James Taylor

Assignment 3

University of Maryland University College

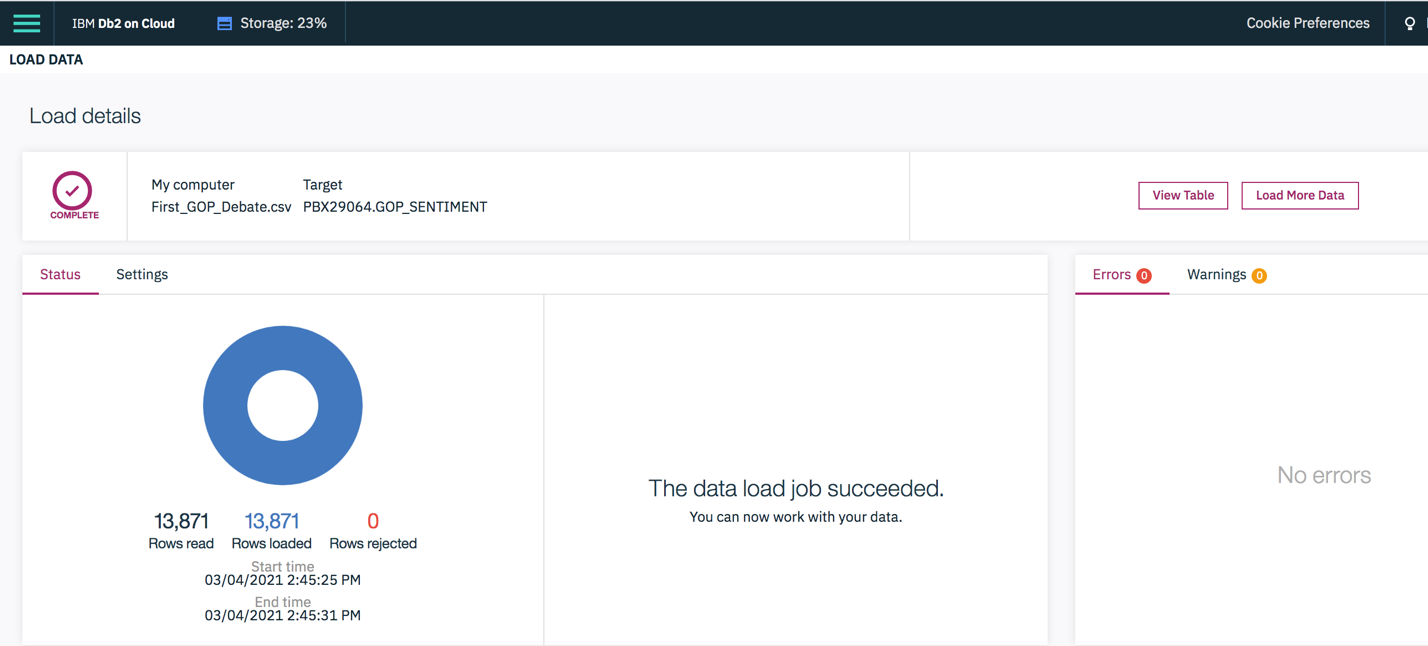
DATA650-9041 Spring 2021

Dr. Gortcheva

jtaylor380@student.umgc.edu

**Introduction**

The goal is to look at the words that are associated with positive and negative sentiment tweets associated with a GOP debate. The sorts of things being talked about in negative vs positive tweets will be the main objective. Various approaches will be used to look at the words associated with each sentiment.

**Data** 

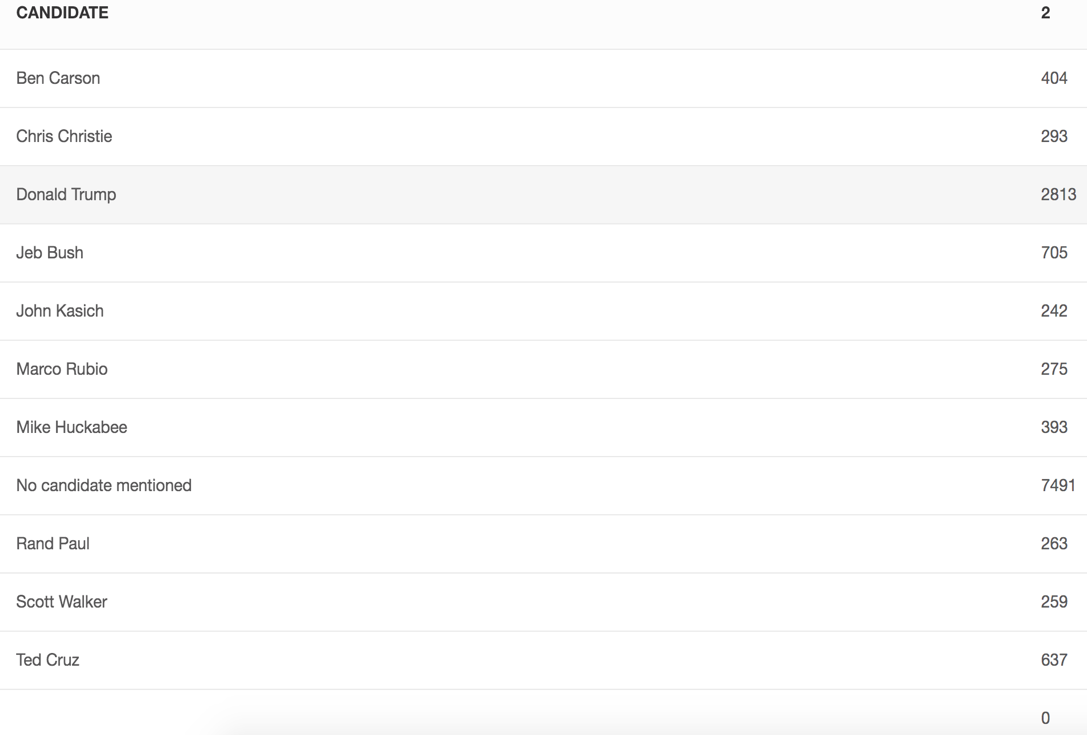
The data was successfully loaded into the IBM Db2 cloud database. No errors occurred or rows rejected during the upload. The data was viewed on the database by the “View Table” ability to confirm the data was organized as desired. The data types of the variables were checked to as well.

