Twitter Sentiment of Democratic Primary Candidates Around the Second Primary Debate

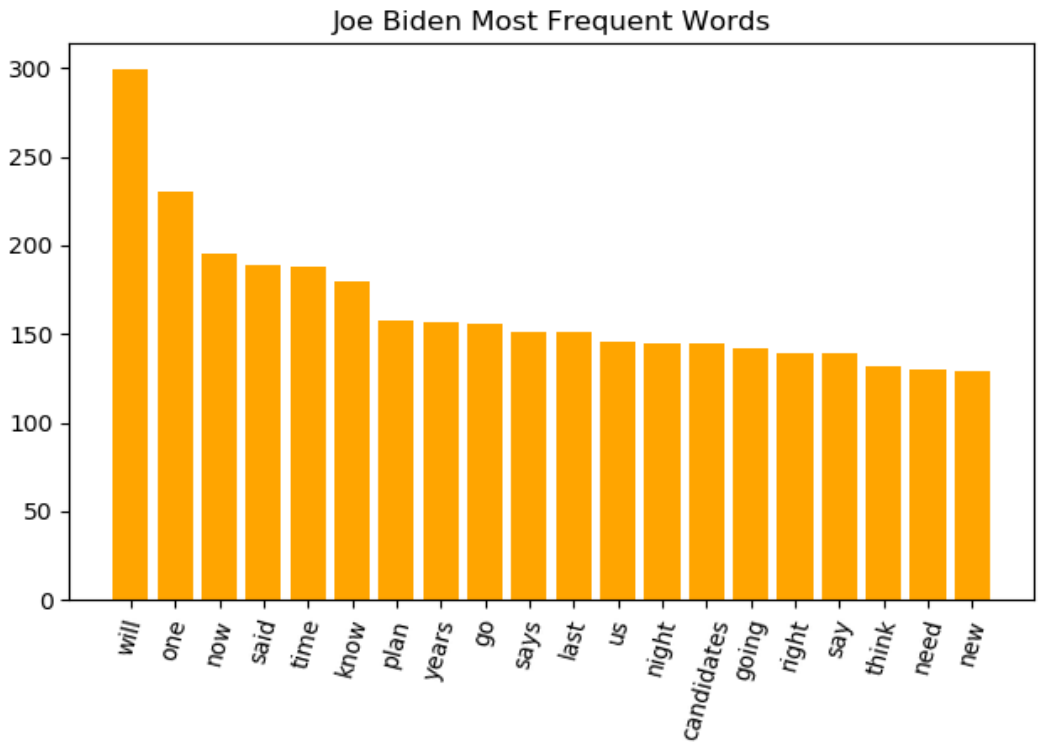
This is a report analyzing the twitter sentiment of the top four candidates competing for the 2020 Democratic nomination for president. Tweets made between 7/22 and 8/6 about each candidate were collected for analysis. The second Democratic primary debate took place over two nights on July 30th and 31st 2019 in Detroit, MI. Tweets were collected around these dates to analyze how sentiment on the four candidates varied over the days surrounding the debates. Tweets were extracted simply by searching for tweets that contained a candidate's full name. Bernie Sanders and Elizabeth Warren debated on July 30th and Joe Biden and Kamala Harris debated on the 31st. These are the candidates who are ranked highest in overall favorability polls. In order to predict the sentiment of these tweets, both unsupervised and supervised machine learning algorithms were trained. In addition to analyzing sentiment changes, I will also be comparing the advantages and disadvantages of these two different ML approaches.

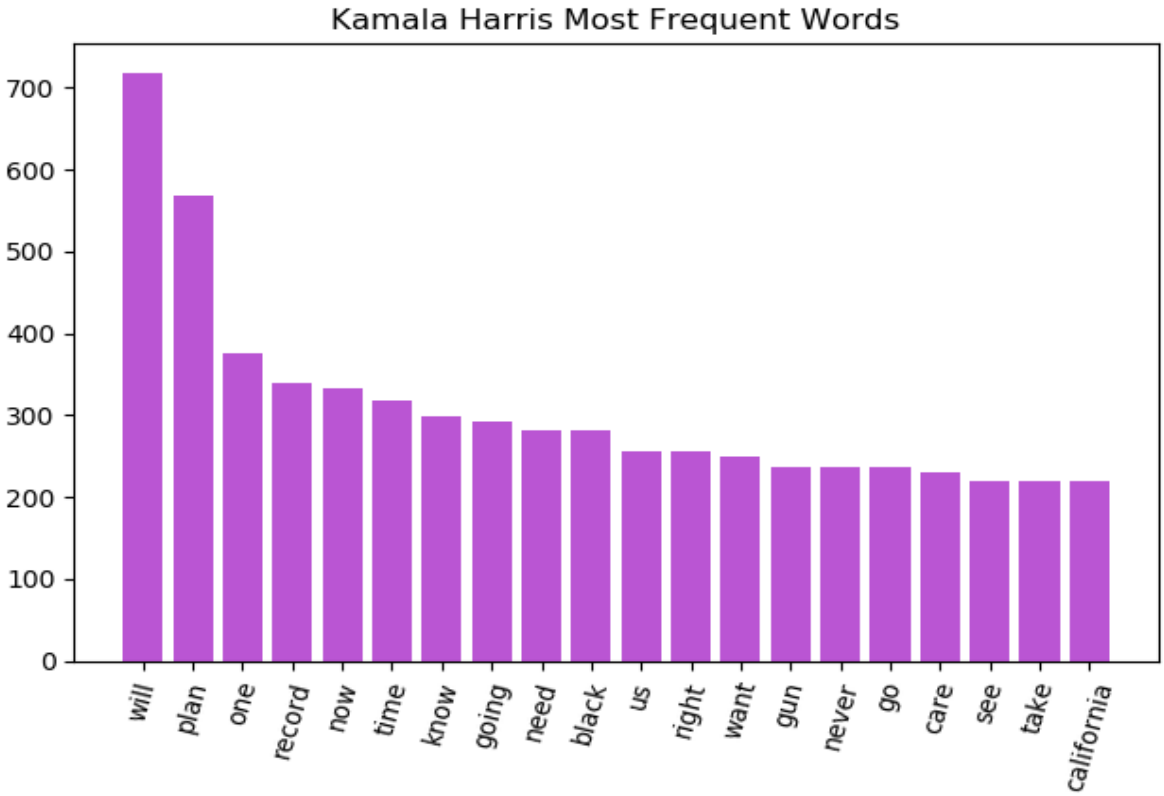
1. Data Analysis

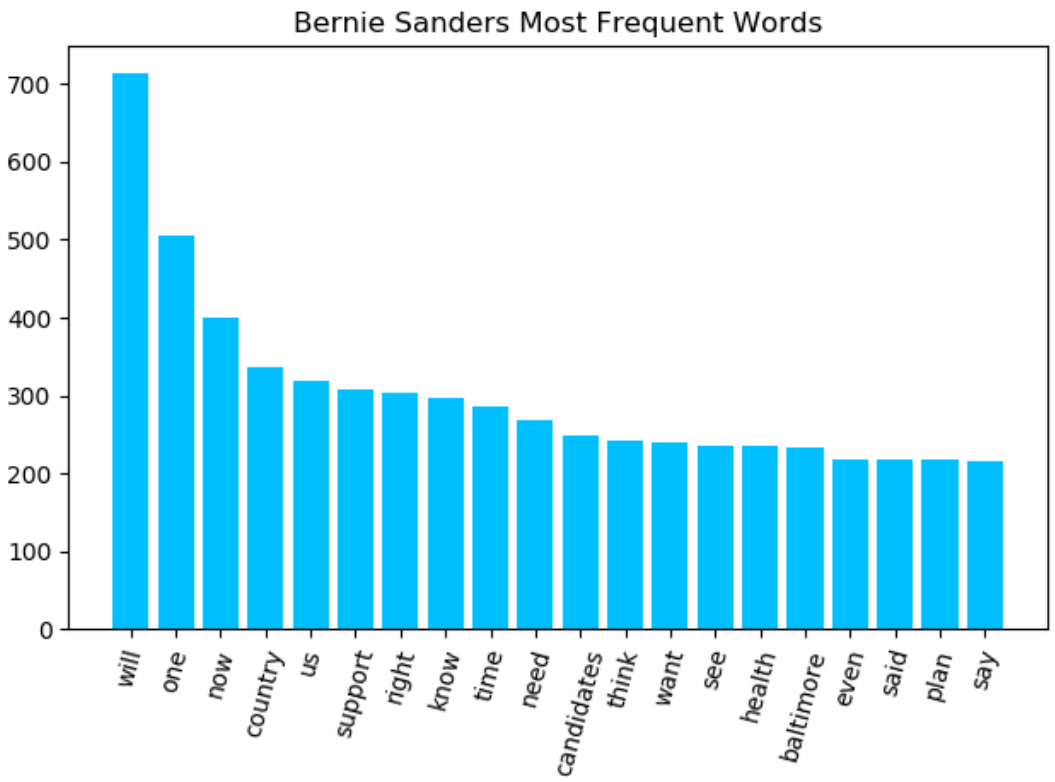
We begin with some explanatory analysis of the tweets. Below we show what the dataset looks like for tweets about Bernie Sanders. The datasets for the other candidates follow the same format. Tweets were scraped using the Octoparse web scraping tool.

