidates.

We produce different plots for both hourly Google search frequencies and for daily Google search frequencies. The daily plot may be more suitable for statistical analysis by providing direct comparison of the differences in searches between October 15th and October 16th to indicate the potential effects of the debate night. On the other hand, the hourly plot may be more suitable as a visualization of rapid fluctuations in the number of Google searches in short time intervals as a result of the debate performances.

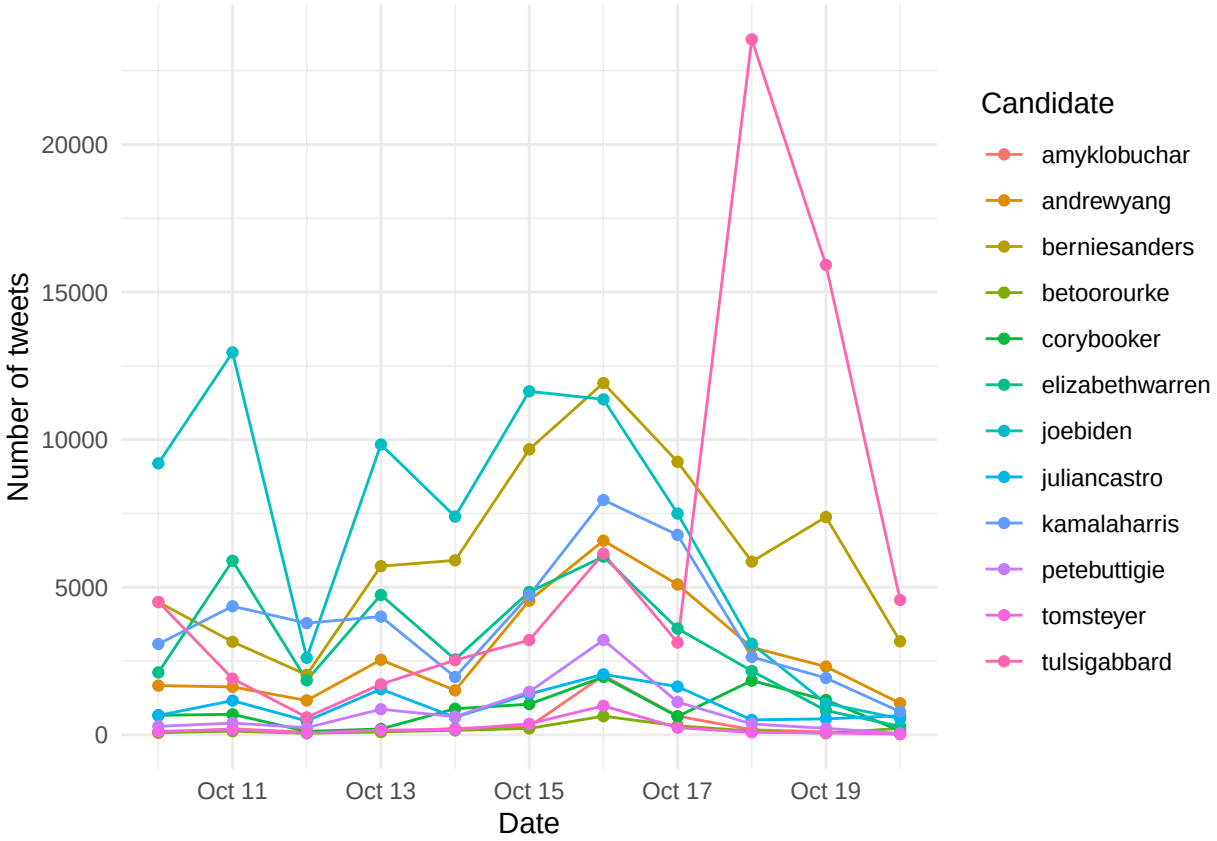
Analysis

1. Twitter API:

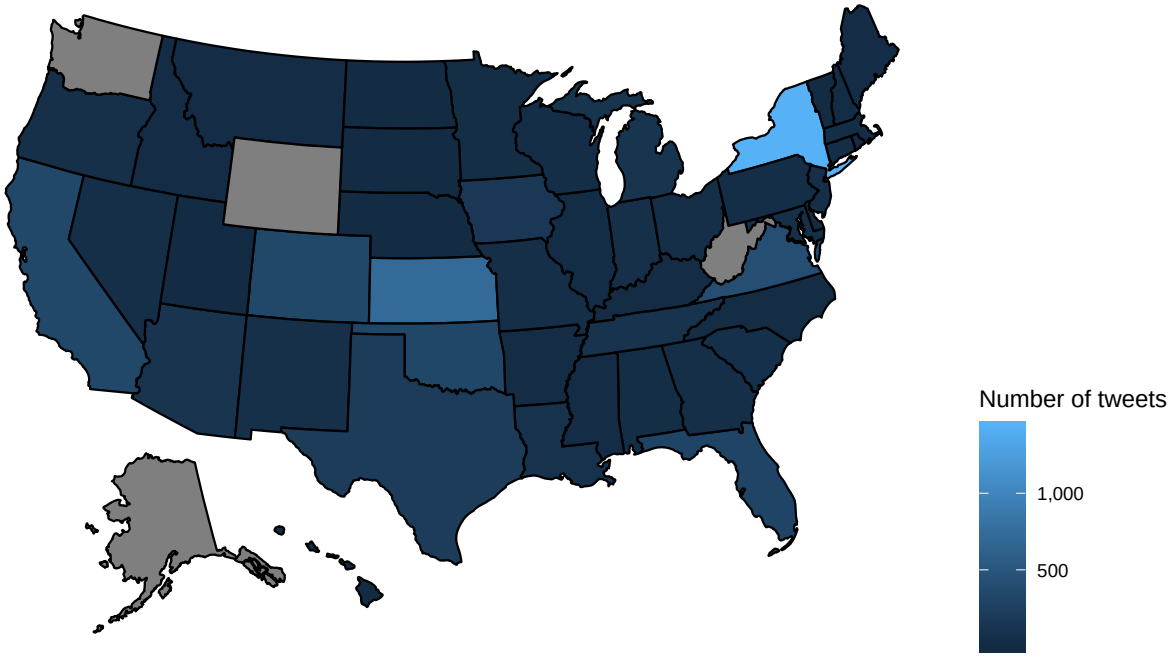
Considering the temporal characteristics of the tweets containing keywords, we graph the number of tweets by time period. It is observed that the number of tweets in our sample increased since October 14th and reached a peak at about 60000 pieces of related tweets on October 16th, then gradually decreased over time. Note that the tweets we used are proportionately sampled for each day, therefore, the sample should reflect a similar trend as the actual full set of tweets. This pattern in tweet counts is intuitive to understand since the targeted presidential debated started on 8:00 pm (ET) on October 15th and ended around 11:00 pm (ET), and many people tended to discuss the 12 candidates from one day before the debate until a day after the debate.



After knowing how the trends of number of tweets changed over time, the above word cloud provides an overview of what tweets said. As we expected, most of the content we collected are highly correlated with our research topic. Hence, changes in the number of tweets and the public support of candidates may be correlated.



Next, we draw a plot to closely look into the trend of each candidate. The number of tweets related to Amy Klocuchar on October 18th and 19th are extremely large, we may treat this as an outlier and not make any direct judgement based on this observation. Overall, there are more tweets related to Joe Biden, Bernie Sanders, Kamala Harris, Cory Booker, Amy Klobuchar, and Andrew Yang, compared to the number of tweets related to the other six candidates. Joe Biden had largest number of tweets until the debate day, and then Bernie Sanders became the most popular candidate after the debate. Kamala Harris had the third most related tweets. The trend of Andrew Yang showed a sharp increase right before the debate, then its decrease was much slower than other candidates. Andrew Yang had second largest number of tweets since October 19th and onwards.



The above map indicates the number of tweets that came from different states. As it covers most of the states in the U.S, the data we collected can be seen as a data with a wide U.S. national coverage.