In the SQL environment of the database, the row count and variables were also checked to make sure the table holding the data was organized correctly and nothing was missing/wrong.

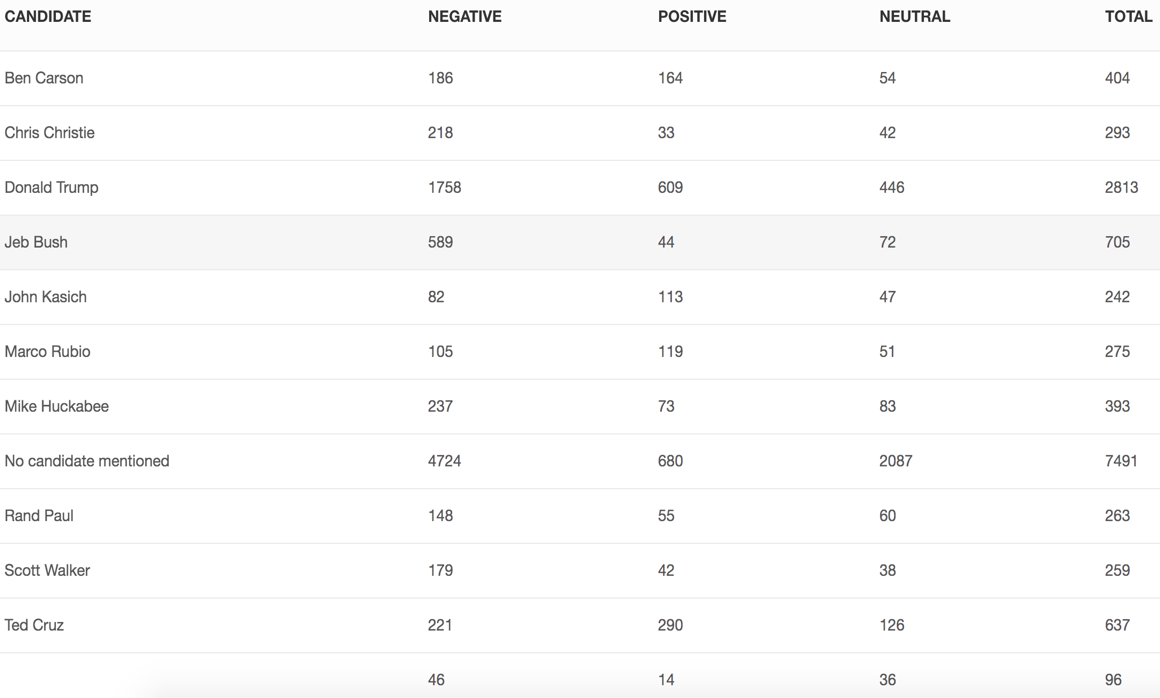
**SQL Exploration**

The tweets are likely mostly from people who think of themselves as Republicans or conservatives. Liberals are likely not to watch the event in mass. This means the negative sentiments are probably not from liberals disagreeing with everything being said, but conservatives pointing out the things they don’t like or think should be fixed.

The below figure is the amount of times a candidate is believed to be mentioned in the tweets in this dataset. This was created using the SQL Editor in the Db2 database. In the tweets, candidates may not be explicitly mentioned. The dataset provides a candidate’s name and the likelihood they are being mentioned, which is why above the word ‘believed to be mentioned’ was used, because it is a prediction in some cases.