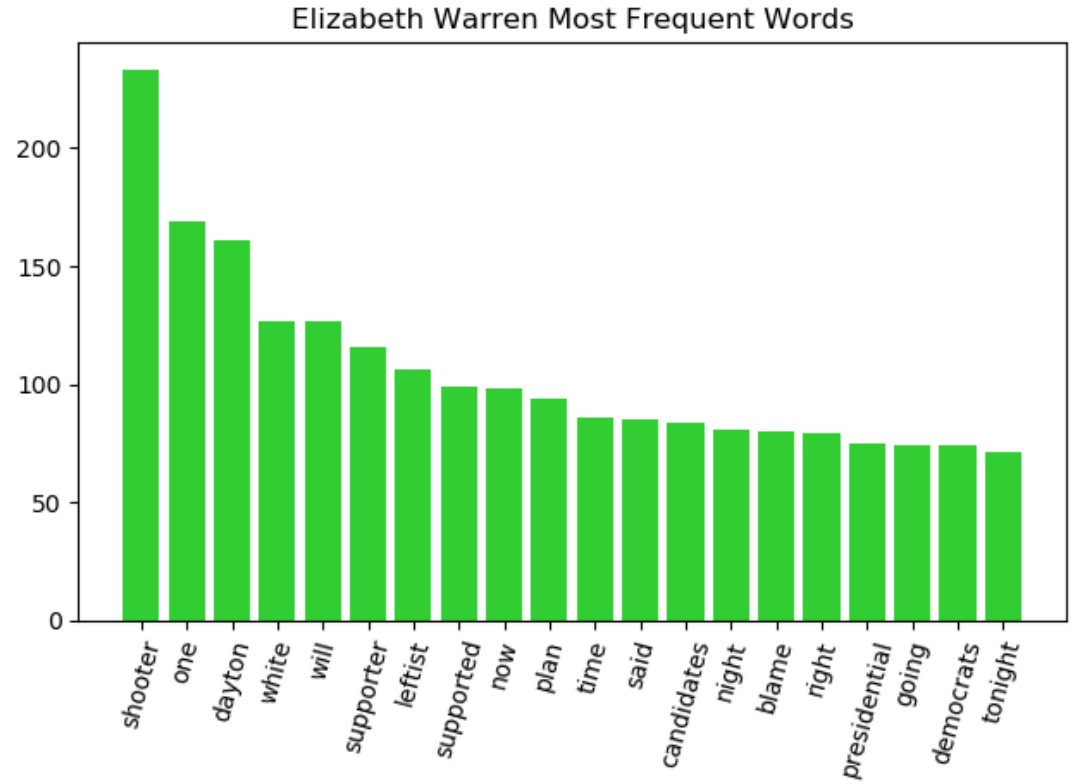
	Tweet	Reply	Date	Username	User
0	Bernie Sanders just went on Joe Rogan.. T...	nan	Aug 6	@DIEGOSILVA	ANDRE
1	We feel the heat, especially in Alaska. Sad..	Replying to @SenSanders	Aug 6	@Alaska4Bernie	Alaska for Bernie
2	Thanks for having Bernie Sanders on your ..	Replying to @joerogan	Aug 6	@philosophrob	Rob
3	Holy shit Joe Rogan got Bernie Sanders o...	nan	Aug 6	@hack_attack96	LJ.
4	We also can not afford to let automakers ..	Replying to @SenSanders	Aug 6	@totalcoverage	Total Coverage
5	Seriously @BernieSanders? I had higher h...	nan	Aug 6	@KatieLujan4	Katie

We can start by examining the most frequent words that were used in the tweets about each candidate:

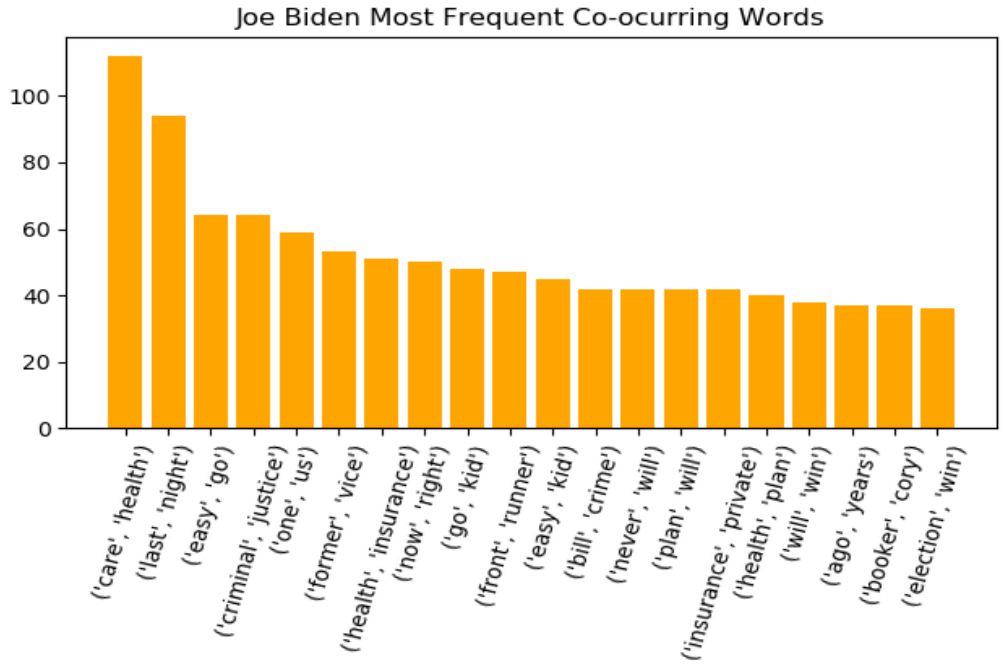

Joe Biden Most Frequent Words
300
250
200
150
100
50
0
will
one
now
said
time
know
plan
years
go
says
last
us
night
candidates
going
right
say
think
need
new

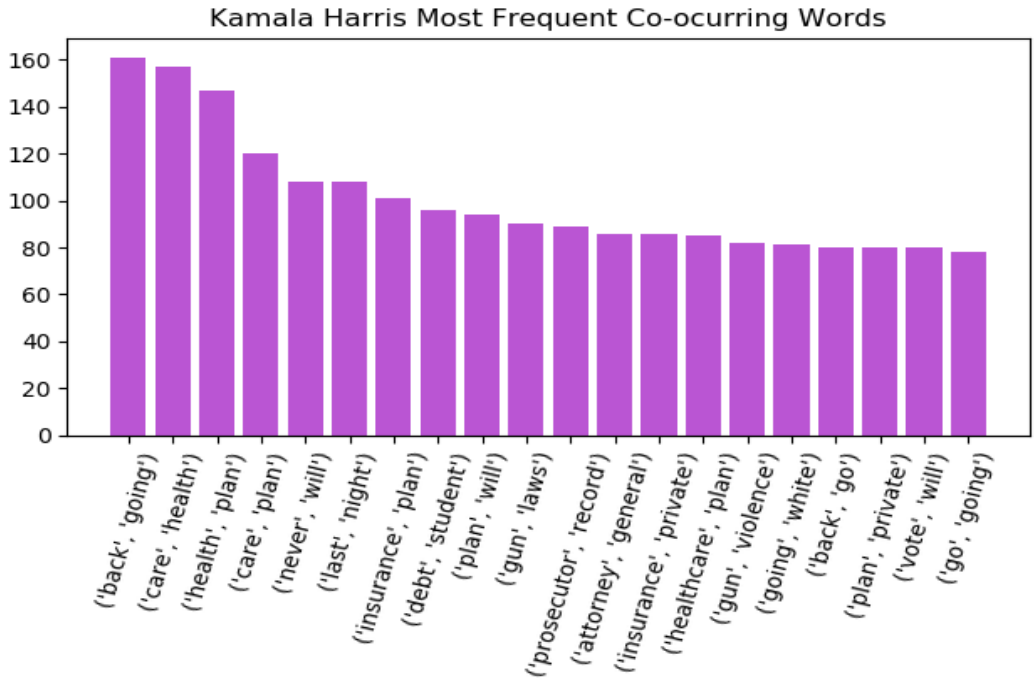

Kamala Harris Most Frequent Words
700
600
500
400
300
200
100
0
will
plan
one
record
now
time
know
going
need
black
us
right
want
gun
never
go
care
see
take
california

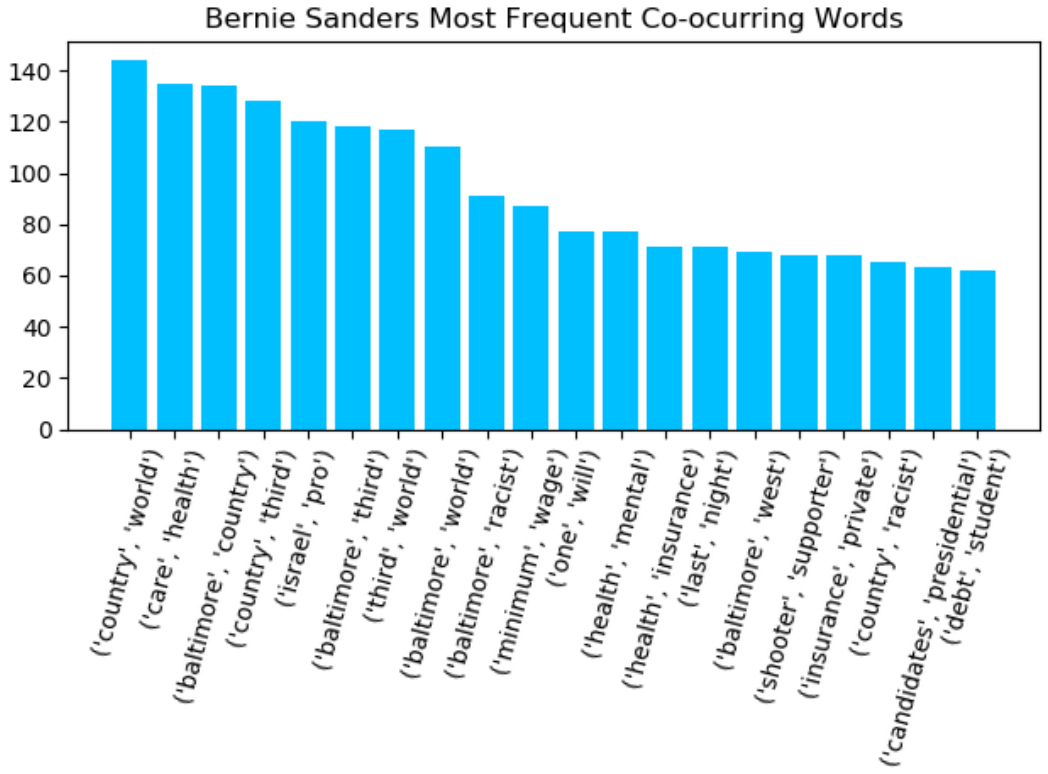

Bernie Sanders Most Frequent Words
700
600
500
400
300
200
100
0
will
one
now
country
us
support
right
know
time
need
candidates
think
want
see
health
baltimore
even
said
plan
say