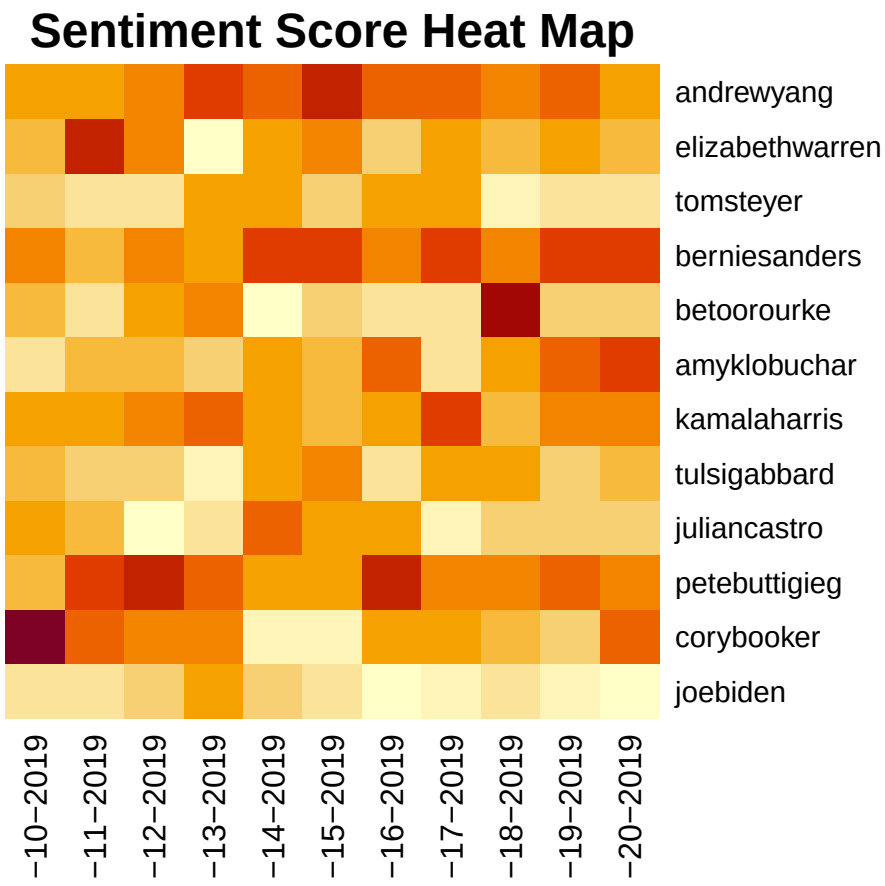
More tweets may not necessarily mean there are more support of candidates. For example, Bernie Sanders had a heart attack in early October, and then there were more tweets and an increased online search volume about him. Obviously, in such a case, having more tweets is not evidence of increased support. Therefore, instead of focusing on the tweet counts, we use sentiment analysis for interpreting tweets that we collected for each candidate. There are three main dictionaries that can be used to score the sentiment, Harvard-IV General Inquirer, Henry's finance specific dictionary and Loughran-McDonald financial dictionary, as well as QDAP. Harvard GI refers to the psychological Harvard-IV dictionary, it assigns to words among others a sentiment in a scale going from -1 to +1 depending on their connotation: positive or negative. This well-known dictionary was already used in other political research whereas the other three dictionaries were more frequently used for financial data. In the present study, Harvard-IV dictionary is considered as a good instrument used to analyze the sentiments of the tweets.

The goal is to compare the sentiments scores we obtained for 12 candidates and to see how these scores are changing over time in the eleven-day period from October 10th to October 20th. First, we create a 12*11 matrix, each row represents one candidate, and each column represents one day. When we analyze the sentiments of all the tweet texts, we encounter an error stating that we have exceeded the memory limits. Therefore, we sample a maximum of 2000 tweets for each candidate, and then compute the mean of the sentiment scores. A for loop is used to iterate through a block of statements.

```
sentiment_score_matrix <- matrix(0, nrow = 12, ncol = 11)
colnames(sentiment_score_matrix) <-
  c("10-10-2019", "10-11-2019","10-12-2019","10-13-2019","10-14-2019","10-15-2019",
    "10-16-2019","10-17-2019","10-18-2019","10-19-2019","10-20-2019")
rownames(sentiment_score_matrix) <-
  c("joebiden","corybooker","petebuttigieg","juliancastro","tulsigabbard","kamalaharris",
```

```
  "amyklobuchar","betoorourke","berniesanders","tomsteyer","elizabethwarren","andrewyang")
for(i in 1:12){
  for(j in 1:11){
    tweets_of_interest <- tweetdata_merged[(tweetdata_merged$candidateID == i) &
                          ((as.numeric(tweetdata_merged$date)) -
                          as.numeric(as.Date("10-9-2019", format="%m-%d-%Y")) == j),]
    sentiments = analyzeSentiment(as.character(tweets_of_interest
                          [1:(min(2000, nrow(tweets_of_interest))),]$tweet_text))
    sentiment_score_matrix[i,j] <- mean(sentiments$SentimentGI,na.rm=T)
  }
}
```

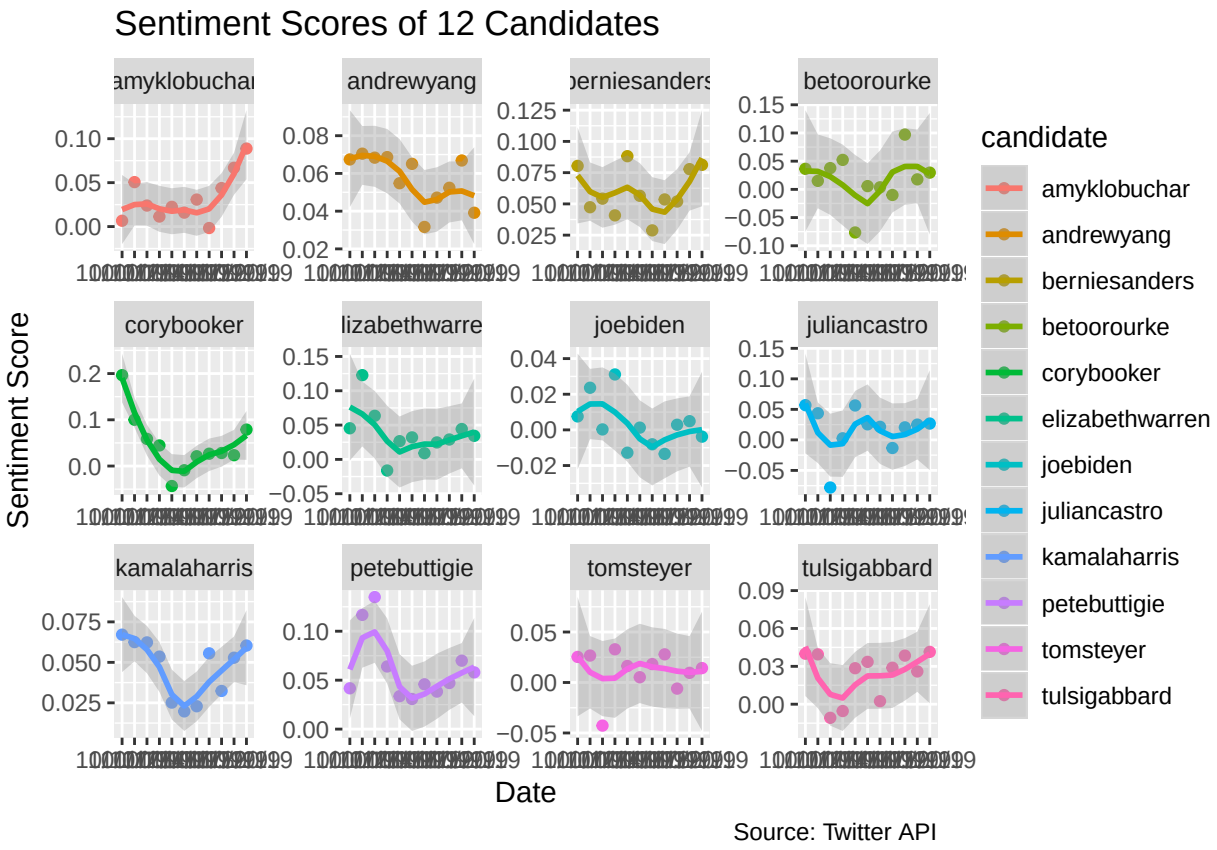
Heatmap is used to visualize the sentiment scores and the change in tweet sentiment over time by a spectrum of colors. A square in dark color represents high sentiment score and a square with light color represents low sentiment score. We can observe that people's sentiment about Bernie Sanders leans more towards positive after the debate. Andrew Yang and Pete Buttigieg have higher overall sentiment scores compared to the remaining candidates. The row for Joe Biden is in very light color, which means he had low to neutral sentiment scores over time.