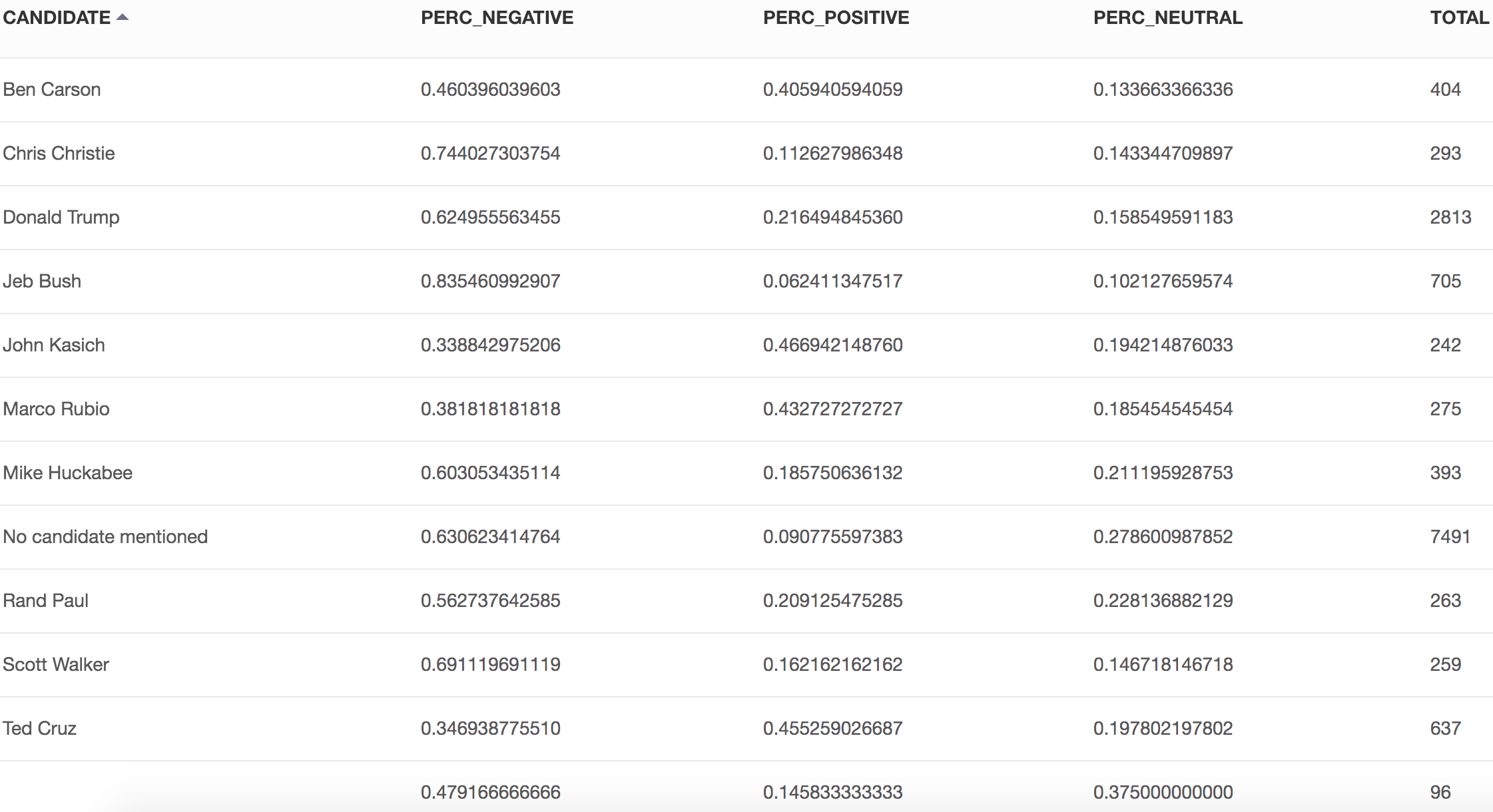


It is most common that a tweet doesn’t refer to any particular candidate. Donald Trump is mentioned first among the candidates provided, showing up 2,813 times. Chris Christie was least mentioned in the tweets.

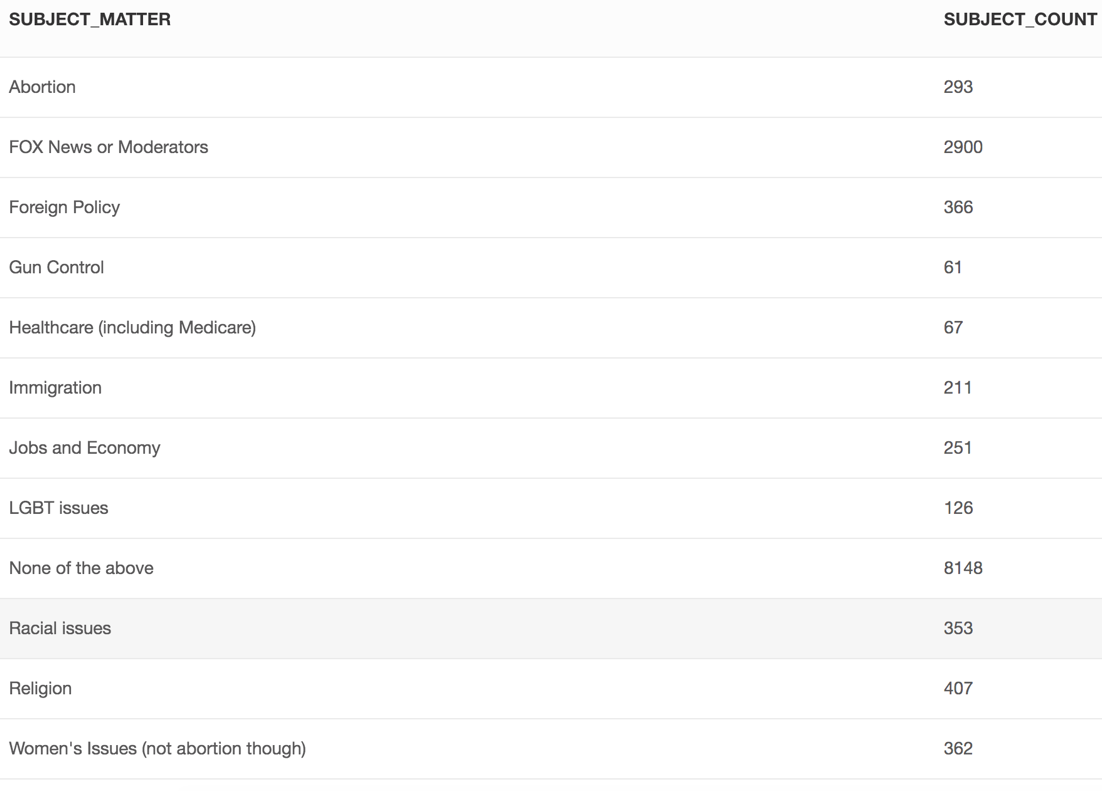


Above is the amount of tweets by sentiment and candidate. Trump has the most positive tweets attributed to him, but also the most negative. Chris Christie appears to be in a bad state, with 218 negative tweets but only 33 positives. Kasich appears to have the best positive tweet ratio but that’s just from guessing based off the numbers above. Jeb Bush has a high negative tweet ratio as well, reporting 589 negative but only 44 positive tweets. For the most part tweets negative and less so neutral and positive. This may be because of the person’s motivation for tweeting. If something makes them angry they may be more inclined to tweet compared to tweeting about something they are indifferent or satisfied about.