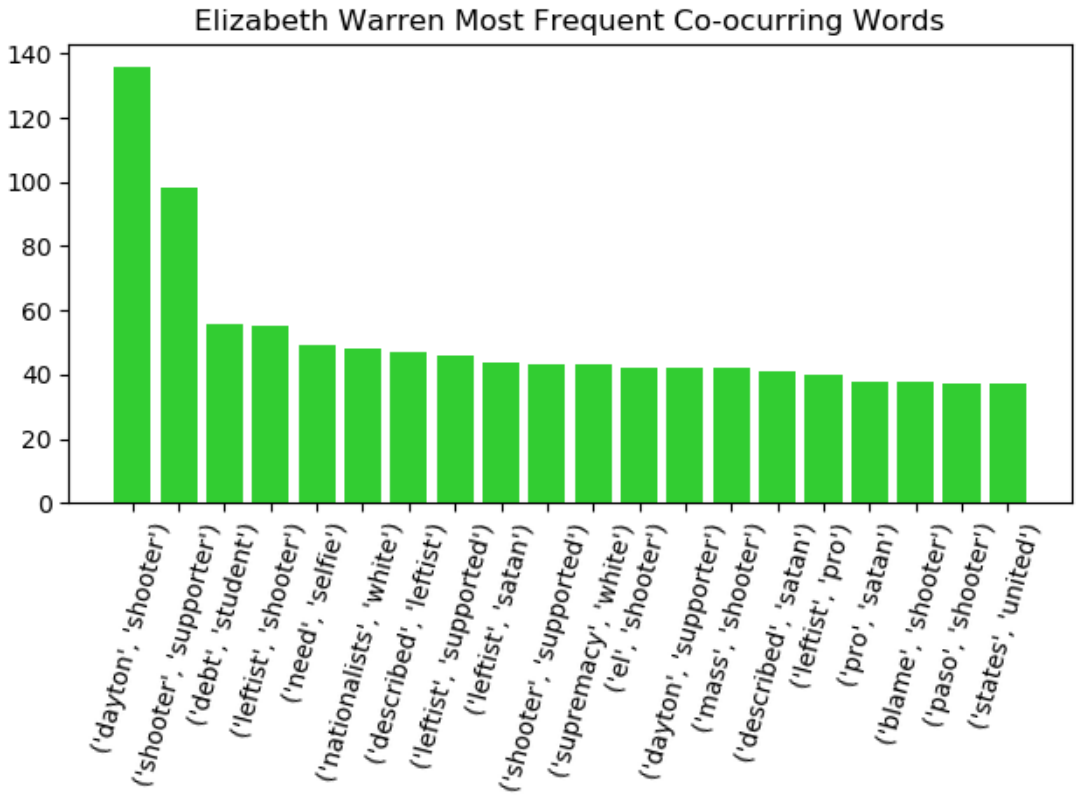

Elizabeth Warren Most Frequent Words
200
150
100
50
0
shooter
one
dayton
white
will
supporter
leftist
supported
now
plan
time
said
candidates
night
blame
right
presidential
going
democrats
tonight

For most of the candidates, the most frequent words are what you would expect of a candidate running for president (words like “plan”, “campaign”, “record”, and “support”. The only thing of note is in the most frequent words in the Elizabeth Warren dataset. Her most frequent word was “shooter” because of the large amount of news articles coming out stating that the perpetrator of a mass shooting in Dayton, Ohio was allegedly a supporter of Elizabeth Warren. We can also look at the words that co-occur most in the datasets:






Bernie Sanders Most Frequent Co-ocurring Words
140
120
100
80
60
40
20
0
('country', 'world')
('care', 'health')
('baltimore', 'country')
('country', 'third')
('israel', 'pro')
('baltimore', 'third')
('third', 'world')
('baltimore', 'world')
('baltimore', 'racist')
('minimum', 'wage')
('one', 'will')
('health', 'mental')
('health', 'insurance')
('last', 'night')
('baltimore', 'west')
('shooter', 'supporter')
('insurance', 'private')
('country', 'racist')
('candidates', 'presidential')
('debt', 'student')


Elizabeth Warren Most Frequent Co-ocurring Words
140
120
100
80
60
40
20
0
('dayton', 'shooter')
('shooter', 'supporter')
('debt', 'student')
('leftist', 'shooter')
('need', 'selfie')
('nationalists', 'white')
('described', 'leftist')
('leftist', 'supported')
('leftist', 'satan')
('shooter', 'supported')
('supremacy', 'white')
('el', 'shooter')
('dayton', 'supporter')
('mass', 'shooter')
('described', 'satan')
('leftist', 'pro')
('pro', 'satan')
('blame', 'shooter')
('paso', 'shooter')
('states', 'united')

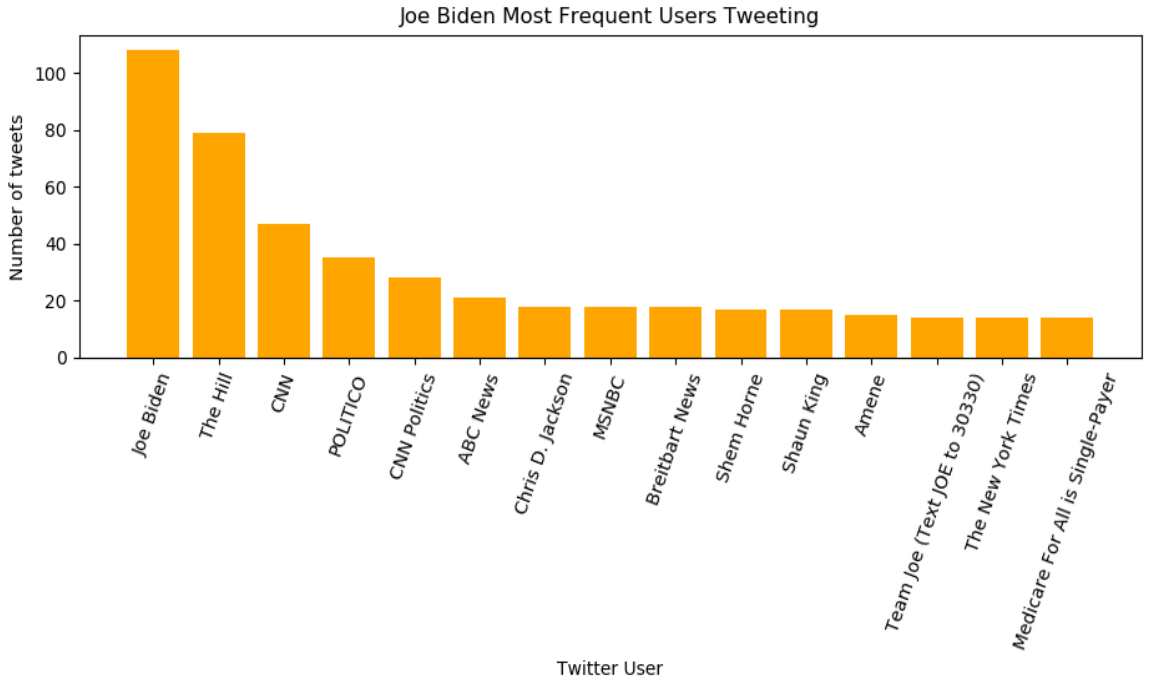
Looking at the most frequent co-occurring words has given us a lot more information about what topics are discussed when talking about the four candidates. The most notable for Joe Biden are ('care','health'), ('criminal','justice') and ('front','runner'). Health care has been the central issue in the election so far, so the topic is likely to discussed along with each candidate Furthermore, during this time period Joe Biden released his criminal justice reform plan, so there were likely many tweets referencing it. Additionally, two national polls were released during this period where Biden was shown as the front runner for the Democratic nomination, explaining the large occurrence of the ('front','runner') pair

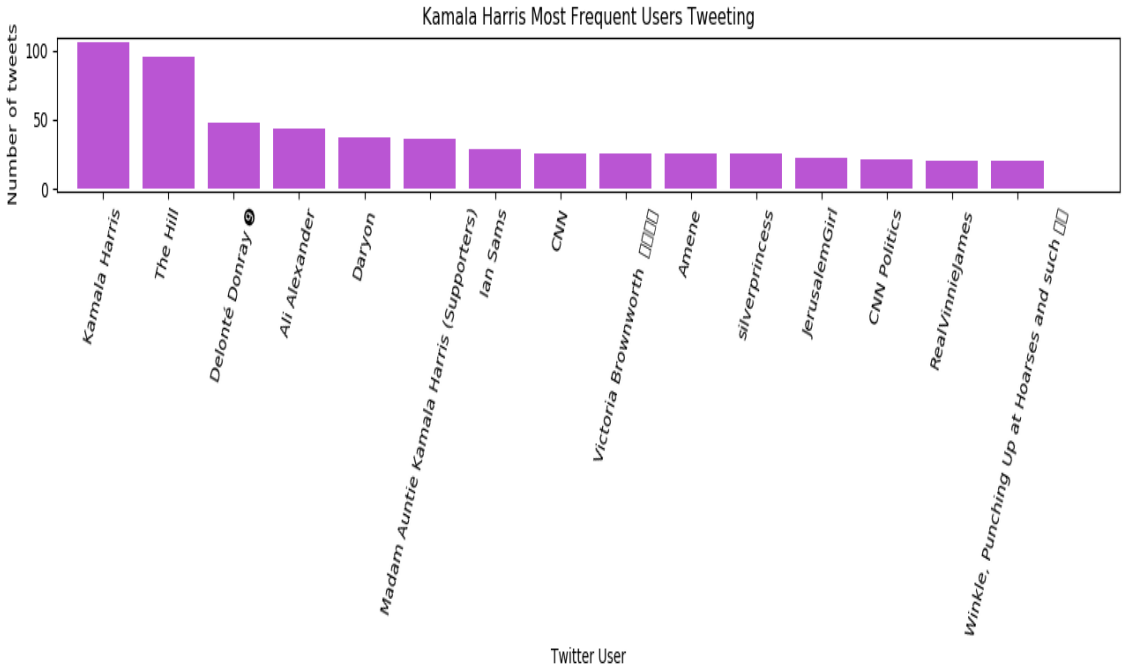
For Kamala Harris, the most discussed topics were ('care','health'), ('care','plan') and ('prosecutor','record'). Prior to the debate, Kamala Harris had just released her comprehensive health care plan. Additionally, many people were questioning her record from when she was a prosecutor in California.