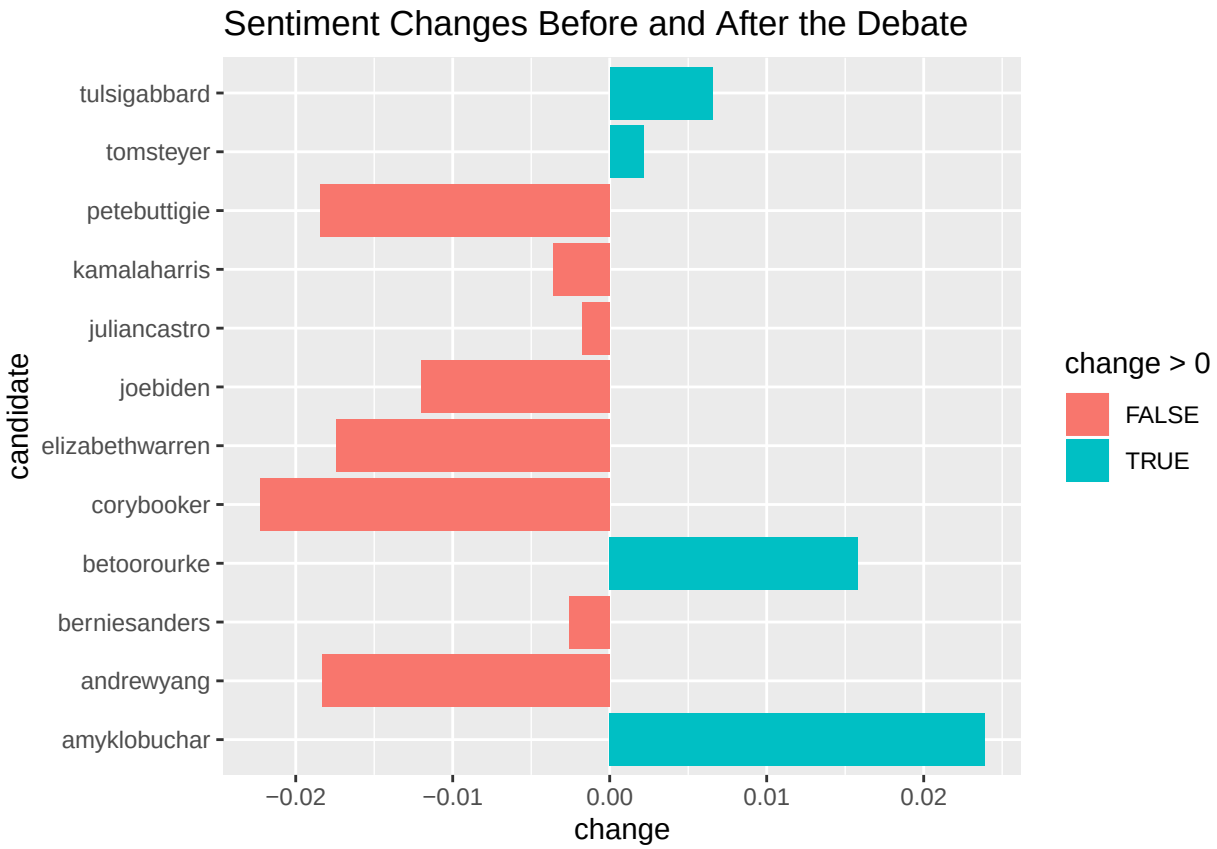
Next, we use ggplot() function to plot the line chart of Twitter sentiment scores, faceting is used to draw separate charts for each candidate. In terms of sentiment change, one interesting finding is that Beto O'Rourke, Cory Booker, Kamala Harris, and Pete Buttigieg have a U-shape changing trend, they gained their most negative tweets during the debate, and then gradually recovered to the original status. We could observe that there are increasing positive views on Amy Klobuchar, Bernie Sanders and Tulsi Gabbard after the debate. Elizabeth Warren, Joe Biden, Julian Castro, and Tom Steyer had approximately constant sentiment scores over time, whereas Andrew Yang's sentiment score decreased before the debate and then remained constant. In terms of the magnitude of the overall sentiment scores, we can conclude that Elizabeth Warren,

Pete Buttigieg, and Tulsi Gabbard had most positive scores among 12 candidates.

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



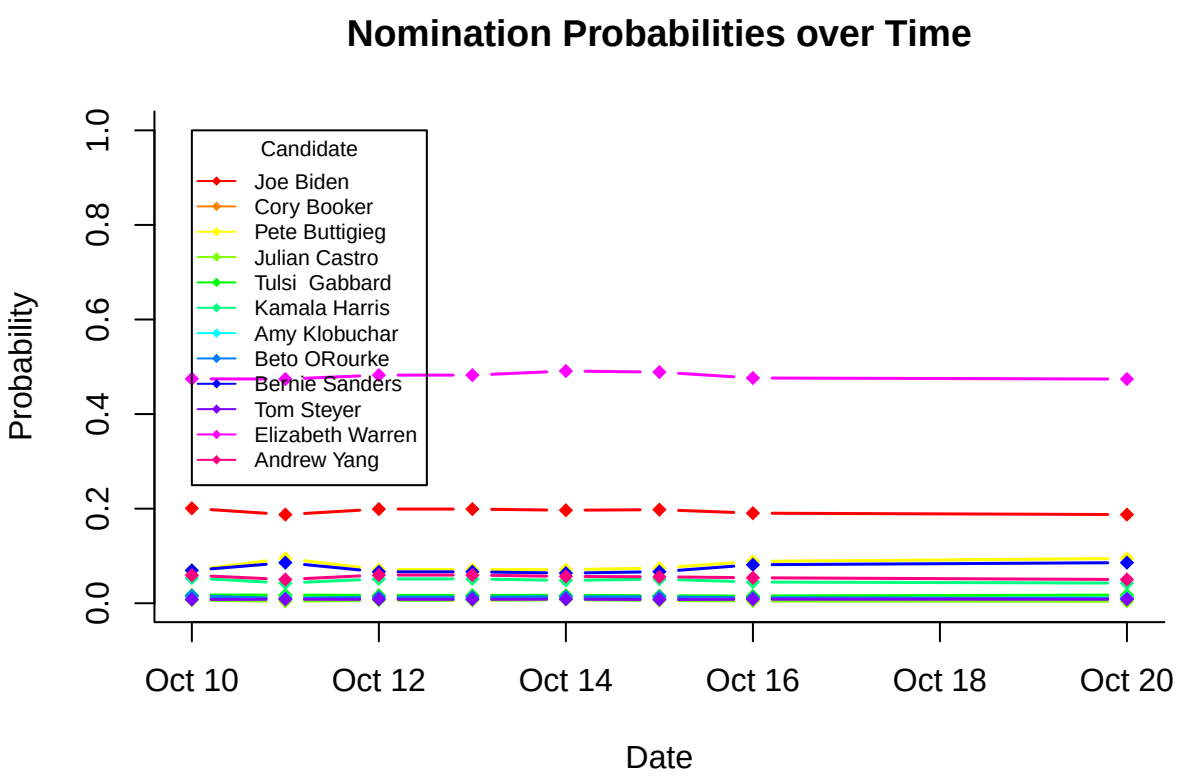
To understand how the debate affected people's sentiment about candidates, we compute the mean sentiment score of the first six days (before the debate), and the mean sentiment score of the last five days (after the debate) for each candidate. Then we focus on the difference of two mean scores. A positive change value means the sentiment for that candidate became more positive. The green bars of the following bar plot shows the positive change and red bars shows the negative change.



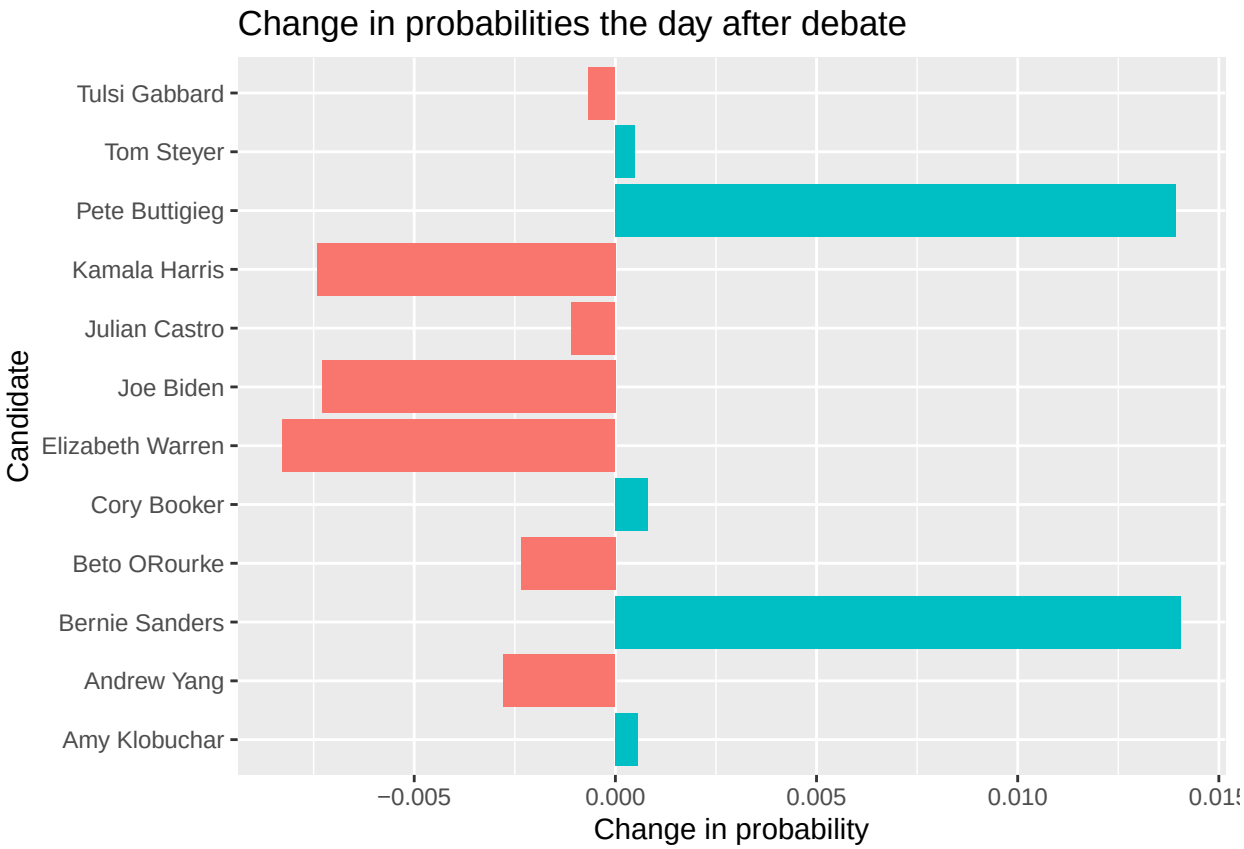
In conclusion, the above changes in tweet counts and Twitter sentiment analysis indicate mixed results. Amy Klobuchar has both an increase in the number of tweets and more favorable sentiment after the debate. Most candidates appear to have more negative sentiments after the debate. More data is needed to make direct conclusions using Twitter counts or sentiments about who performed well and who performed poorly in the debate.