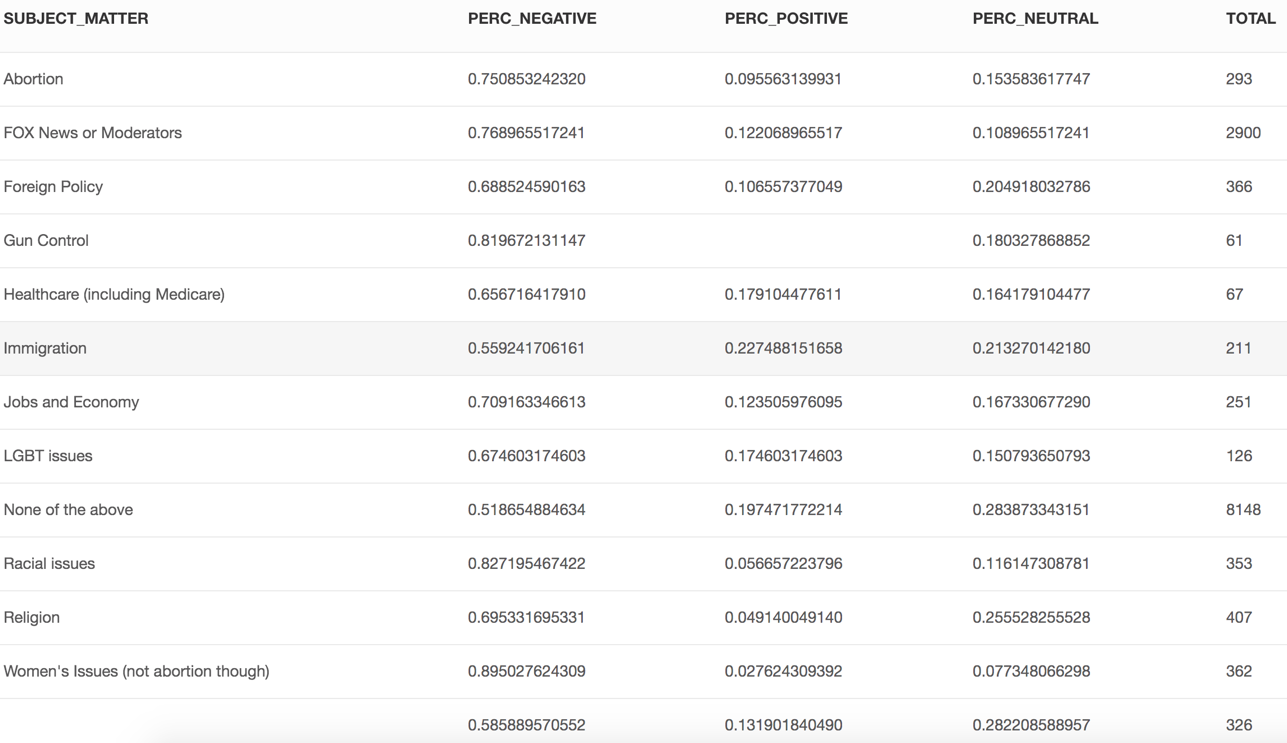
To get a better understand how many tweets about a candidate are positive and negative, the ratio was calculated. This was done by taking the frequency of a given sentiment and divide it by the frequency of all tweets about that candidate. This returns the proportion of sentiments for a candidate. The results are below.



