**DS710\_FA19**

**Final Project**

Chris Obermeier

**Hypothesizes:**

Null Hypothesis: Twitter is able to show the same rankings of popularity as the polling is able to do during the democratic debating period in time.

Alternative Hypothesis: Twitter’s reactions are not linked to what polling data will show, and are not a substitute for it.

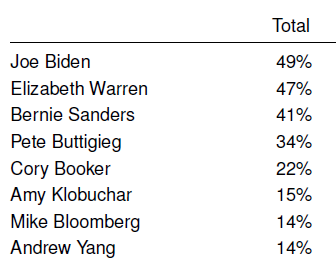
**Introduction:**

Currently (December, 2019), Democratic debates and polling for the 2020 presidential campaign are taking place. The current climate of Twitter is very active during this time and shows that a lot of the debates are being live-tweeted and talked about during this time. Additionally, 2 hash tags are being used on Twitter officially by the Democratic Party for these debates “#demdebate” and “#democraticdebate.”

**Methodology:**

The official hash tags can be used to search the “Twitterverse” using their own API, along with other words or names used that the candidates identify with to not only identify tweets about the campaign. Additionally, we can use RStudio to identify whose campaign that it is that is being talked about with each tweet and ultimately whether or not it is a positive or negative tweet.

Each tweet will be taken from the Twitter API, cataloged, and then analyzed using the following steps:

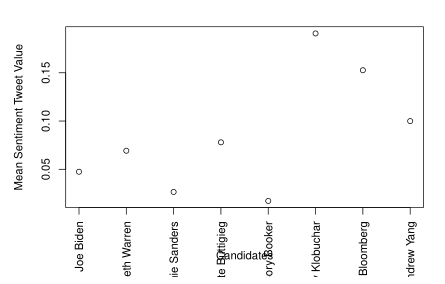
1. Only grab tweets with the official hash tags
2. Categorize which candidate each tweet is about.
3. Figure out if the tweet is positive or negative
   1. Score the tweets based on numeric value with high numbers being positive and lower numbers being more negative.
4. Take a mean of each candidate’s positive and negative values and assign the value to each candidate.
   1. I chose to use the “sentimentr’ library offered by RStudio’s supporters to do this scoring.

The values that are being used as the polling numbers were gathered from the Economist / YouGov poll on December 11th at 12:00 PM. The polling was done during December 7-10th on 1500 US Adult citizens registered either as a Democrat or Independent. Each person was polled on their probability to vote on each of the democratic candidates.

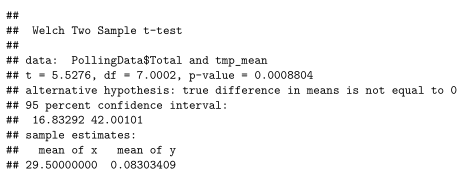
Figure - Economist / YouGov December 11th 2020 Presidential Poll

The voting was further broken down by gender, age, race, party ID, income, and region, but those drill-down metrics were not used on this study. Instead this study is only focusing on the total percent likelihood for each person to vote for a candidate.

Additionally, only the top 8 candidates were used to keep the study as simple as possible due to the large amount of candidates still running for the 2020 Democratic presidential nomination.

**Conclusion:**

Twitter’s tweeting did not even come close to reflecting the polling data as seen in the Economist / YouGov study. In fact, it was almost the complete opposite based on the data that was pulled from Twitter and ranked than what the polls reflected. The highest mean of positivity was reserved for the bottom 3 polling candidates (Amy Klobuchar, Mike Bloomberg, Andrew Yang), and the worst was reserved for the third highest candidate (Bernie Sanders).

A paired t-test on each of the outcomes was done, and the result was that the P-value was well under 0.05 and null hypothesis was rejected. The conclusion is that Twitter’s sentiment is not reflective of the polling statistics generated by doing traditional polling.

**Limitations:**

One of the limitations of the study was that each candidate did not have the same amount of tweets for them. In fact, looking at the data, it seems that by having fewer tweets for each candidate their own tweets then are taken into account, which tends to score very high on the positivity scaled. Future tests would take into account using the same amount of data for each candidate and also removing official or sponsored Twitter accounts from analysis.

A second limitation was that of the scoring library itself. I did not write how this was scored which led to some difference in how a layman may score them. This was a needed limitation, but looking into further efforts to make weighting more precise would be worth the time investment in the future.

Finally, another situation was that I wanted to do this during the actual live debates to reflect more responsiveness to the issues being talked about. Twitter’s API did not let me go back to the previous debate during the time of the November dates, so in the future I would make sure that I was able to do it live.