2. Oddschecker webscraping:

The Oddschecker implied nomination probabilities of candidates over time is shown by the following line chart. As we focus on only a limited period of time from October 10th to October 20th, the probability differences between different candidates appear mostly stable. Elizabeth Warren has the highest probability at around 50%, Joe Biden has the second rank with probabilities around 20%, and he is followed by Bernie Sanders and Pete Buttigieg both at around 10% probabilities of winning the nomination.



While the overall comparison of probabilities across candidates do not appear to initially suggest clear patterns of the debate performance, taking a careful look at the changes over time between shortly before and shortly after the debate for each candidate can be indicative of the debate performance. The next bar plot represents these changes in probabilities for each candidate the day after the debate night. Green bars indicate an increase in probabilities and red bars indicate a decrease in probabilities shortly after the debate.



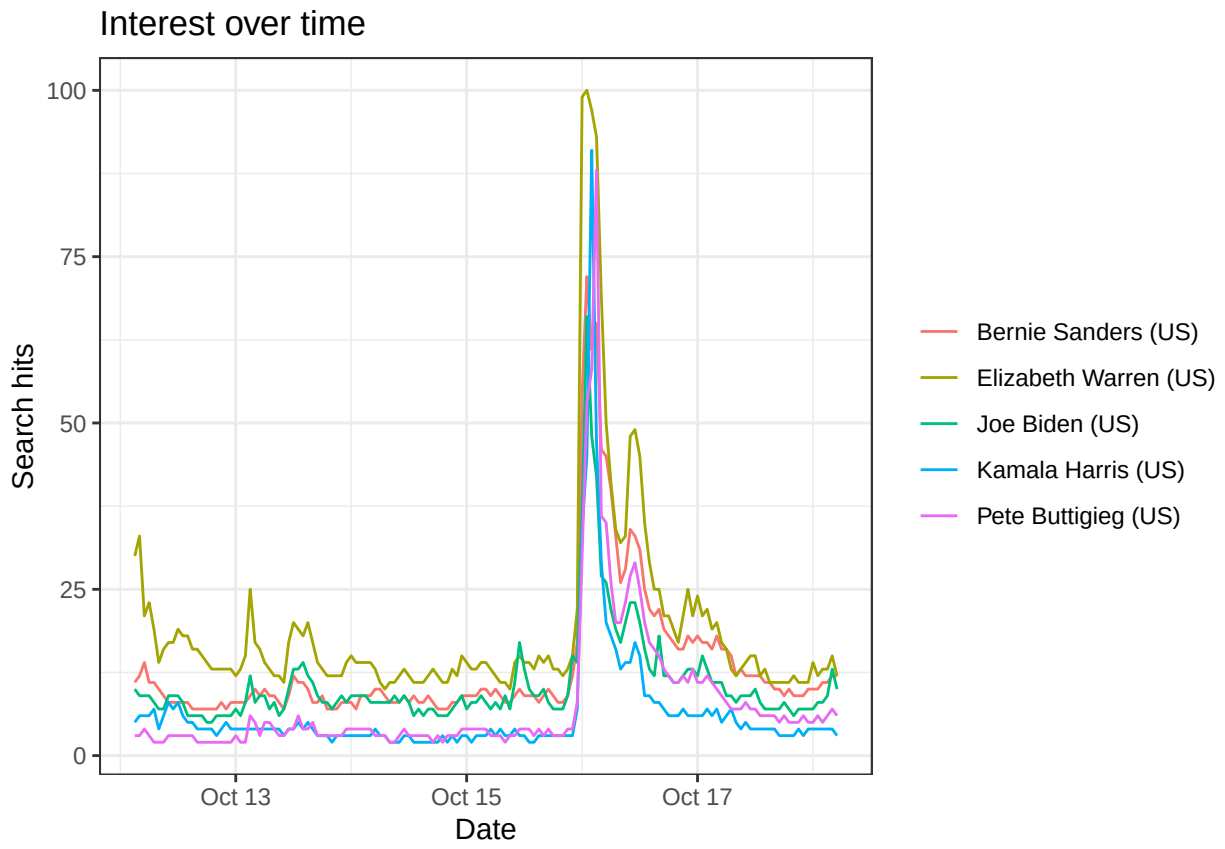
This comparison of Oddschecker data from 8pm on October 15th (the start of the CNN debate) and 7pm on October 16th (the evening following the debate night) is likely to reflect perceived debate performances of respondents. The implied probabilities for Elizabeth Warren, Joe Biden and Kamala Harris have each dropped by close to 1%, whereas the probabilities for Pete Buttigieg and Bernie Sanders have both increased by more than 1%.

These increases and decreases in implied probabilities of winning the nomination matched almost perfectly with political analyses by mainstream news organizations of debate performances of each candidate, as can be seen by comparison with the CNN editor analysis by Chris Cillizza. The political analysis article indicated Pete Buttigieg, Amy Klobuchar, Bernie Sanders and Andrew Yang as "winners" of the debate and Joe Biden, Kamala Harris, Elizabeth Warren and Tom Steyer as "losers" of the debate which is generally consistent with the Oddschecker finding. This justifies that the Oddschecker data can be a very good real-time indicator of dynamic shifts in voter support before any opinion poll result is released.

3. Google Trends API:

Lastly, we analyze the trends in Google searches of each candidate and the potential implications of these search frequencies on their voter support. When google search statistics between October 10 and October 20 is gathered for each candidate, there is a clear spike in Google searches on the night of October 15 (during the debate) and the next day. Also note that the time zone in these plots is given by UTC (Coordinated Universal Time) by default, whereas the Eastern Time (ET) in the US during October is defined as UTC-4. i.e. there is a 4-hour time zone difference. This means that the start of the debate which is 8 pm (ET) on October 15th is displayed as 12 am (UTC) on October 16th in the Google search plots.

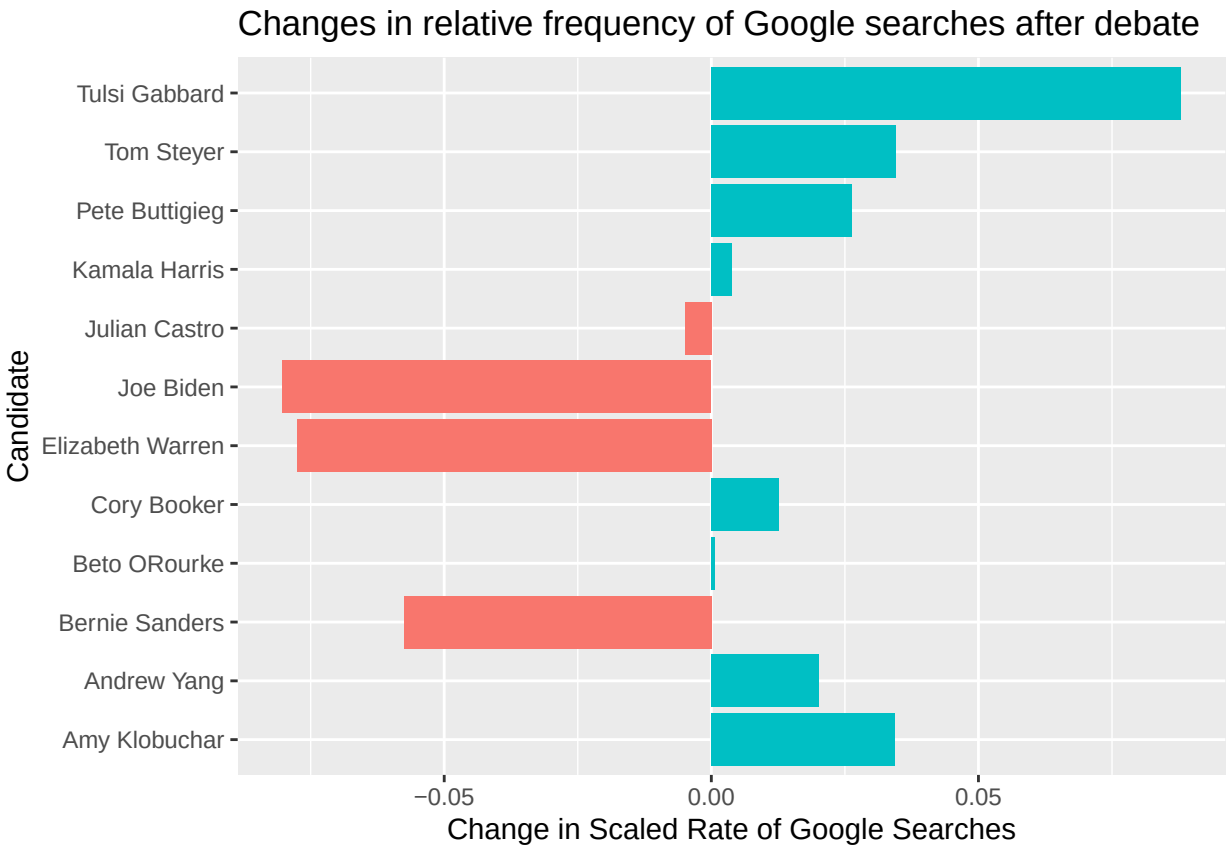
The following is the hourly frequency of Google searches for the top 5 leading candidates. Noting that gtrendsR package only allows up to 5 distinct searches to be displayed on a single plot, we cannot show all the candidates in one plot. Other Google trends plots for different groups of candidates, for hourly as well as daily frequency of Google searches, can be seen in the appendix.