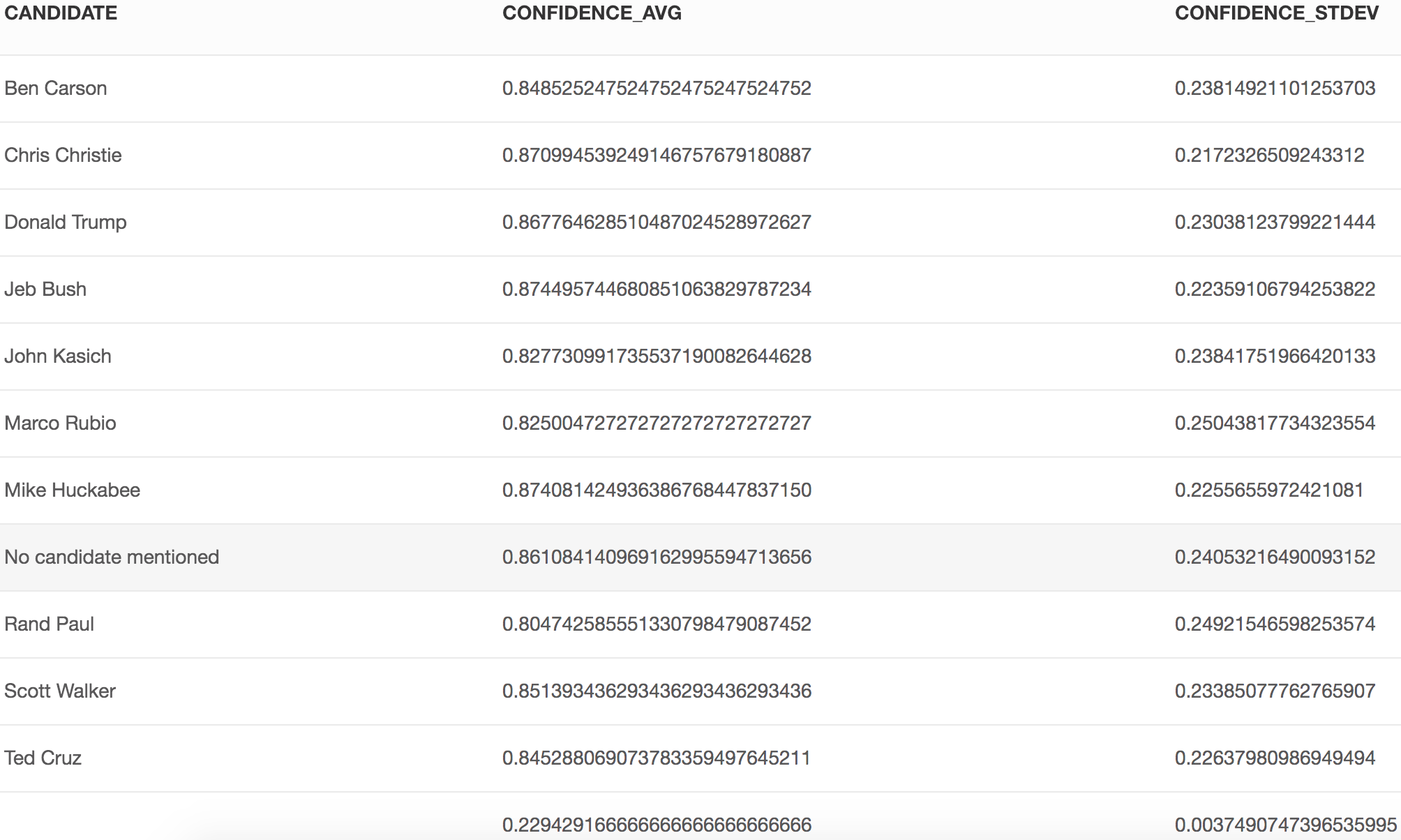
This is a better view of the tweets’ sentiments on the candidates because it controls for the amount of tweets candidate got. For example, Trump had more than three times the amount of negative tweets than the second most negative tweet amount (Bush) but this shows Bush having the worse negative tweets ratio. Bush was only able to get 6% of his relevant tweets to be positive. This is hardly the results a presidential-runner would hope for. Kasich and Cruz had about 45% of tweets about the be positive which is the best of the competitors. Those two had very low negative ratios as well.



Above are the counts by tweet subjects. The most frequent subject was not listed, this was because it was unclear what the subject was or it wasn’t one of the specified categories. The FOX network and its moderators of the debate where the hot topic far and away, being mentioned seven times more often in thee tweets than the second most frequent topic which was religion.



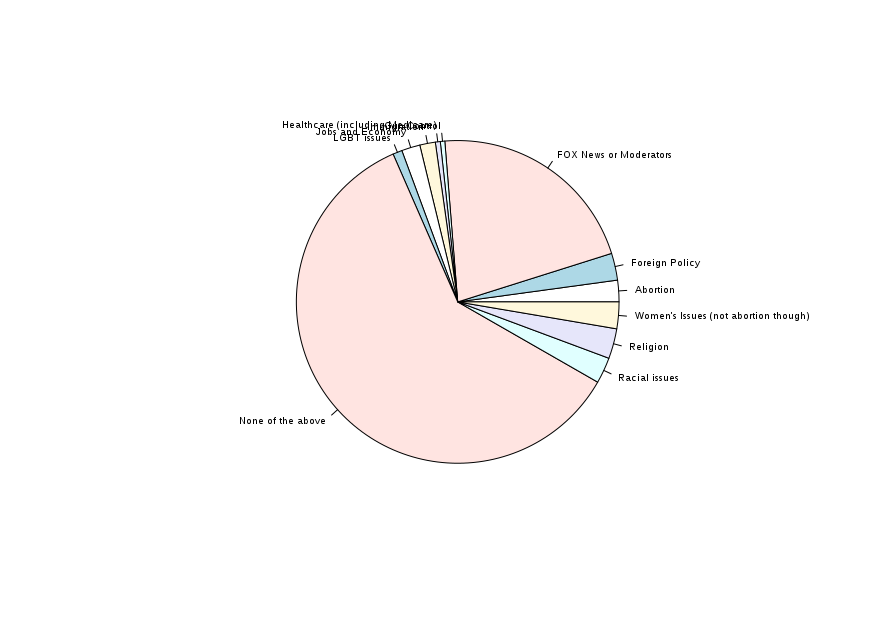
The above figure shows the proportion of sentiments across each subject. They mostly lean negative. For gun control, there were no positive tweets which is why it’s blank. Nearly 90% of women’s issues tweets were negative in nature. It really goes to show the political landscape on twitter where issues and problems are the dominant topic. Every topic given was majority negative.



