I chose to scrape as many comments from posts reaching “hot” status for each of the most popular subreddits dedicated for the five candidates most likely to win the 2020 Presidential Election, as set by the betting markets. These candidates are Donald Trump, Elizabeth Warren, Bernie Sanders, Joe Biden, and Pete Buttigieg. Reddit comments provide a different and useful view into the candidates' supporters because the data aren’t bound by a list of users or 140 characters at a time like many Twitter data sets are. I set the program to collect as many comments from each subreddit that it could. I did this for two main reasons. One was for a much richer data set, I standardized most of the descriptive statistics to combat any discrepancies. The second reason was that I knew the subreddit for Joe Biden, whose supporters are notoriously quiet online, would pull far fewer comments since his most popular subreddit, r/JoeBiden, had by far the fewest “members” (people who subscribe to a subreddit) at around 2,200 while others like Trump’s had about 781,000 members and Sanders with 338,000. I still felt it was important to include him since he’s leading most polls in the Democratic primary. For more in-depth analysis, I will use this data set to create a Naïve Bayes classifier to see which words are the most informative for each subreddit. When comparing Democrats’ subreddits to Trump’s it’s clear that Trump supporters are more likely to talk about the impeachment inquiry, and Democratic supporters are much more likely to talk about the ongoing Democratic primary, as well as policies that their candidates support.

One question that I wanted to answer through this data set was whether the candidates’ subreddits matched my conceptions of the candidates. This seems to have shone through in terms of length of comments and lexical diversity. Elizabeth Warren, who is the most policy-focused candidate, appears to have much lengthier posts with a much higher lexical diversity than the rest of the candidates. On the other side of that spectrum, Trump, who doesn’t read his daily intelligence briefings and prefers “killer graphics” to written reports1, sees a subreddit whose commenters offer a terseness well-suited for their president. r/The\_Donald has by far the fewest words per comment and the least diverse lexicon. The second thing I wanted to investigate was if the subreddits were talking about hot topics at differing levels. I chose impeachment and healthcare for this and found that the two furthest left candidates, who both support a similar Medicare-For-All bill, see healthcare used the most, while the two moderate Democrats see a slight decline, and Trump sees a precipitous drop off. While comments containing impeachment in Trump’s subreddit occur much more often than any Democratic candidate.

**Candidate Subreddit Descriptive Statistics**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | r/SandersForPresident | r/ElizabethWarren | r/Pete\_Buttigieg | r/JoeBiden | r/The\_Donald |
| Tokens | 63,931 | 81,470 | 87,765 | 30,465 | 51,323 |
| Comments | 35,744 | 39,290 | 25,849 | 6,661 | 16,365 |
| Total words used | 1,221,741 | 3,004,812 | 2,001,073 | 505,161 | 677,143 |
| Words per comment | 35.18 | 50.76 | 39.71 | 38.92 | 21.69 |
| Lexical diversity | 19.11 | 36.88 | 22.80 | 16.58 | 13.19 |
| Percent impeachment | 0.003% | 0.006% | 0.008% | 0.009% | 0.062% |
| Percent healthcare | 0.080% | 0.074% | 0.031% | 0.060% | 0.003% |

When looking at the top words used on each subreddit, many of the top words are similar. Each subreddit is talking about their candidate and what people like and think about any given topic. Further down more interesting words occur. Pete Buttigieg’s subreddit was talking about Iowa before he started polling first there, Joe Biden’s subreddit has talked a lot about California Senator Kamala Harris, who went at him during the first debate for his past busing policies, and r/The\_Donald is the only subreddit to see curse words in its top 50. All Democratic candidates see Trump in their top 50, but most see it in the bottom half of this list. All except r/Joe\_Biden where it’s the sixth most popular word. So far most national polls have had Biden as the most likely Democratic candidate to beat Trump, and many Democratic voters have put their support behind Biden mainly for this reason. This may explain why his subreddit talks about Trump so much more often than the other Democrat supporting ones.

**Candidate Subreddit Word Rankings with Stop Words Removed**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Ranking | r/SandersForPresident | r/ElizabethWarren | r/Pete\_Buttigieg | r/JoeBiden | r/The\_Donald |
| 1 | bernie | warren | pete | biden | like |
| 2 | people | would | like | like | would |
| 3 | like | people | people | would | people |
| 4 | would | like | think | think | get |
| 5 | get | think | would | people | trump |
| 6 | think | get | one | trump | one |
| 7 | even | bernie | get | joe | think |
| 8 | one | one | really | one | even |
| 9 | make | going | also | get | know |
| 10 | know | even | even | warren | see |
| 11 | also | really | see | going | going |
| 12 | much | also | know | even | us |
| 13 | really | see | good | really | make |
| 14 | want | much | going | bernie | got |
| 15 | going | tax | warren | good | want |
| 16 | good | good | much | see | never |
| 17 | right | make | campaign | also | say |
| 18 | see | plan | make | much | need |
| 19 | us | want | want | want | could |
| 20 | need | need | way | support | president |
| 21 | could | way | said | sanders | time |
| 22 | still | still | still | vote | heard |
| 23 | warren | know | time | make | really |
| 24 | sanders | go | could | still | good |
| 25 | way | lot | lot | know | right |
| 26 | time | could | buttigieg | could | go |
| 27 | vote | pay | say | way | fucking |
| 28 | money | say | candidate | say | said |
| 29 | support | support | biden | lot | still |
| 30 | go | sanders | right | candidate | way |
| 31 | many | time | first | voters | someone |
| 32 | trump | money | support | need | also |
| 33 | say | actually | need | win | shit |
| 34 | pay | right | candidates | candidates | much |
| 35 | actually | better | many | said | every |
| 36 | lot | biden | someone | go | actually |
| 37 | someone | health | go | harris | take |
| 38 | better | take | new | actually | back |
| 39 | take | wealth | actually | better | new |
| 40 | new | candidate | bernie | someone | guy |
| 41 | sure | taxes | got | time | first |
| 42 | every | someone | take | right | thing |
| 43 | first | campaign | iowa | many | something |
| 44 | campaign | many | better | left | years |
| 45 | work | insurance | black | polls | many |
| 46 | healthcare | said | something | democratic | left |
| 47 | biden | trump | sure | far | sure |
| 48 | said | candidates | trump | polls | state |
| 49 | got | supporters | love | president | look |
| 50 | government | public | back | debate | democrats |