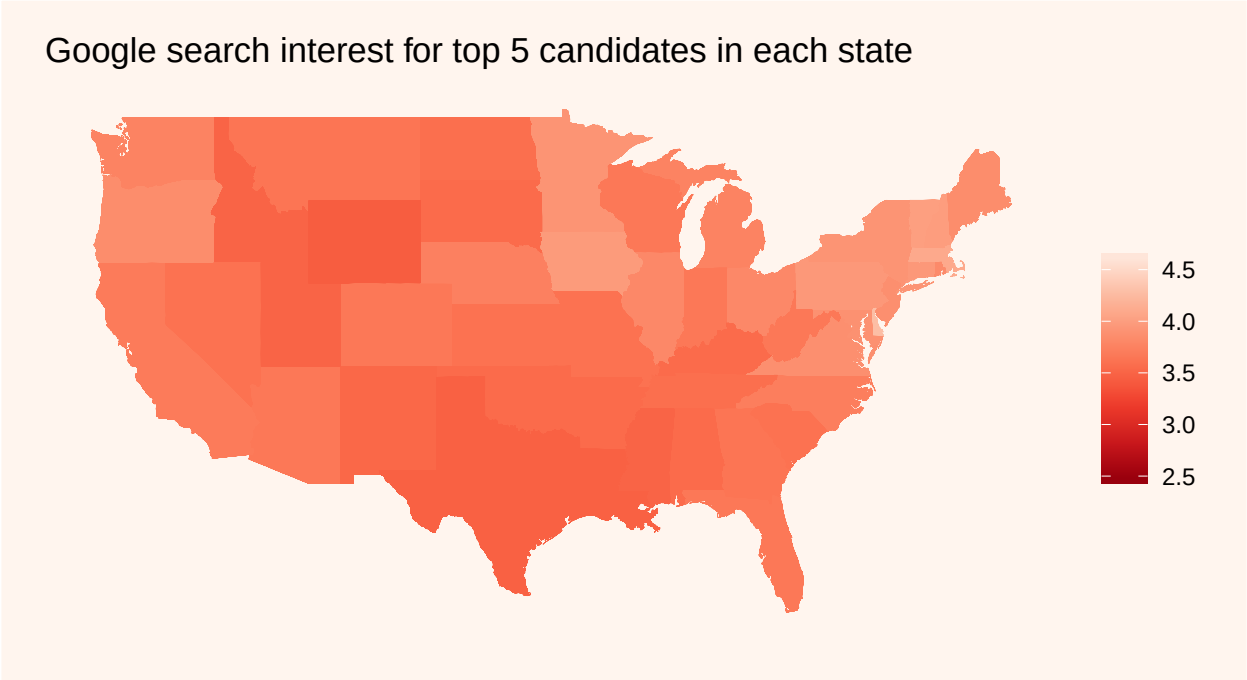
High relative volume of searches for a candidate may correlate with increasing voter support for that candidate. However, it is not immediately clear how to compare the debate effect between different candidates since all of the 12 candidates actually observed major increases in their Google search counts during and shortly after the debate. We have observed that the total number of combined searches across all the 12 candidates on October 16th was $535/111 = 4.82$ times higher than the total searches on October 15th. If we purely look at the differences in searches for each candidate, then we usually observe biggest increases for candidates who were already popularly searched on Google. For instance, Elizabeth Warren who was searched the most on October 15th at a rate of 25 units, saw an increase in searches to reach 79 units of searches on October 16th, which indicates a difference of 54. In contrast, Amy Klobuchar has 2 units of searches on October 15th and it rose to 28 units on October 16th, corresponding to a difference of 26. While the actual difference in search counts is higher for Elizabeth Warren, the search ratio for Amy Klobuchar is much higher at $28/2 = 14$ than that of Elizabeth Warren at $79/25 = 3.16$. This indicates that checking the pure differences in counts may be misleading as it will be disproportionate between candidates who were already popular and those who are only gaining popularity after the debate.

To overcome this issue, we compute the relative shares of each candidate's searches as a proportion of the total searches of all candidates on a given day. This calculation indicates that searches for Elizabeth Warren accounted for 22.5% of all searches of candidates' names on October 15th while it dropped to 14.8% of all searches on October 16th. In contrast, Amy Klobuchar's searches have increased from being 1.8% of all searches on October 15th to 5.2% of all searches on October 16th. We infer that this "scaled" measure of Google searches better reflects changes in candidates' relative popularity before and after the debate.

The bar plot shows these "scaled" differences in Google searches for each candidate. Green bars indicate a relative increase whereas red bars indicates a relative decrease. We observe that the relative shares of Google searches of the top 3 candidates Joe Biden, Elizabeth Warren and Bernie Sanders have all dropped, whereas the relative shares by Tulsi Gabbard, Tom Steyer, Amy Klobuchar and Pete Buttigieg have increased after the debate.



Lastly, the following visualization of the Google search frequencies across different states on the U.S. map shows that every state contributes towards the search counts, but states in the eastern coast (which are more populous and more democratic leaning states) have higher proportion of searches of the democratic candidates' names than those of the middle and southern U.S. states.



Discussion of the final model

Each of the three approaches presented in this paper could potentially be used to predict likely changes in U.S. voter support for each of the candidates. In this final section, we attempt to create a final model of estimated change in opinion poll numbers of candidates as a linear combination of the changes observed in Twitter, Oddschecker and Google trends data. We expect increasing sentiment scores of tweets, increasing frequency of scaled Google searches and increasing implied probabilities indicated by Oddschecker to each correlate positively with an increasing voter support for that candidate. One of the most reliable established tools to measure public support for candidates is through traditional surveys using nationally representative samples. Treating the results by Morning Consult opinion polls as our benchmark measure of US voter intentions, we analyze how well each of the three approaches in this paper perform at predicting these voting intensions before any opinion poll data is even released.

Morning Consult is a technology and media company focusing on survey research (www.morningconsult.com). There were some national survey results of voting intentions in democratic presidential primaries, both before and after the debate of October 15th. The first survey was conducted during October 7-13 using a large nationally representative sample of size 15,683. It was released on October 14th shortly before the debate. The second survey was conducted during October 16-20 again using a large national sample of 11,521 eligible voters and was released on October 21st. Although some questions of these surveys were particularly aimed at respondents from early voting states, the survey question of interest for our analysis asks for the voting intentions at a national level, where the outcome variable is defined as the candidate that is respondent's first choice of vote if an election were to be held that day. It should be noted that high quality survey data at state-level voting intensions is not available for most of the states, except for a few early voting states at this time of the election cycle. Hence, our analysis using our final model is conducted at national level.