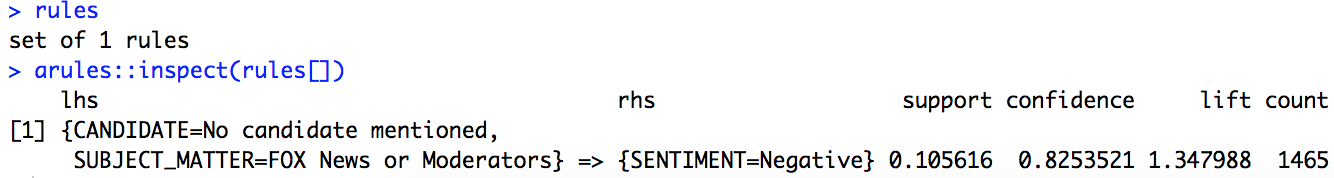
Above is the average and standard deviation of the candidate confidence measure. The data includes data where it is confident in a tweets reference towards a candidate. Because the average confidence value is high, the numbers of how many times a candidate is mentioned is trustable.

**RStudio**

The dataset was loaded into R for more analysis. Various techniques were used to analyze the data further.

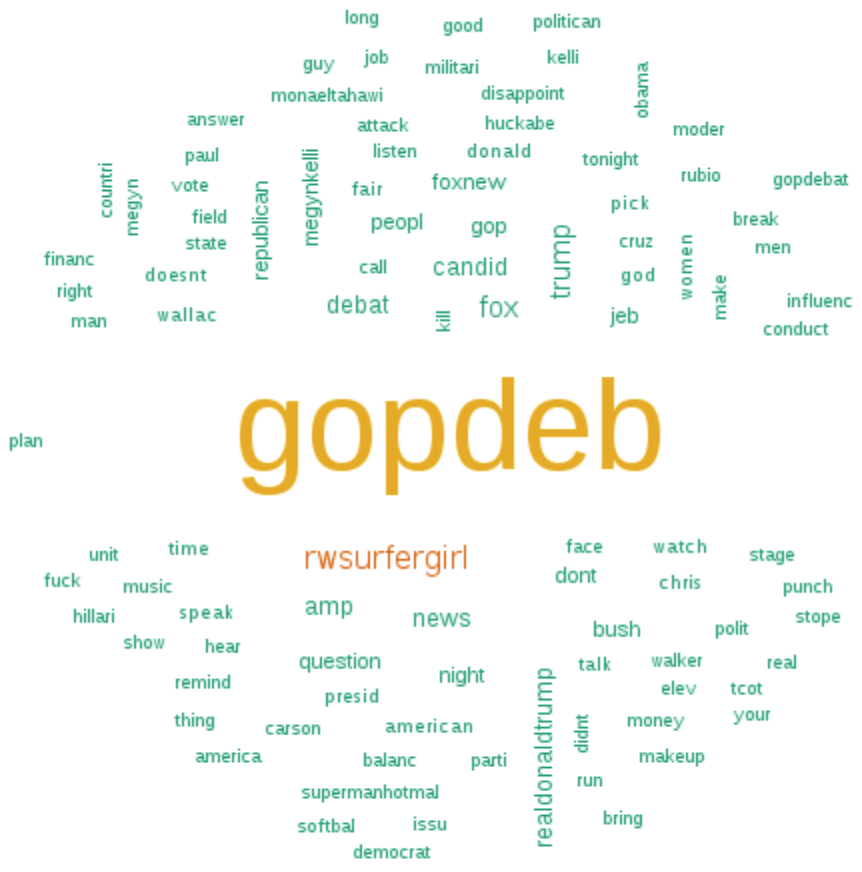


The above is the pie chart of the tweet subject count. It reveals visually how often FOX or its moderators were mentioned. This was somewhat understandable, there was public debate on FOX News having a preference on the GOP candidates. Some candidates publicly stated they thought FOX was biased and excessively covering negative news on certain candidates to help candidates they did like. One reason this may be the case is the moderators where there for the entire event, as opposed the specific topics that were topics during the debate. The topics debated were in segments with time limits so each one could get covered during the event. Since viewers were already aware the moderators may be biased, any slight perceived bias during the debate was probably more likely to be noted. The subjects on the right side of the pie take up remarkably the same size of the pie. This may be related to time limits on subjects at the debate.