To further compare these five subreddits, I’ve created a Naïve Bayes Classifier and compared individual subreddits to each other as well as grouping subreddits based on ideological similarities to compare against one another. More so than looking at top 50 words, this provided a better look into what subreddits are more like to talk about than others. The first comparison I ran was Sander’s subreddit against Trump’s. When looking at the top 20, most words that appear for Bernie deal with the Democratic primary and his healthcare plan, and words that deal with the impeachment inquiry go for Trump. Some of these words for Sanders include Medicare, doctors, Iowa, caucus, and polls. While words for Trump include, Ukraine, (R-OH4 congressman) Jim [Jordan], testimony, and (D-CA28 congressman) Adam [Schiff]. During congressional hearings on impeachment, Jim Jordan has aggressively defended the president while Adam Schiff has gone after the president’s behavior.

Accuracy = .6894

|  |  |  |
| --- | --- | --- |
| **Word** | **Candidate** | **Score** |
| bernie | Bernie | 171.7 |
| pete | Bernie | 99.8 |
| ukraine | Trump | 97.1 |
| medicare | Bernie | 96.9 |
| doctors | Bernie | 47.1 |
| jim | Trump | 46.1 |
| sanders | Bernie | 46 |
| iowa | Bernie | 42.4 |
| jordan | Trump | 41.6 |
| witness | Trump | 40.9 |
| caucus | Bernie | 38.2 |
| healthcare | Bernie | 35.6 |
| islam | Trump | 33.4 |
| testimony | Trump | 32.6 |
| ukrainian | Trump | 28.1 |
| insurance | Bernie | 27.7 |
| donors | Bernie | 25.8 |
| polls | Bernie | 24.7 |
| adam | Trump | 22.7 |
| hunter | Trump | 22.7 |

Next, I combined the four Democrats together and compared them to the Trump subreddit by itself. One aspect that becomes obvious is that Trump supporters aren’t talking about South Bend, Indiana mayor and current favorite to win the Iowa caucus, Pete Buttigieg. With a score of 751.4 to 1, this is the single largest discrepancy in the whole analysis. It didn’t make the written report, but even when comparing Warren against Trump, “Pete” is the number one term with a score of 255.9:1 in favor of Warren. The top 20 most informative features for Democrats are dominated by candidate names and healthcare discussion. Trump’s list is a bit more diverse this time around (never thought I’d use “Trump” and “diverse” in the same sentence) including non-impeachment words such as “drudge”, the last name of Matt Drudge, writer and owner of the right-wing conspiracy blog “Drudge Report”, “tds” which is short for Trump Derangement Syndrome, I had never heard this term before, but apparently is used to describe people with a deep hatred for the current president, and “Islam” which most likely has to do with Islamophobia shared by both the president and his supporters.

Accuracy = .8482

|  |  |  |
| --- | --- | --- |
| **Word** | **Group** | **Score** |
| pete | Democrats | 751.4 |
| islam | Trump | 155.5 |
| jordan | Trump | 133.9 |
| hearsay | Trump | 131.9 |
| bernie | Democrats | 120.4 |
| drudge | Trump | 118.5 |
| taylor | Trump | 103.1 |
| tds | Trump | 78.5 |
| iowa | Democrats | 67.4 |
| bureaucrats | Trump | 60 |
| medicare | Democrats | 56.8 |
| witnesses | Trump | 51.8 |
| mints | Trump | 50.8 |
| drumpf | Trump | 47.7 |
| sanders | Democrats | 46.2 |
| warren | Democrats | 45.1 |
| liz | Democrats | 44.8 |
| hoax | Trump | 44.6 |
| diplomats | Trump | 44.6 |
| beto | Democrats | 41.9 |

Within the code that is attached to this assignment there are a few more examples, including progressive Democrats vs the moderate Democrats, Warren vs Sanders, and Warren vs Trump. If I were to do this again, I’d limit the responses to a specific date range since most topics of conversation have to do with current news surrounding their specific candidate. Next, I would implement a sentiment analysis and plot sentiment over time to see if supporters can predict poll results and eventually primary/caucus results. Right now, this isn’t going to be my capstone, but if I continue to struggle getting my current capstone idea to work, I believe this is a strong backup plan since 29 states hold their primary or caucus before April which would give me plenty of chances to pull data, analyze it, and try to model it.

1[https://www.washingtonpost.com/politics/breaking-with-tradition-trump-skips-presidents-writte-intelligence-report-for-oral-briefings/2018/02/09/b7ba569e-0c52-11e8-95a5-c396801049ef\_story.html](https://www.washingtonpost.com/politics/breaking-with-tradition-trump-skips-presidents-written-intelligence-report-for-oral-briefings/2018/02/09/b7ba569e-0c52-11e8-95a5-c396801049ef_story.html)