Our final model is a linear model with 3 parameters of estimates of change by twitter sentiments, oddschecker

probabilities and scaled google searches. All of these three predictors are available at a national level, and they are used to predict the changes in the national-level benchmark Morning Consult surveys. Considering that data has been collected for only 1 debate period from 12 candidates, the degrees of freedom of this model is 12-3 = 9. This model can be substantially improved if further data is collected from other debate periods and from other election cycles.

```
model1 <- lm(Change_in_Morning_Consult_Poll ~ Change_in_Twitter_Sentiment +
          Change_in_Probabilities_by_Oddschecker + Change_in_Scaled_Google_Searches,
          data = combined_results)
summary(model1)
```

```
##
## Call:
## lm(formula = Change_in_Morning_Consult_Poll ~ Change_in_Twitter_Sentiment +
##     Change_in_Probabilities_by_Oddschecker + Change_in_Scaled_Google_Searches,
##     data = combined_results)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0096995 -0.0043244 -0.0003503  0.0034281  0.0118218
##
## Coefficients:
##                                       Estimate Std. Error t value
## (Intercept)                          -0.001591   0.002359  -0.675
## Change_in_Twitter_Sentiment          -0.189540   0.173958  -1.090
## Change_in_Probabilities_by_Oddschecker  0.157429   0.329425   0.478
## Change_in_Scaled_Google_Searches        0.122598   0.051648   2.374
##                                      Pr(>|t|)
## (Intercept)                             0.519
## Change_in_Twitter_Sentiment             0.308
## Change_in_Probabilities_by_Oddschecker    0.646
## Change_in_Scaled_Google_Searches          0.045 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.007808 on 8 degrees of freedom
## Multiple R-squared:  0.453,  Adjusted R-squared:  0.2479
## F-statistic: 2.209 on 3 and 8 DF,  p-value: 0.1647
```

The model output shows that the coefficient estimates of the variables of oddschecker probabilities and google searches are both positive. This indicates that whenever a candidate is observed to have higher implied probabilities by political betting companies in Oddschecker, it is also more likely that the upcoming opinion polls will produce results of higher voting intensions for that candidate. i.e. the oddschecker data from the night of the debate and the day following the debate can be used to generate preliminary estimates of which candidates are more likely to gain more support in ongoing opinion polls before those survey results are announced several days after the debate.

Similarly, the positive estimate for the scaled google searches parameter indicates that whenever a candidate is observed to have a more substantial increase in the proportion of Google searches of his/her name among all the candidates, it is more likely that he/she will also perform better in the ongoing national surveys of voter intensions. i.e. google search trends can similarly be used to generate preliminary estimates of changes in opinion poll numbers before the actual survey results are announced. It is also worth noting that the effect of the google searches parameter is statistically significant with a p-value less than 0.05. Thus, google trends is shown to significantly correlate with survey results and it is a much faster and cheaper tool than the expensive surveys for which the results become available in no less than a few days.

The results suggested by three different data sources share some common features that Bernie Sanders, Pete Buttigieg and Amy Klobuchar have increased chances to win the election after the debate. Based on the regression model, Twitter sentiment score does not show the same direction with the polling numbers when predicting the election result. The possible reasons can be related to uncertainty of auto-coded sentiment analysis. There are various dictionaries to compute the sentiment score and selection difference may produce different results. Auto-coding is a powerful tool, but the judgement is based only on the words due to the "bag of words" approach of dictionaries, and not the whole context and the overall meaning. For example, some tweets with metaphors or sarcastic/ironic tweets are very likely to be misjudged. Also, since there is a substantial number of re-tweets in the data, some of these may cause duplication and there is no readily available standard way to categorize such tweets.

Finally, this study leaves space open for further research. Since many people may not use Twitter or post any political tweets, and the proportion of public political tweets is almost entirely dominated by just around 10% of Twitter users as discussed in the introduction section, under-coverage is a very serious issue. A very large group of people is excluded from the Twitter study and the target population does not coincide with the sampling frame. In addition, there were some occasional internet disconnections during our Twitter streaming which may have affected the number and quality of tweets we collected. Furthermore, many tweets miss the location information, or the location is manually defined. If we have external sources to locate the tweet users, a state-level prediction could be very informative in terms of presidential elections. Despite the rich literature on using web data to predict election results, the computations involved and features of this kind of found data have still not been fully explored. Methods such as weight adjustments based on the target population may further improve the accuracy of the predictions by partially resolving issues of non-representativeness of the web samples.

References

Cillizza, Chris. "Chris Cillizza's Winners and Losers from the Fourth Democratic Debate." CNN, Cable News Network, 16 Oct. 2019, www.cnn.com/2019/10/15/politics/who-won-the-democratic-debate/index.html.

"Morning Consult Political Intelligence 10.14.19." Morning Consult, Oct. 2019.

"Morning Consult Political Intelligence 10.21.19." Morning Consult, Oct. 2019.

Nickerson, et al. "Political Campaigns and Big Data." Journal of Economic Perspectives, www.aeaweb.org/articles?id=10.1257%2Fjep.28.2.51.

Romano, Lois. "Obama's Data Advantage." POLITICO, 9 June 2012, www.politico.com/story/2012/06/obamas-data-advantage-077213.

"Small Share of U.S. Adults Produce Majority of Political Tweets." Pew Research Center for the People and the Press, 14 Nov. 2019, www.people-press.org/2019/10/23/national-politics-on-twitter-small-share-of-u-s-adults-produce-majority-of-tweets/.

"Social Media Mining: The Effects of Big Data In the Age of Social Media." Yale Law School, law.yale.edu/mfia/case-disclosed/social-media-mining-effects-big-data-age-social-media.

"Third Presidential Debate of 2016 Draws 71.6 Million Viewers." Nielsen, www.nielsen.com/us/en/insights/article/2016/third-presidential-debate-of-2016-draws-71-6-million-viewers/.

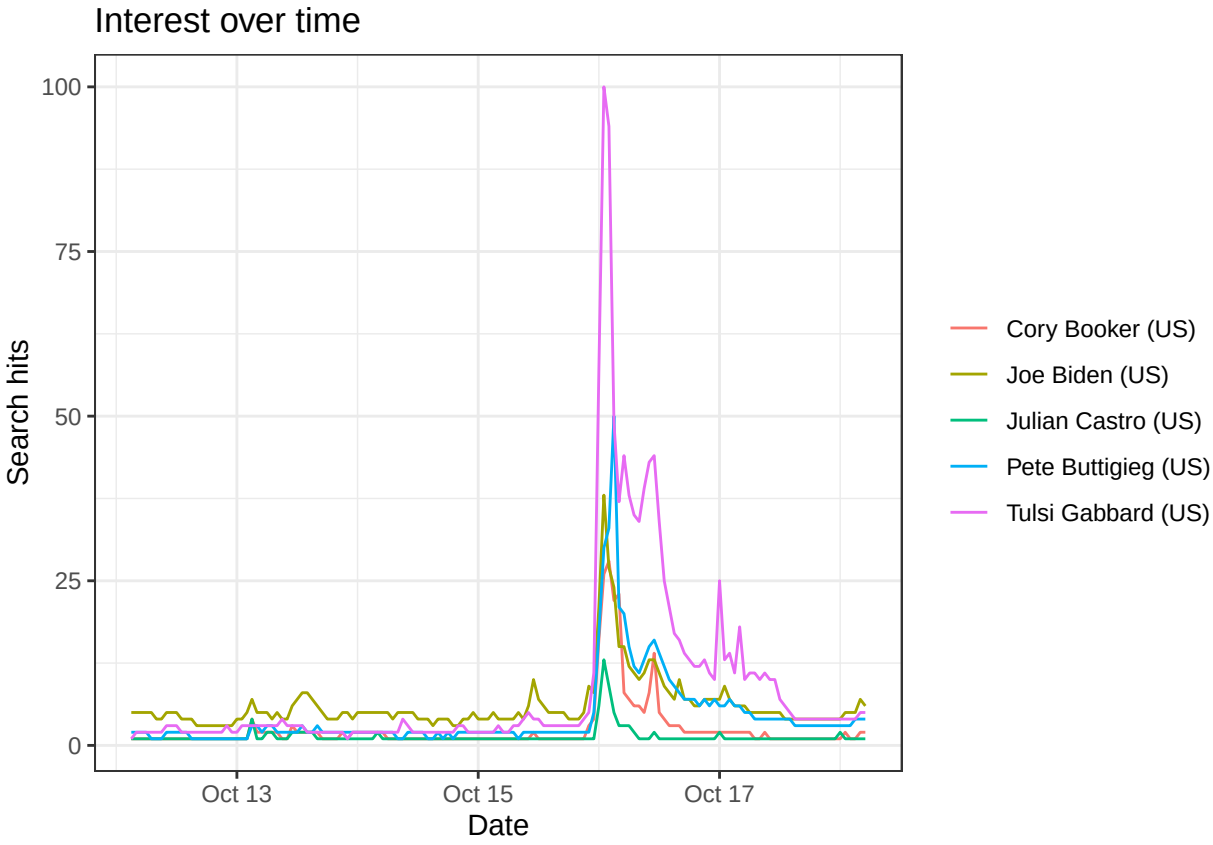
"US Presidential Election 2020 - Democrat Candidate Betting Odds." Oddschecker, 2019, www.oddschecker.com/politics/us-politics/us-presidential-election-2020/democrat-candidate.