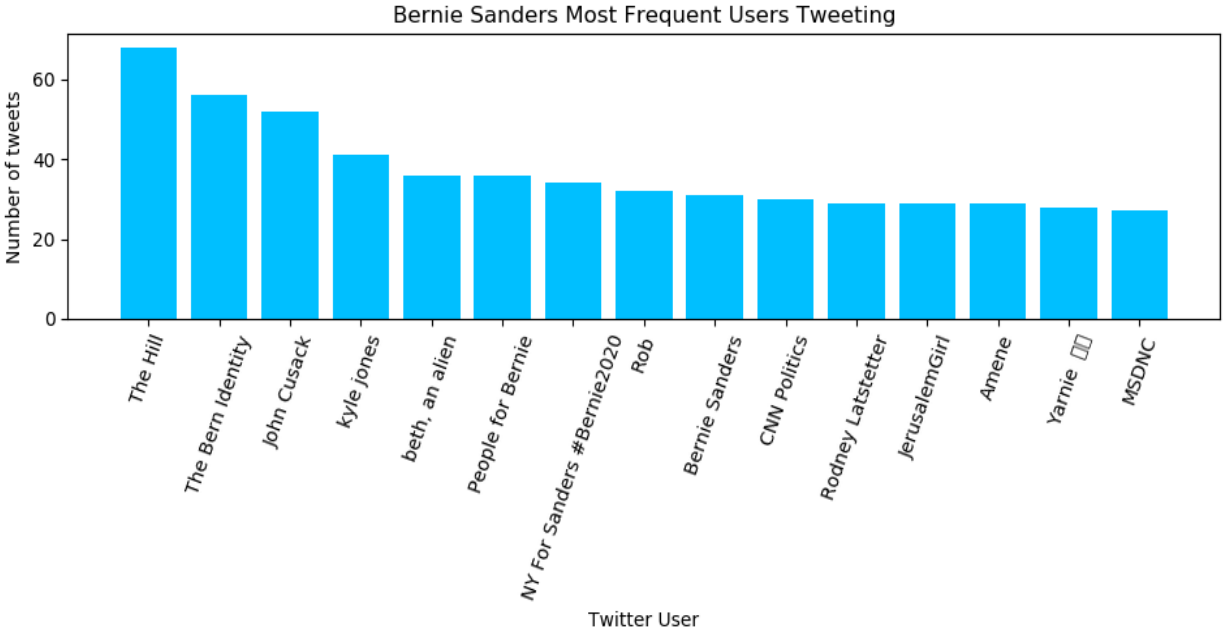
Bernie Sanders' most discussed topics were ('care','health'), ('israel','pro'), and pairs involving 'baltimore'. Bernie has made health care the central issue of his campaign, so it is no surprise that it is one of the most discussed topics in relation. Furthermore, during this time period Bernie Sanders was discussing the problems in Israel and his plans for the issue. There were also many news stories about quotes made by Sanders on the poverty and crime rates in Baltimore.

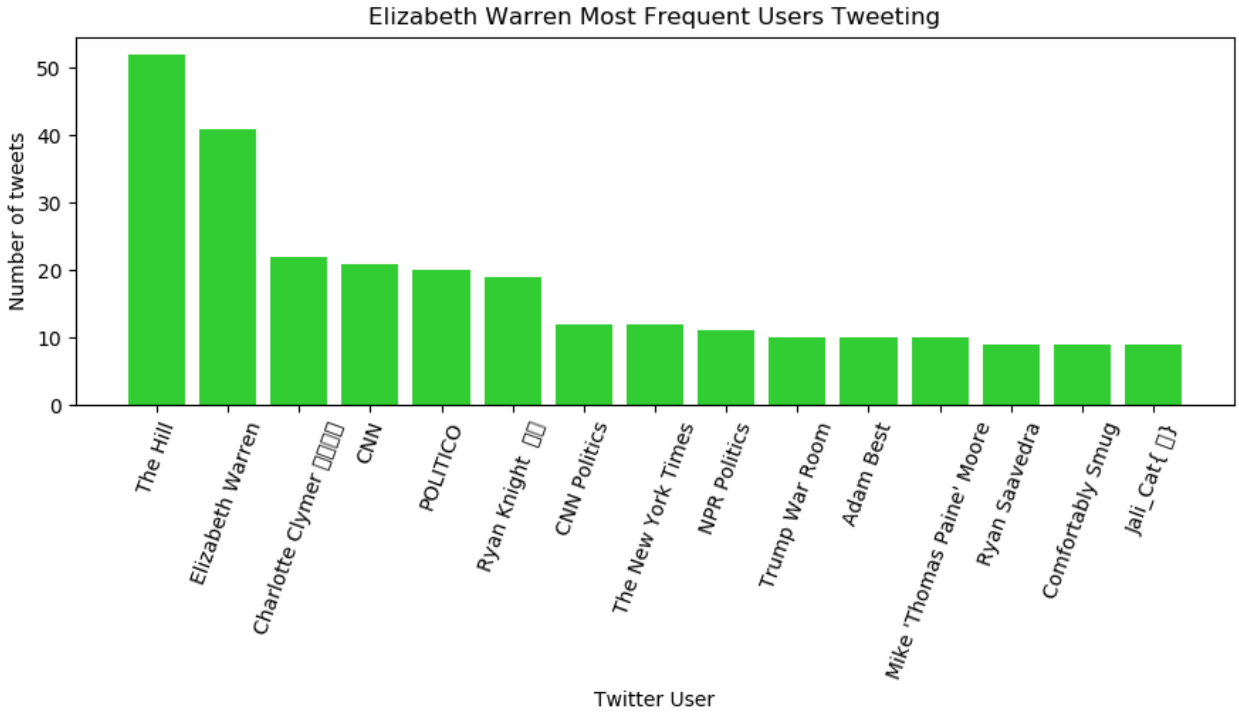
Elizabeth Warren's tweets seem to have been dominated by the stories on the Dayton Shooter, as well as her student debt relief plan. During this time a mass shooting took place in Dayton Ohio, and many news articles came out alleging that the shooter was a supporter of Warren.

We can also look at the users most frequently tweeting about each candidate during this time period:




Kamala Harris Most Frequent Users Tweeting
Number of tweets
100
50
0
Kamala Harris
The Hill
Delonté Donray
Ali Alexander
Daryon
Madam Auntie Kamala Harris (Supporters)
Ian Sams
CNN
Victoria Brownworth
Amene
silverprincess
JerusalemGirl
CNN Politics
RealVinnieJames
Winkle. Punching Up at Hoarses and such
Twitter User


Bernie Sanders Most Frequent Users Tweeting
Number of tweets
60
40
20
0
The Hill
The Bern Identity
John Cusack
Kyle jones
beth, an alien
People for Bernie
NY For Sanders #Bernie2020
Rob
Bernie Sanders
CNN Politics
Rodney Latstetter
JerusalemGirl
Amene
Yarnie
MSDNC
Twitter User



For most of the candidates, the most frequent users are the candidates themselves, as well as The Hill, CNN, and other major news outlets. Bernie Sanders has the most users who are not news organizations.

I will now discuss the models I trained to perform sentiment analysis on the tweets for each candidate and present their predictions.

2. Unsupervised Approach

I began by trying an unsupervised machine learning method to predict the sentiment of these tweets. The method I followed is the one presented by Peter Turney in the below paper:

https://www.aclweb.org/anthology/P02-1053

In this approach, semantic orientation is determined by examining how "close" words in the tweet are to positive or negative words. The Pointwise Mutual Information (PMI) between two words is calculated between each word in our tweet vocab set and a list of positive and negative words.

$$PMI(word_1, word_2) = log_2\left[\frac{p(word_1 \,\&\, word_2)}{p(word_1)\,p(word_2)}\right]$$

Where $p(word_1 \& word_2)$ is the probability that $word_1$ and $word_2$ will co-occur, and $p(word_1)p(word_2)$ is the probability that the two words will occur if they are statistically independent. If we divide these values, we are able to get a measure of the degree of **statistical dependency** between the words. Taking the log of this quantity represents how much information we gain about the presence

of one of the words when the other is present in a document. The semantic orientation is then calculated as

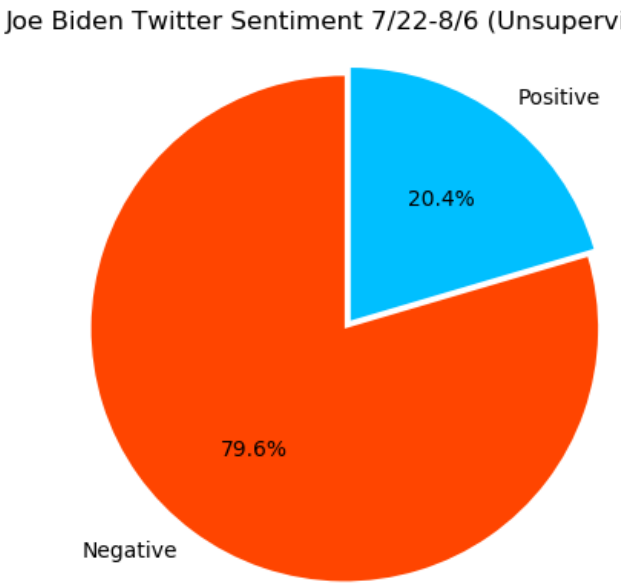
$$SO(phrase) = PMI(phrase, posWord) - PMI(phrase, negWord)$$

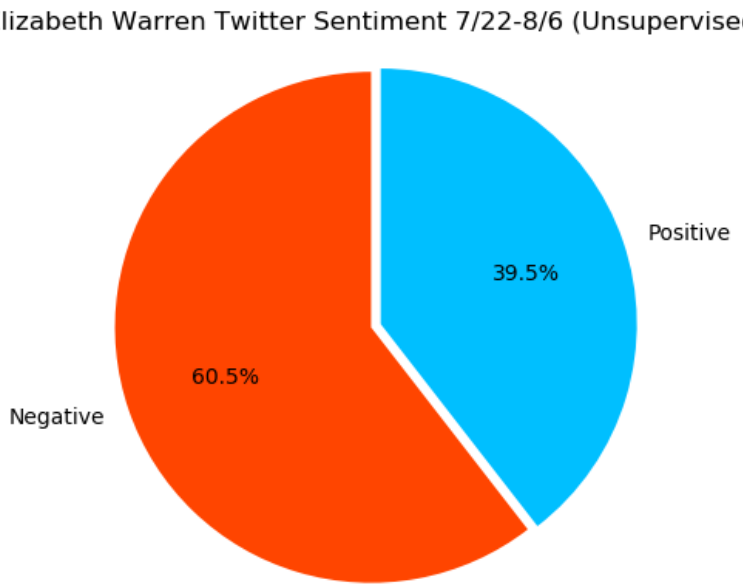
This quantity is summed over all the positive and negative words in our list to determine which group of words our phrase is most similar to. I obtained this list of positive and negative words from the below paper:

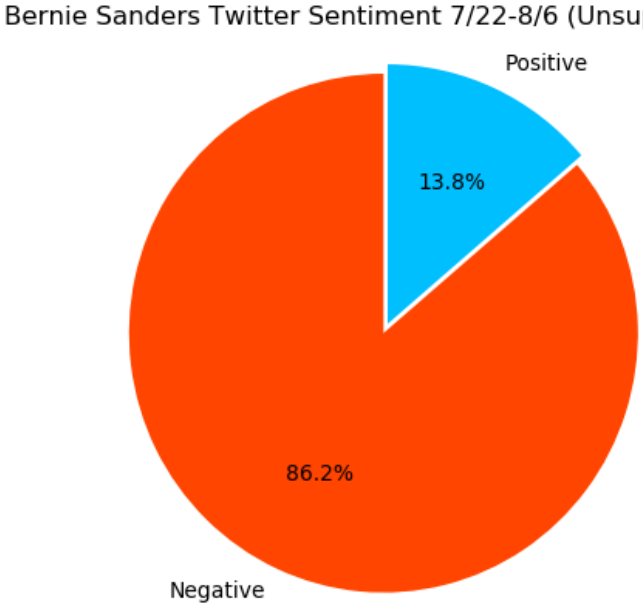
Minqing Hu and Bing Liu. "Mining and Summarizing Customer Reviews."
Proceedings of the ACM SIGKDD International Conference on Knowledge
Discovery and Data Mining (KDD-2004), Aug 22-25, 2004, Seattle,
Washington, USA

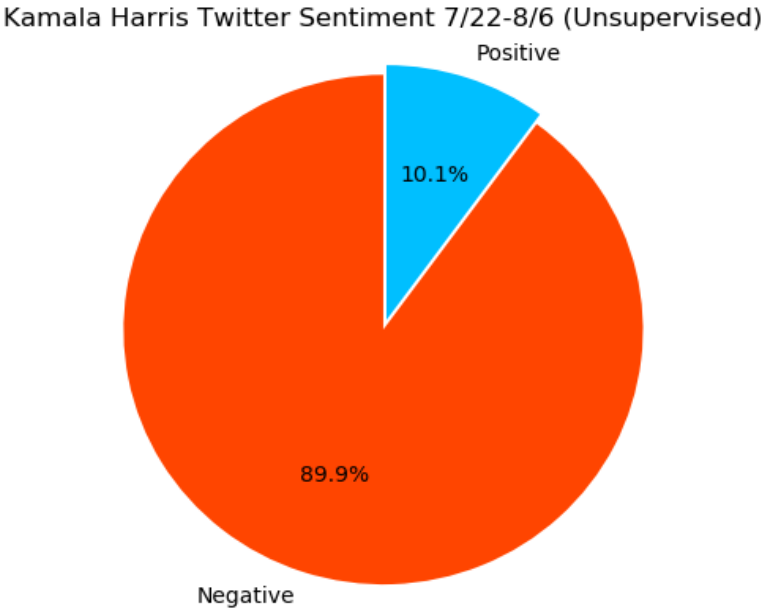
We can now look at the predictions of the unsupervised model on our datasets on the next page.

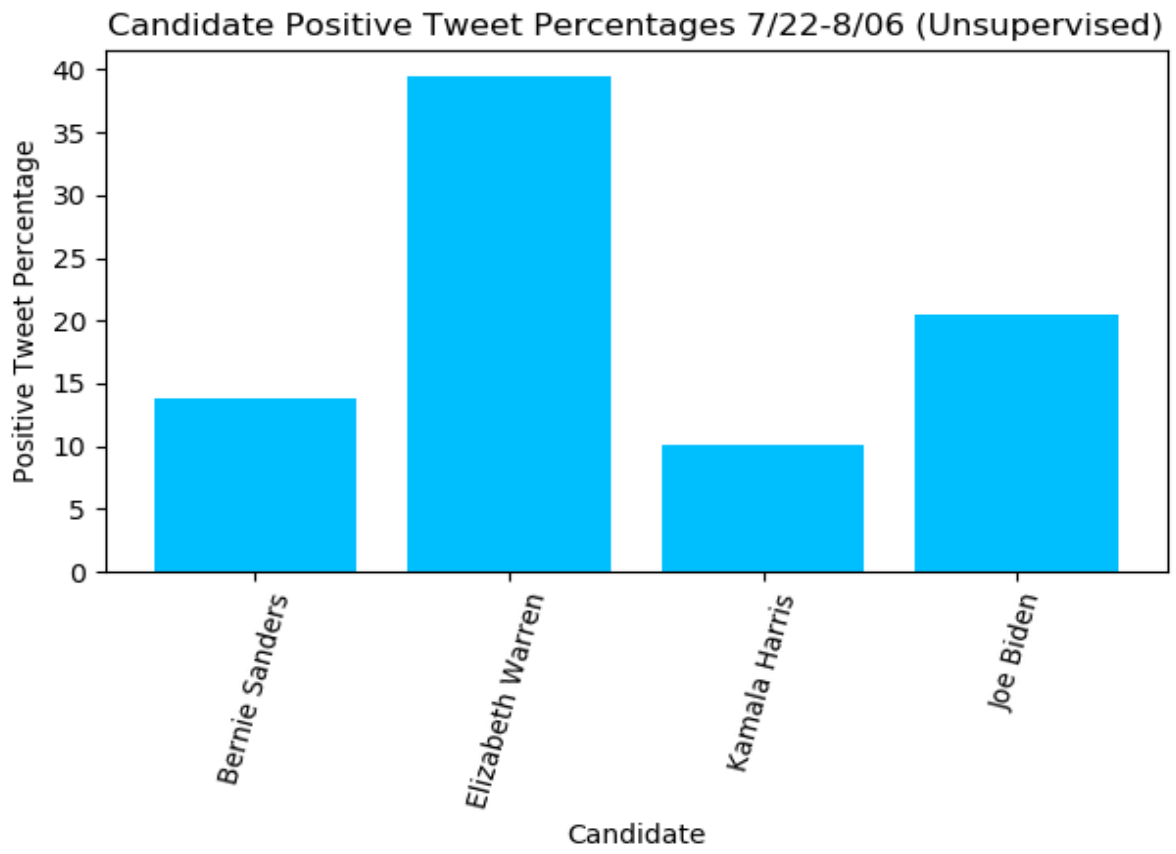
We can take a look at the overall percentage of positive and negative tweets throughout the entire time period:


Joe Biden Twitter Sentiment 7/22-8/6 (Unsupervised)
Positive
20.4%
79.6%
Negative


Elizabeth Warren Twitter Sentiment 7/22-8/6 (Unsupervised)
Positive
39.5%
60.5%
Negative

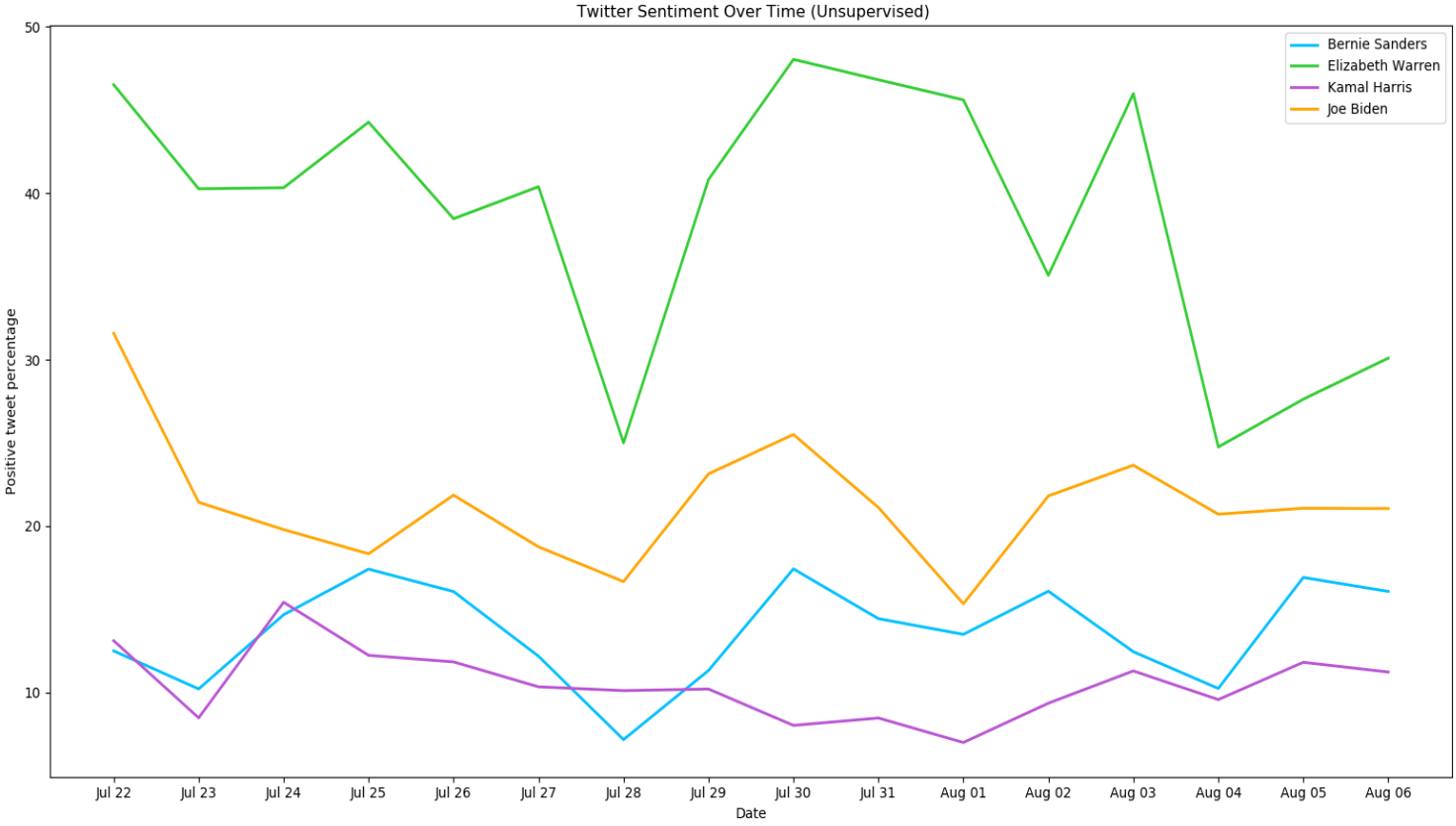

Bernie Sanders Twitter Sentiment 7/22-8/6 (Unsupervised)
Positive
13.8%
86.2%
Negative


Kamala Harris Twitter Sentiment 7/22-8/6 (Unsupervised)
Positive
10.1%
89.9%
Negative



We can see that Elizabeth Warren and Joe Biden had the highest percentages of positive tweets during the whole time period, with Elizabeth Warren being significantly higher than the other 3 candidates.

Let's look at how tweet sentiment changed throughout the period:



This graph gives us a clearer picture of how the sentiment changed over time, as we are able to see that Elizabeth Warren's positive tweet percentage has consistently been higher throughout the period. Joe Biden's positive tweet percentage has also consistently been higher than the other 2 candidates. The second Democratic Debate took place on July 30th and 31st, so looking at sentiment on August 1st should give us an idea of twitter's opinion on each candidate's performance. Although Biden had the 2nd highest positive tweet percentage in the whole period, we can see that this debate actually hurt him relatively more when compared to other candidates. Kamala Harris seems to be the least affected by the debate, whereas other candidate's positive tweet percentages took a dip after the debate. We can also observe that Elizabeth Warren, Joe Biden, and Bernie Sanders' positive tweet percentages were the most volatile.