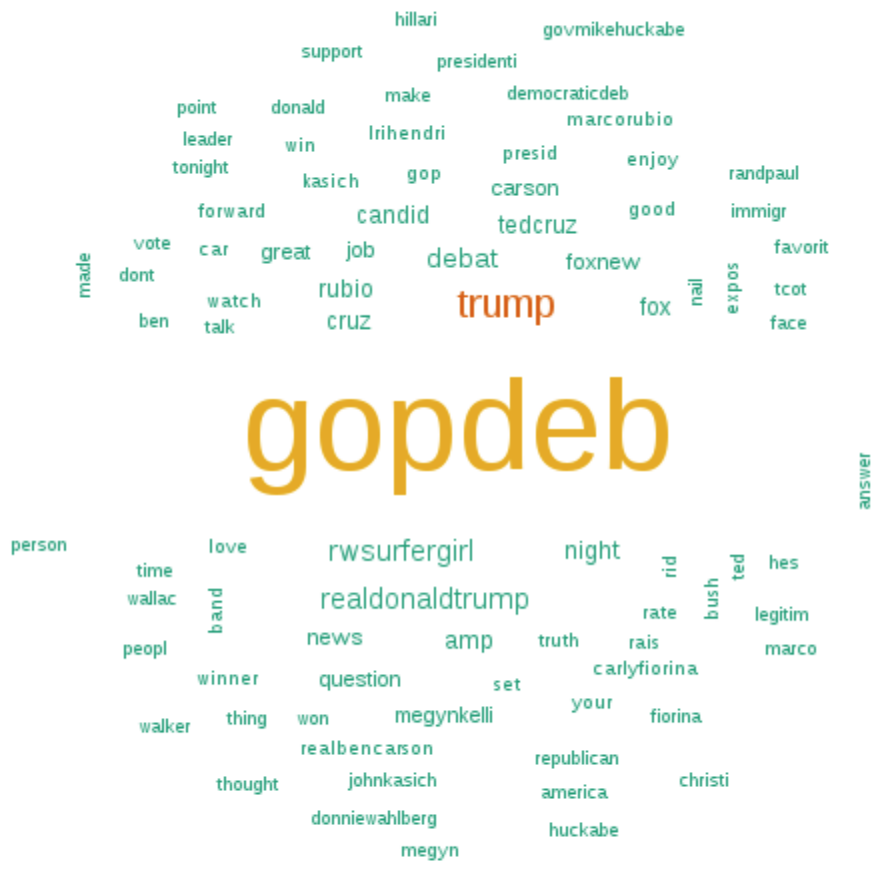
****

The apriori algorithm looks through the data to find correlations in the variables. The variables available to find correlations was candidate, sentiment, subject matter and retweet count. There was only one rule found by the algorithm that satisfied its threshold, which was when the subject matter of a tweet was Fox News or its moderators and no candidate was mentioned, 82% of them are of negative sentiment. This may be evidence of people complaining about perceived bias by Fox or its moderators.

Some data preprocessing occurred to inspect the text of the tweets. Characters related to URLs, punctuation, symbols and numbers were removed. Stop words were also removed, which are common words that are not of interest to our analysis like ‘the’ and ‘it’. The ‘english’ and ‘SMART’ stop words from the ‘tm’ package. The remaining words were extracted to their stems, meaning the root of the word. Al this allows better counting of the words we may find interesting.

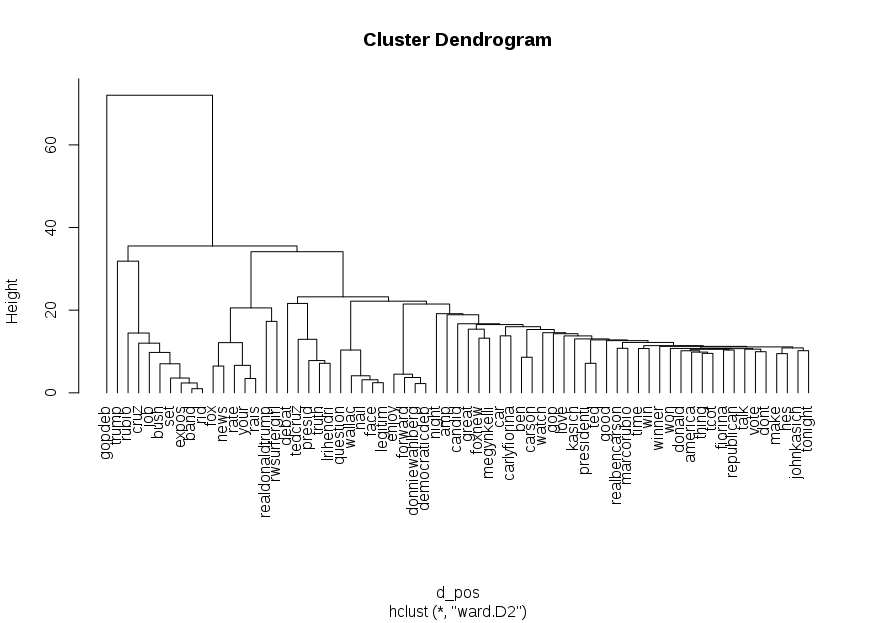
****

The above wordcloud is from the text from the tweets that were marked negative. The most popular negative term was ‘gopdeb’, which was likely a hashtag before the preprocessing removed the symbol. It is likely that it was used to pull the tweets from Twitter to get tweets related to the political event. The second most frequent term was ‘rwsurfgirl’ which was probably an account before the @ was removed. As far as who it was I’m not sure, it was apparently a Trump supporter who had one of their tweets showing dissatisfaction with Obama read on *Jimmy Kimmel Live* TV show by Barack Obama himself. The clip of him reading the mean tweet about himself went viral and people talked about it. The word ‘wallac’ is likely refereeing to Chris Wallace, a FOX presenter. One concerning thing was ‘punch’ being in the top 100 words.