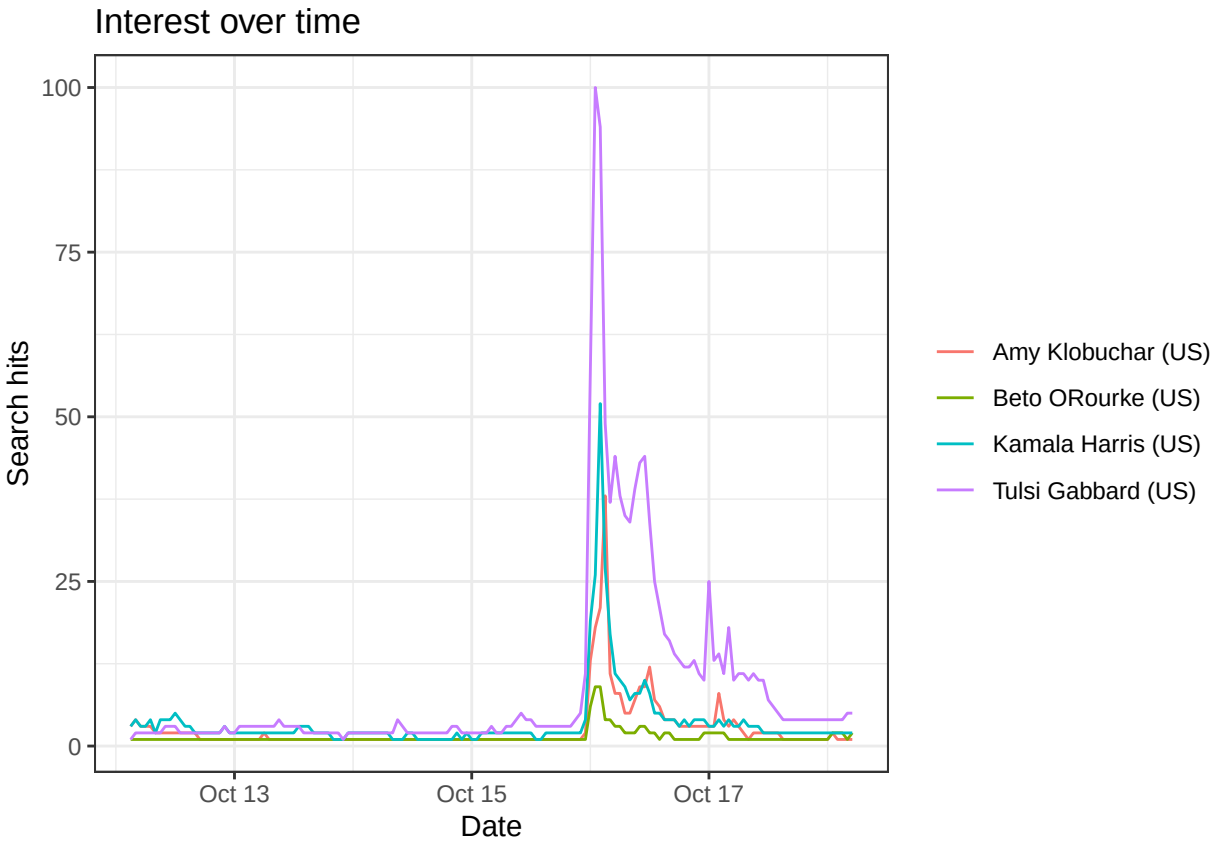
"Warren's Support Persists Despite Attack-Heavy Democratic Debate in Ohio." Morning Consult, 17 Oct. 2019, morningconsult.com/2019/10/17/warrens-support-persists-despite-attack-heavy-democratic-debate-in-ohio/.

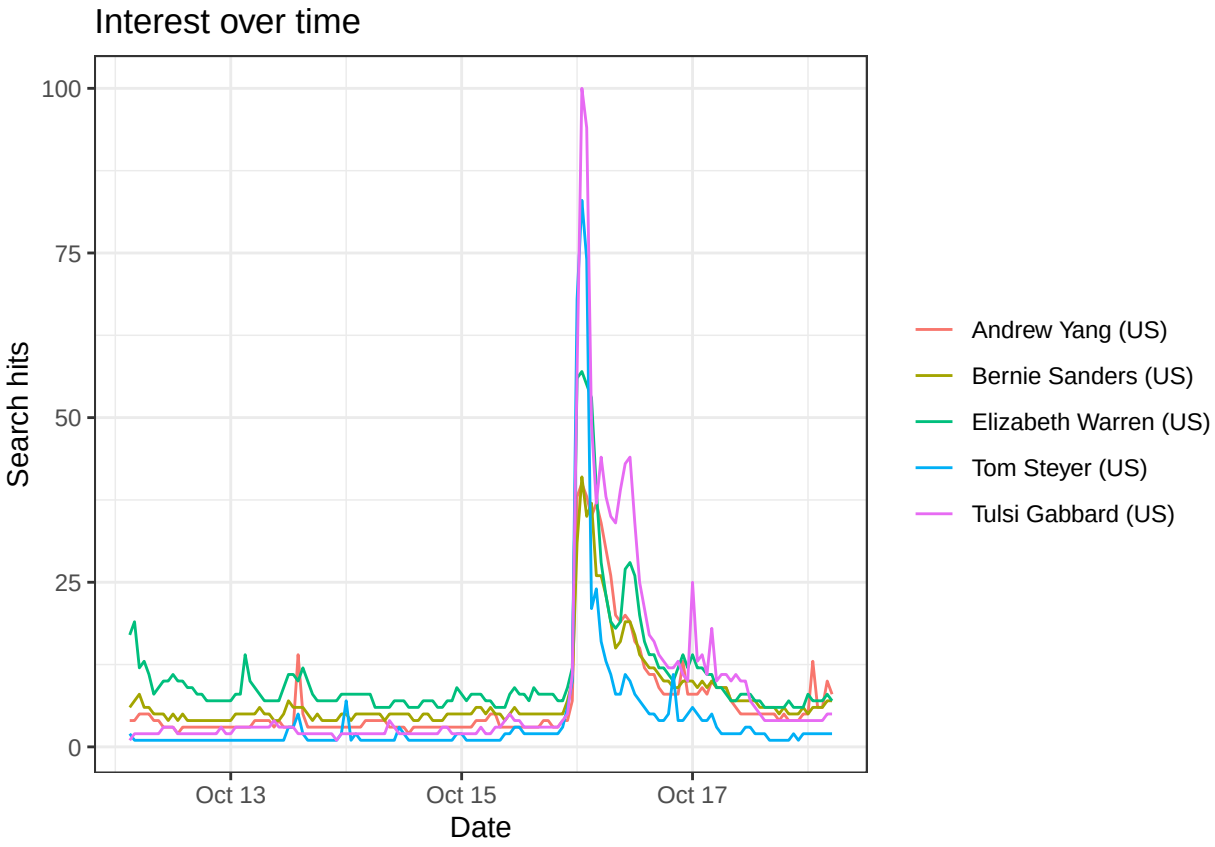
Wojcik, Stefan, and Adam Hughes. "How Twitter Users Compare to the General Public." Pew Research Center: Internet, Science & Tech, Pew Research Center, 24 July 2019, www.pewresearch.org/internet/2019/04/24/sizing-up-twitter-users/.

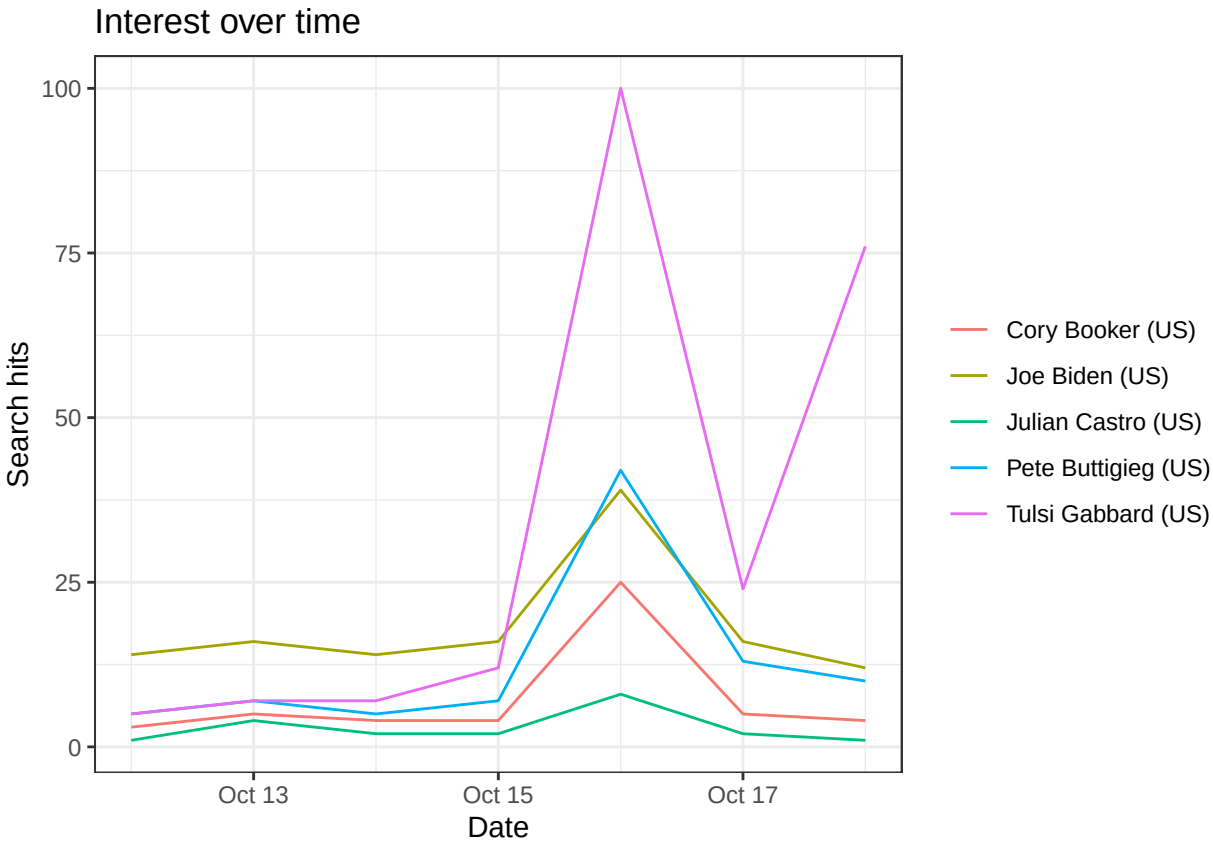
Wu, Nicholas. "ABC Says 14 Million Viewers Watched the Third Democratic Presidential Debate." USA Today, Gannett Satellite Information Network, 13 Sept. 2019, www.usatoday.com/story/news/politics/elections/2019/09/13/how-many-viewers-watched-third-democratic-presidential-debate/2312880001/.