Although this is a useful method for conducting sentiment analysis when choosing to use an unsupervised method, it can oversimplify things and many of the nuances of determining a word's semantic orientation are lost. In the next section, I will be training a deep learning classifier for this problem. However, the unsupervised approach is a fairly good option for when there is a lack of labeled training data available.

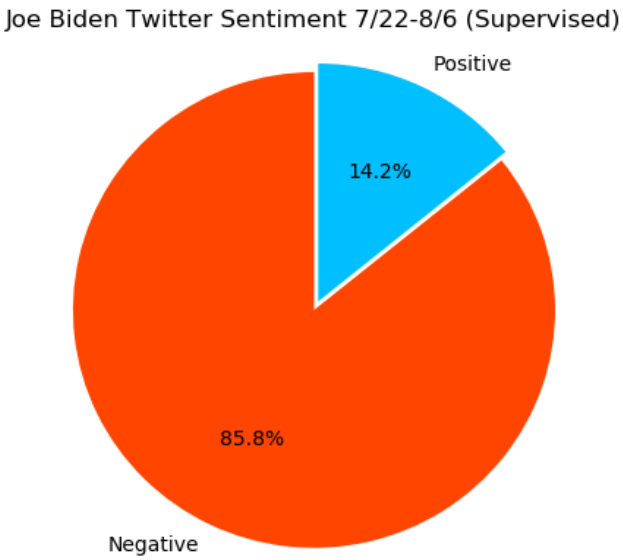
3. Supervised Approach

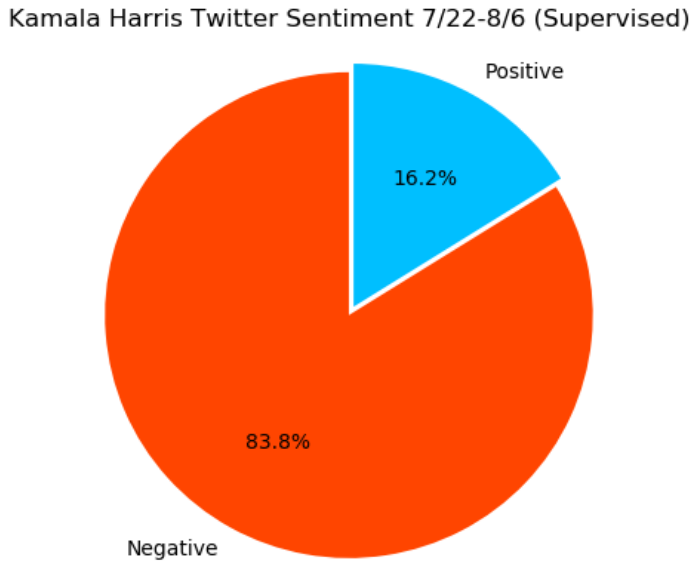
In order to produce a model that would be able to recognize the complexities in the tweets better, I trained a 1 layer Recurrent Neural Network (RNN) using a subset of the Sentiment140 dataset containing 1.6 million tweets found on Kaggle. This is a dataset containing tweets that have been labeled as either "Positive" or "Negative". Having tweets that have a "Neutral" label would have been useful for this problem, but I found a lack of publicly available labeled data, which is a common problem when performing sentiment analysis. Because of limited computing power, I used around 400,000 samples from this dataset.

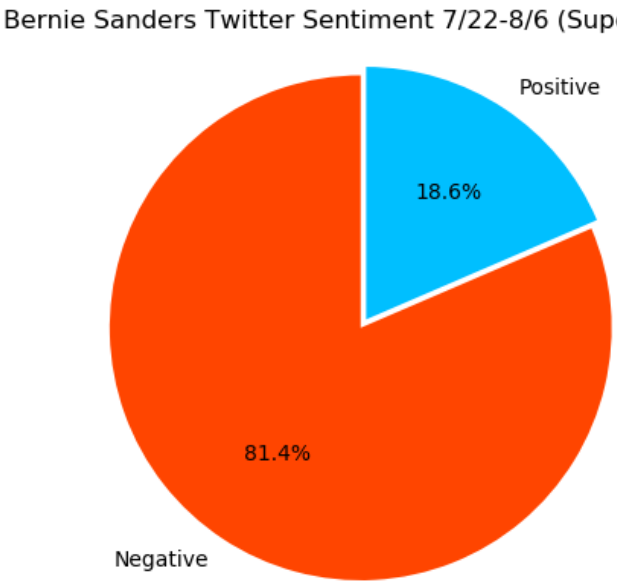
An RNN was chosen for this application because of its better performance in predicting sequence-based data compared to other deep learning architectures. All the tweets in the dataset were padded to be 28 words long, and the RNN architecture contained a learned Embedding layer with 150 features. An LSTM unit with 50 neurons, 70% dropout percentage and L2 Regularization with λ=0.7 was used, as well as a dense final layer. After training, the RNN was able to achieve 84% validation accuracy.

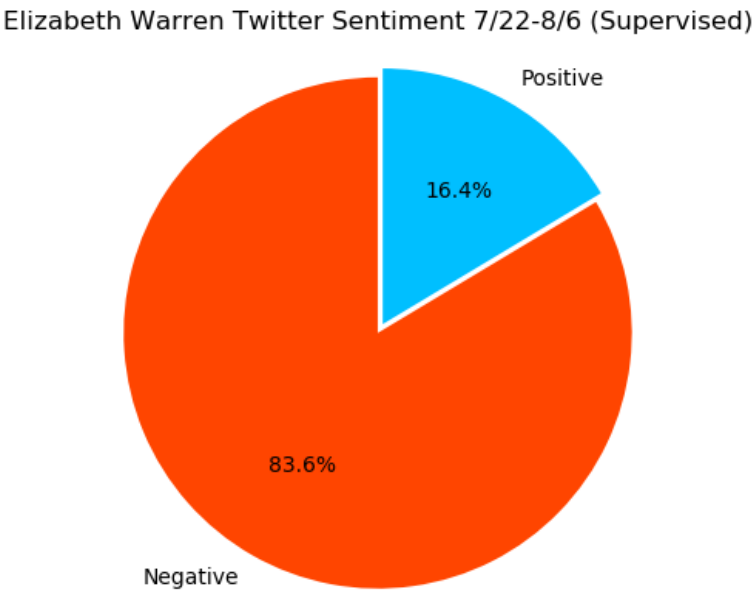
Before training the RNN model I attempted to use traditional machine learning algorithms such as LinearSVC and Naïve Bayes but was only able to achieve maximum 73% validation accuracy.

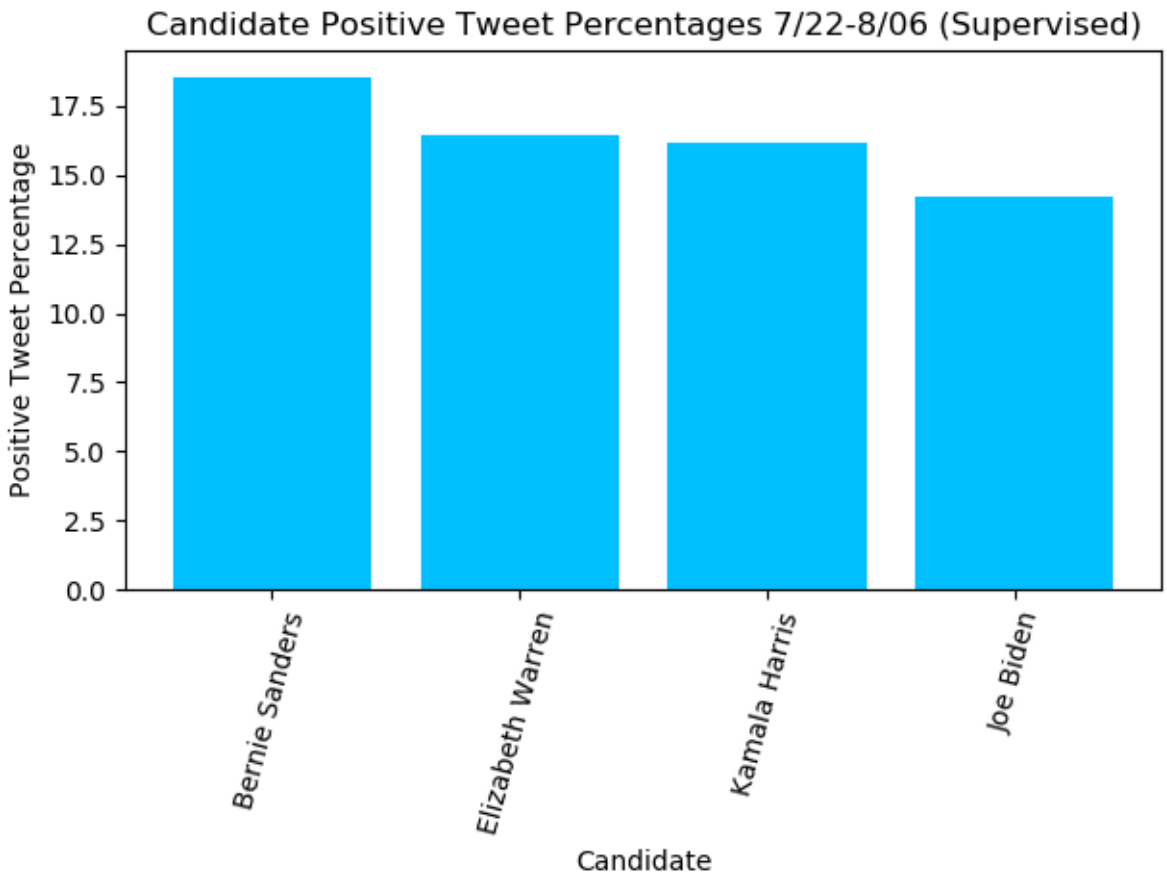
We can now look at the sentiment results returned by our supervised model:





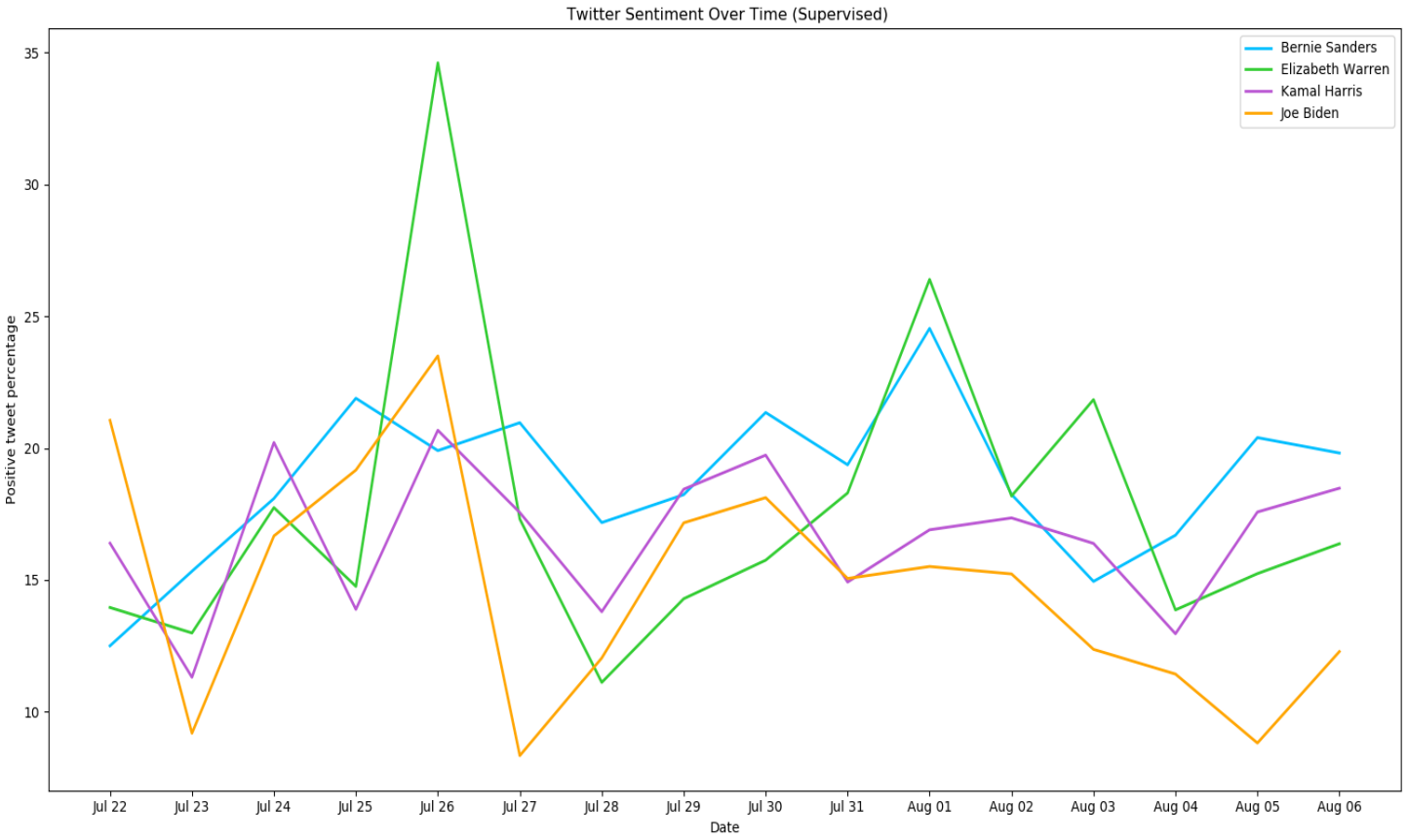




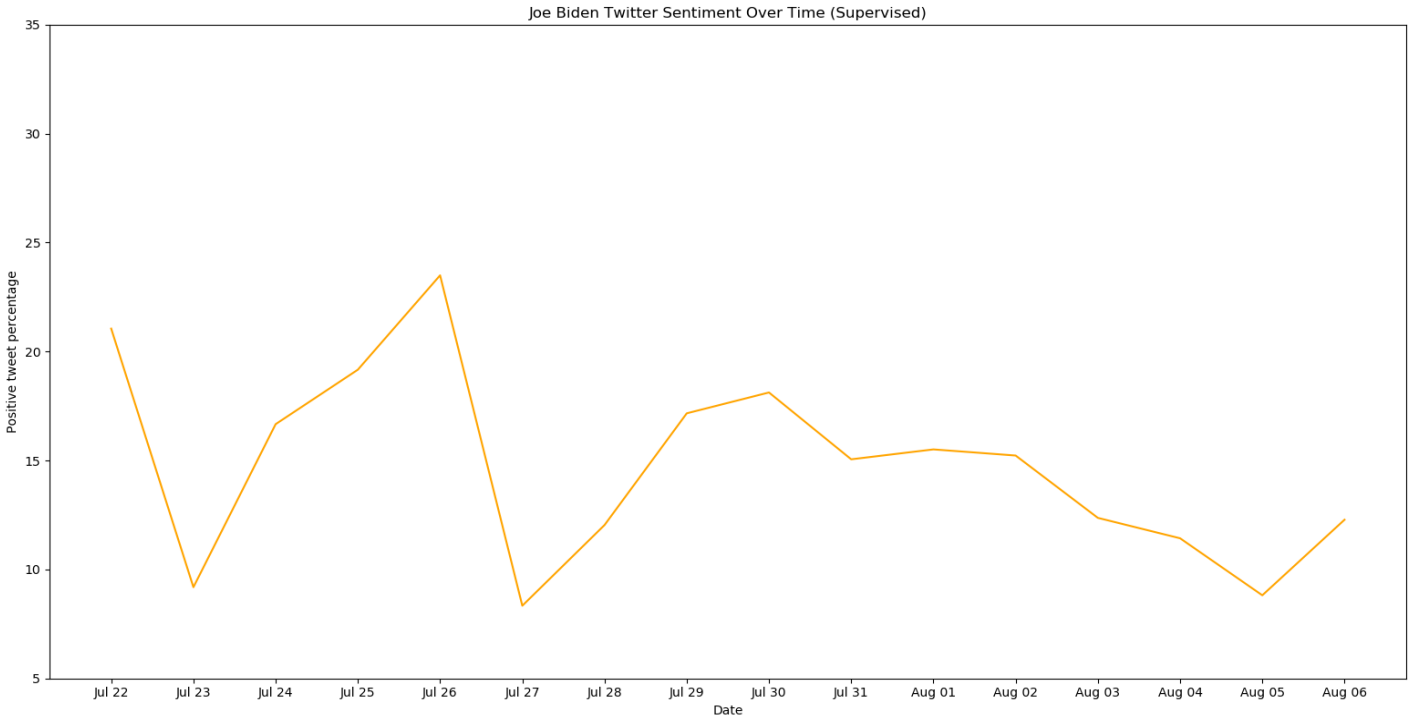


We can see a clear difference in the predictions of the unsupervised and supervised model. The supervised model has predicted positive tweet percentages to be very similar, which is more likely to be accurate for these four candidates, as it is unlikely there would be drastic differences in positive tweet percentages in this early stage of the election, and given the support that all 4 candidates have. In our supervised model, Bernie Sanders has the largest positive tweet percentage, although the difference is not extremely significant.

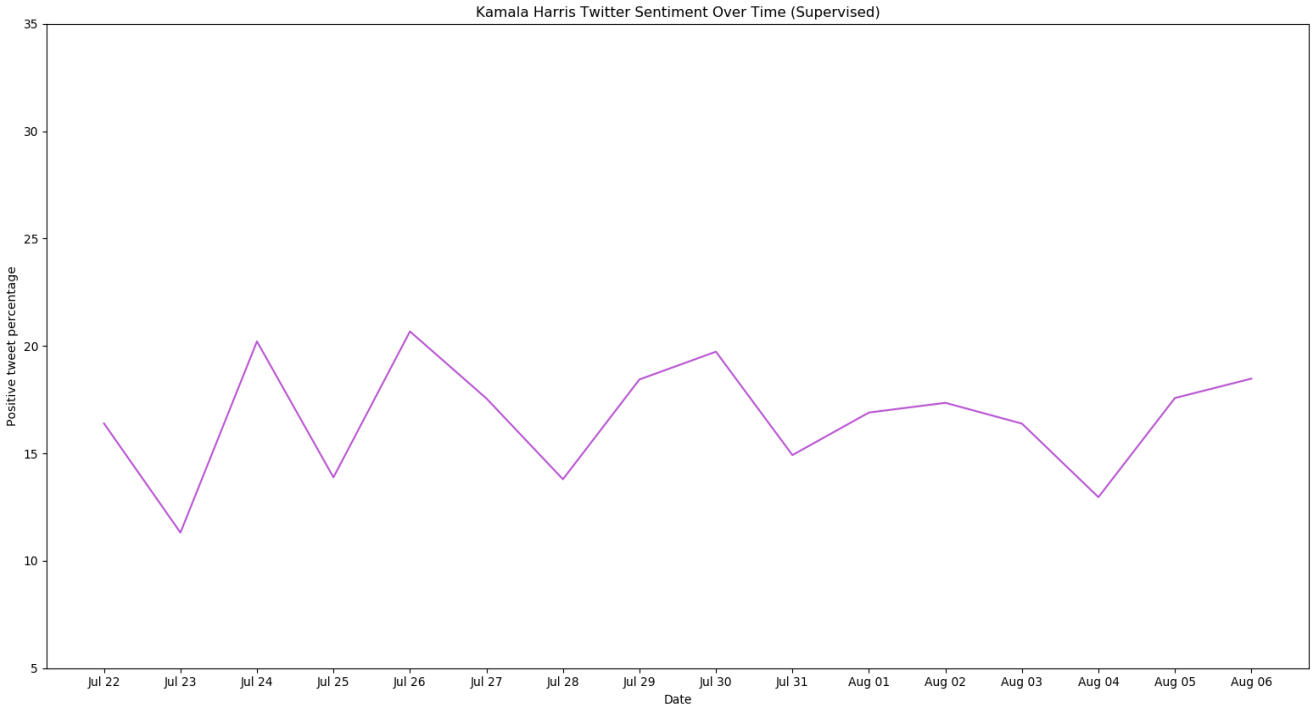
Let's take a look at how the sentiment changes throughout the time period for the four candidates:



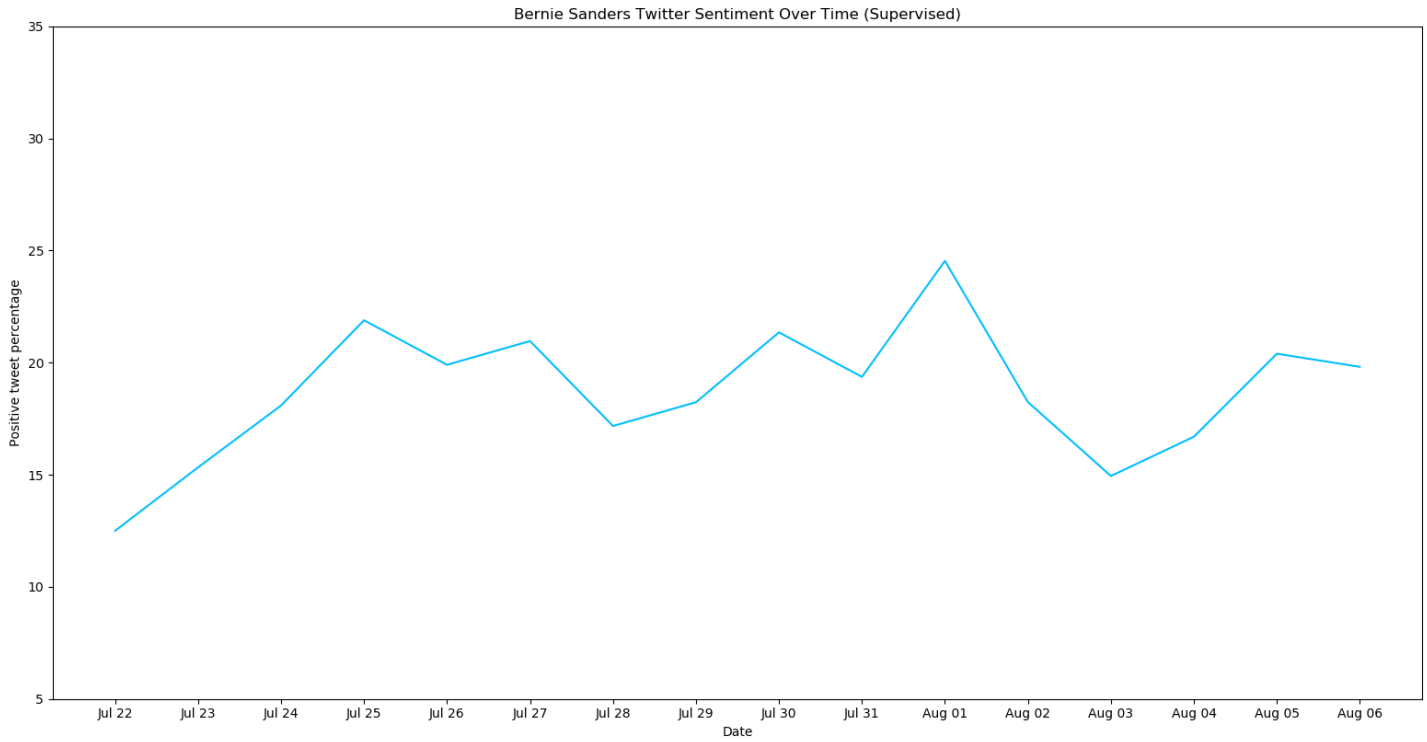
The sentiment over time is also much more similar between candidates as compared to our unsupervised model. From the above graph, it seems that Bernie Sanders has had the most consistently high positive tweet percentage in the period, except for the extremely sharp rise for Elizabeth Warren on July 26th. The unsupervised and supervised models both seem to follow the various news events that took place during the time period. I will now break these down by looking at the sentiment over time for each candidate in detail.



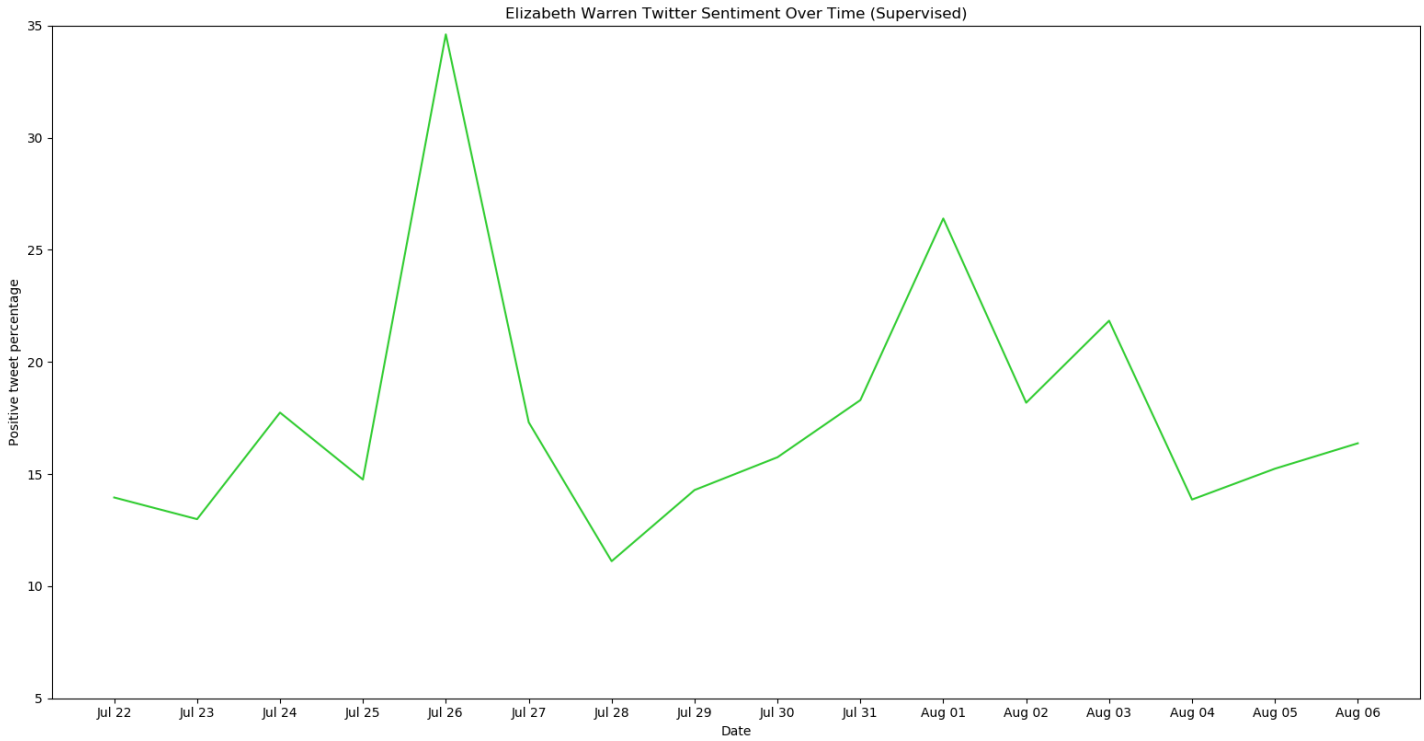
We can observe that Joe Biden's positive tweet percentage ranged from about 8% to 23% during this time period. We notice that between July 23rd and July 26th, his positive tweet percentage increased significantly. The news surrounding Biden during this time seems to be fairly mixed, but the bump may be due to a new poll released on July 26th showing that Joe Biden could have a chance at beating the incumbent president Donald Trump in Ohio during the general election. But then the positive tweet percentage drops very sharply on July 27th, when another candidate, Cory Booker, expressed criticism of Biden's criminal justice reform plan. After recovering, the positive tweet percentage seems to stabilize some over the next week. This period included the second primary debate (Joe Biden debated on July 31st), so this might be an indication that Biden's performance in the debate did not have a large impact on voter sentiment, judging based on twitter. However he faces another decrease in positive sentiment on August 5th, coinciding with Biden's advocation for a federal gun buyback program.



Kamala Harris's twitter sentiment seems to be the most stable of the candidates during this time period. The positive tweet percentage ranges from around 11% to around 21%. For most of this period, her positive tweet percentage seems to oscillate from day to day, so I will only analyze the days around the debate. Kamala Harris debated on the second night of the debate, July 31st. The day after, August 1st, her positive tweet percentage increased very slightly, and remained relatively stable until dipping some on August 4th. This might be an indication that her debate performance did not do much to affect her twitter sentiment. There doesn't seem to be one clear reason for this dip in positive tweet percentage.



Bernie Sander's positive tweet percentage during this time period ranged from around 12.5% to around 25%. From July 22nd to 25th, there was a steady increase in positive tweet percentage. There is not one clear reason for this increase, but Sanders did make an appearance on Jimmy Kimmel on July 25th, and there was also a poll released by the Economist showing that Sanders and another candidate, Andrew Yang, had the most support from voters who voted for Donald Trump in the 2016 general election. Sanders' positive tweet percentage remains fairly stable, until we notice a sharp peak on August 1st, the day after both nights of the debate had concluded. Compared to Kamala Harris and Joe Biden, Bernie did seem to be helped by his debate performance. However there is a steep drop off that takes place the next two days. This drop off coincides with a Washington Post editorial that came out on August 1st sharply criticizing both Bernie Sanders and Elizabeth Warren's health care proposals. The article stated that their proposals "do not meet a baseline degree of factual plausibility", and this criticism by one of the country's largest news organizations could be the cause of the dip in Sanders' positive tweet percentage.



Elizabeth Warren's twitter sentiment during this period is the most volatile of the 4 candidates, but she does have the highest positive tweet percentages when compared to the other candidates. She experiences a sharp peak in her twitter sentiment on July 26th, which is the same day that she announced she received 1 million donations solely from grassroots donors, a major milestone for her campaign. However after this large increase, her positive tweet percentage dips down to its lowest point during this period, on July 28th. This coincides with allegations directed at her campaign for allegedly exploiting free labor by offering unpaid fellowships in her campaign. However her positive tweet percentage peaks again the day after the debates had concluded. The Washington Post editorial released on August 1st did not seem to coincide with a positive tweet percentage decrease as it did for Sanders. Warren did have a slight decrease between the debate and August 4th, as her positive tweet percentage stabilized some.

4. Conclusion

Through this analysis of tweet sentiment of the top 4 candidates for the 2020 Democratic nomination, we were able to gain insight into the sentiment on each candidate in the current stage of the election. We were able to track sentiment changes coinciding with the second Democratic primary debate, as well as look at the issues most frequently discussed in relation to each candidate. By analyzing the most frequent words used when tweeting about the candidates, we were able to affirm the fact that health care is the central issue in this primary election. This was also clear from the amount of time spent discussing the issue during the debate itself. We were also able to track the progression of news events

during the time period and see how twitter sentiment changed in reaction to the events. The 2020 election is still in very early stages so the standings of each candidate could change significantly in the months before the primary election, but we were able to see how volatile twitter sentiment can be in such a short time period. We were also able to see how much of an effect a candidate's debate performance can have on the sentiment of voters.