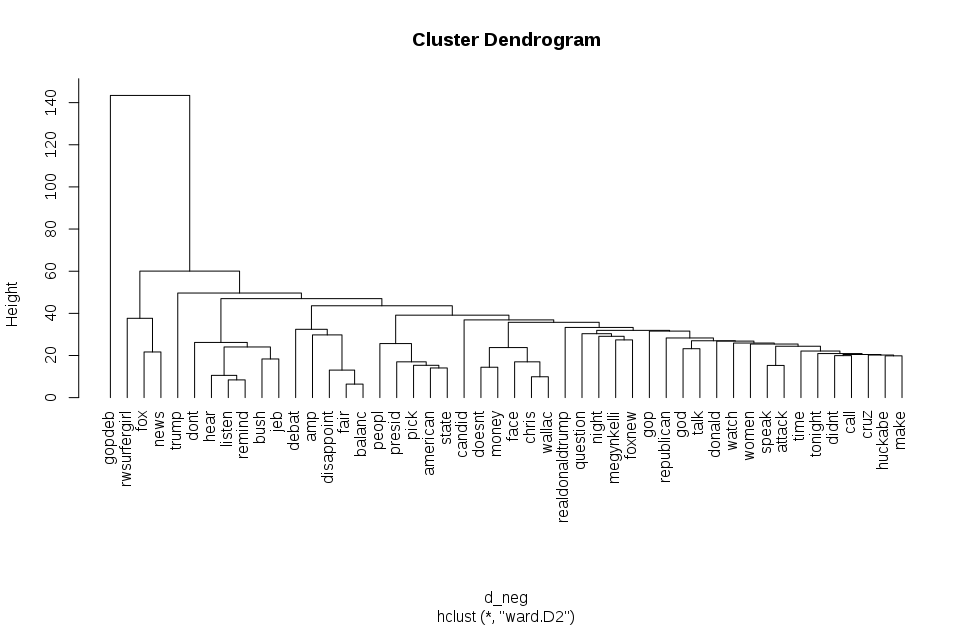


The above wordcloud is from text data from tweets marked as positive sentiment. It has some similar result to the negative frequency counts. The probably hashtag ‘gopdedeb’ is again the most referenced word. Trump was referred to a variety of ways but still to be some of the most common words. Cruz and Rubio managed to have their names near the top of the positive word frequency list. Some information that helped confirm this data is showing positive results is words like ‘good’ and ‘great’ and ‘love’ appear. The word ‘question’ is evidence that a significant amount tweets are about a questioned posed at the debate, along with ‘truth’.

The twitter account ‘rwsurfgirl’ is again one of the most prevalent words. This goes to show commonly talked about things may still show themselves in positive and negative tweets because it is common enough that there will be positive and negative sentiments about the topic.



A dendrogram of the positive words is above. This shows what words appear together frequently. Not all of these correlations are meaningful. For example, ‘don’t’ and ‘vote’ appear next to eachother above. Does that mean the tweets are telling people not to vote, or not to vote for a particular candidate? It is hard to tell just from this. Some correlations make sense. The word ‘ben’ and ‘carson’ are highly correlated because a candidates name was Ben Carson. Another clear correlation is Megyn Kelly and FOX News, as she was a presenter for them. Kasich appears frequently with the word president. People may have felt his answers and presentation were fitting of a president.



The above dendrogram are words from negative tweets that appear together frequently. One interesting correlation is between ‘money’ and ‘doesn’t’. People may be making references to what money can’t win you an election even though campaign adds can be bought and campaigns survive by paying staff. Another is ‘disappoint’, ‘fair’ and ‘balanced’ are correlated. This may be people expressing their dissatisfaction at FOX not being unbiased like they advertise. The word ‘women’ appears near ‘attack’ and ‘donald’, probably because of his controversies with women.

In the SQL section, it showed that only 2% of tweets about Women’s issues (not abortion) were positive and about 90% negative. This is the most negative subject provided in the dataset. From looking at negative tweets on the women’s issues subject, they appear mostly to be women showing displeasure at how some of the candidates talk about women. Below is the 30 most common words found in the tweets (after the same preprocessing mentioned earlier). Lots of attention went to Trump and his history of derogatory terms when referring to women he doesn’t like. Trump’s name appears a variety of ways in the below list. The words he used in referencing those women also appear, like pig and fat, which were mostly directed to a female comedian in a 2006 public war of words. The topic was brought up during the debate however, bring attention to it. Trump also made disparaging comments about Megyn Kelly earlier, who was one of the moderators of the debate. The most common use of the word ‘men’ appears to be mentioning that all 10 of the candidates at the debate were men. There were complaints about diversity and the fact that women’s issues were being debated but no women were present.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| gopdeb | 278 | issu | 37 | twitterland | 31 |
| women | 213 | call | 36 | make | 26 |
| men | 70 | control | 36 | polit | 24 |
| trump | 66 | social | 34 | donald | 23 |
| bodi | 61 | realdonaldtrump | 32 | stage | 23 |
| monaeltahawi | 55 | hold | 32 | pig | 22 |
| amp | 53 | boy | 31 | fat | 20 |
| gop | 48 | code | 31 | megynkelli | 19 |
| woman | 41 | sexul | 31 | hate | 18 |
| misogyni | 41 | tight | 31 | govern | 18 |

**Conclusion**

The analysis was able to uncover some of the topics that were in negative and positive sentiment tweets. Positive tweets talked about specific candidates, probably praising them and their performance during the debate. The presence of certain words showed evidence that some tweets were referencing how well they answered questions or spoke truthfully. Donald Trump’s name was quite prevalent as well. For negative sentiment tweets the big subject were Fox, its moderators and Trump. Fox had already faced accusations of bias in how it portrays candidates and it appears the debate didn’t help quell them. Trump’s controversies showed themselves in the negative sentiment analysis, with his words for women getting the most attention. This is one reason why the Women’s Issues subject was so negative. The attention was on Trump’s history of unpleasant words. There are many issues women have in the country, but the tweets were vastly on Trump’s words.

There are some limits on this analysis. One is like the Trump and women controversy, attention is what drives tweets. If people get riled up on certain topics, those will be what is seen in the tweets because that’s what’s on the public attention. Tweets are not a great data source for opinions on policies. Another limit is these are not the opinions of just Republicans. The word ‘KKKorGOP’ was a significantly frequent word and Republicans likely wouldn’t liken their party in such a way, providing evidence that non-Republicans were tweeting as well. More topics for positive and negative may have been uncovered if less words were included. Some blame falls on the inclusion of hashtags. This analysis treats all words equally, but words used in the sentence may not be used the same way as hashtags in the tweet. Treating these words as equal is probably not right.

More research should be done including different ways to categorize positive and negative tweets. The data provided had the sentiments given, but there’s more than one approach to extract a sentiment. There could be more granular sentiment done, like including very positive and very negative along with positive and negative. Things like bi-grams and tri-grams could also be used to find relationships in words.