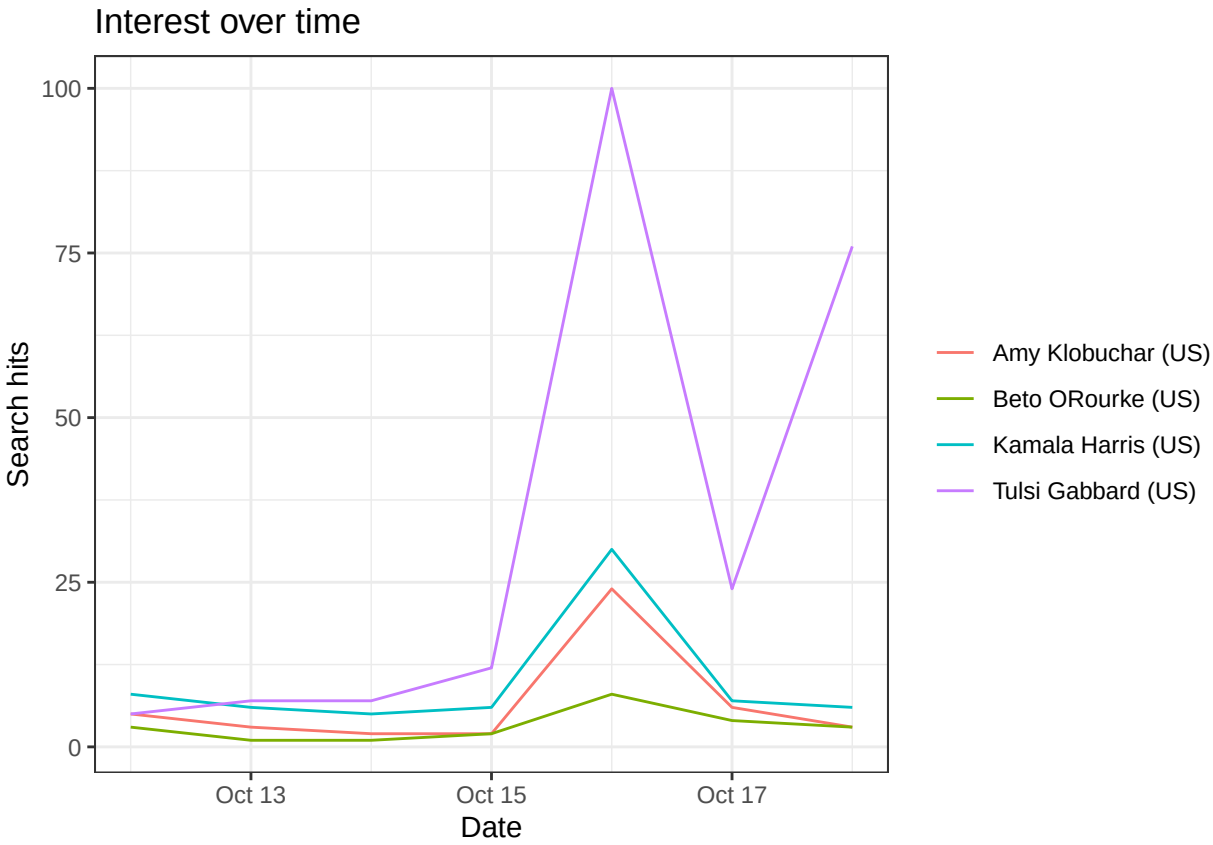
Appendix: Additional Google Trends Plots

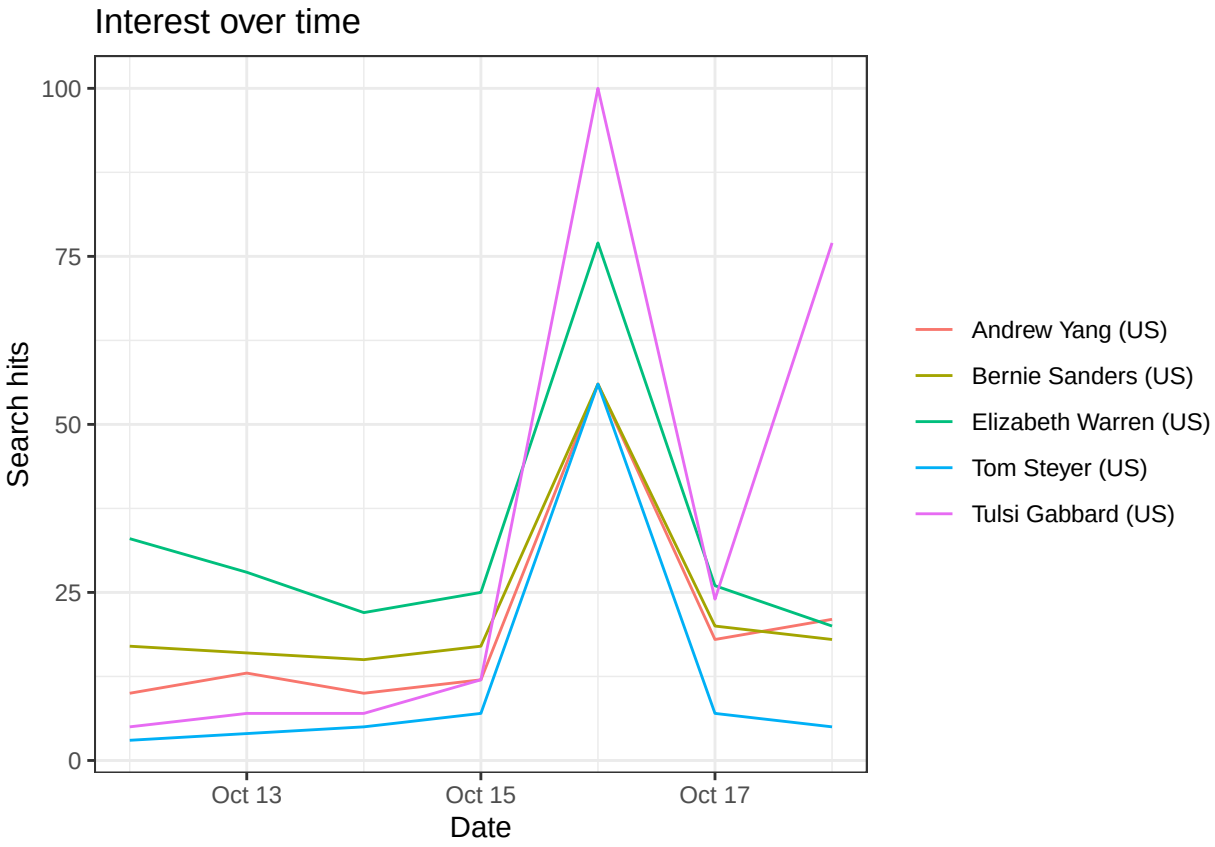





Interest over time
Search hits
100
75
50
25
0
Oct 13
Oct 15
Oct 17
Date
Andrew Yang (US)
Bernie Sanders (US)
Elizabeth Warren (US)
Tom Steyer (US)
Tulsi Gabbard (US)


Interest over time
Search hits
100
75
50
25
0
Oct 13
Oct 15
Oct 17
Date
Cory Booker (US)
Joe Biden (US)
Julian Castro (US)
Pete Buttigieg (US)
Tulsi Gabbard (US)


Interest over time
Search hits
100
75
50
25
0
Oct 13
Oct 15
Oct 17
Date
Amy Klobuchar (US)
Beto ORourke (US)
Kamala Harris (US)
Tulsi Gabbard (US)


Interest over time
Search hits
100
75
50
25
0
Oct 13
Oct 15
Oct 17
Date
Andrew Yang (US)
Bernie Sanders (US)
Elizabeth Warren (US)
Tom Steyer (US)
Tulsi Gabbard (US)