Final Report

Rashed Aldulijan

rashedul@colostate.edu

DSCI-320

Colorado State University

Dec 10, 2022

Table of Contents

Introduction

Twitter's influence in the 2020 US Presidential Election has grown significantly. Based on tweets, debates have raged, and candidates have risen and fallen. In this project, I will generate word clouds for several possible presidential candidates from several articles and people's tweets, and based on that, I will decide who has more words matching between their articles and people's comments about them. Then I will classify people's comments by polarity, which means how positive and negative the tweet is, and subjectivity, which measures how subjective or objective the tweet is. Based on that, I will predict who is going to win between these 6 candidates. The 6 candidates are (D) Joe Biden, Gretchen Whitmer, and Gavin Newsom; (P) Ron DeSantis, Mike Pence, and Nikki Haley.

Collecting data

The first step on my project is to collect data. I collected several articles and speeches transcript for each candidate. Approximately, I collected 3000 words for each candidate. After that, I put these data into a single file for each candidate, so there is 6 files total.

Word Clouds

After collecting some data, now it is the time to generate word clouds. I used WordCloud function from wordcloud library to generate these word clouds. Here are the word clouds that I got.

After generating these word clouds, I used twitter API, which is a collection of programmatic endpoints that can be used to comprehend or participate in Twitter conversations, to generate word clouds for the last 100 tweets about each of the candidates, so 600 tweets total. The first thing I did before I generated the word clouds is I put people's tweets about these candidates into data frames with a hundred observations each. Here is an example of the data frames.

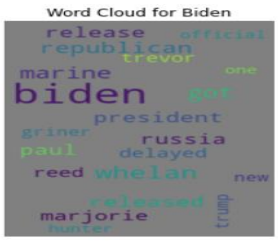
	Text
0	For those with their knickers in a twist about the Biden administration trading an arms dealer for Brittney Griner,… https://t.co/I0DxfQgQHq
1	RT @NoLieWithBTC: This is US Marine Trevor Reed, who Biden got released from Russia. His release was delayed by Republicans like Marjorie T…
2	RT @Blake_Allen13: Biden right now https://t.co/YV5sdYdzg5
3	RT @ChuckCallesto: BREAKING REPORT: Felony Arrest Warrant Issued For Trans nonbinary Biden Official Sam Brinton FOR ANOTHER ALLEGED THEFT…
4	RT @EndWokeness: Jill Biden urges Americans to get another booster before Christmas\n\n"Your president and I care about you and want to make…

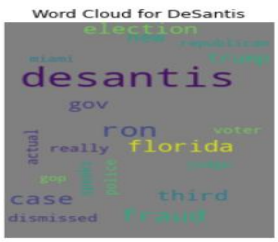
However, people tweets usually subjective. It includes emojes and punctuatuion and words that are useless for my analysis. Therefore, I decieded to clean their tweets from emojes and punctioation and some other things. Then, I added a column into the same data frame called

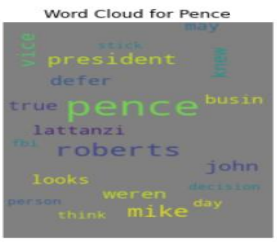
“cleaned tweets”. Here is an example on how the new data frames look like.

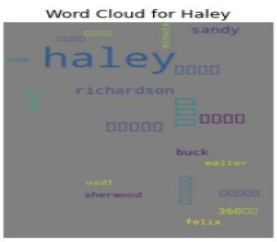
	Text	cleaned tweet
0	For those with their knickers in a twist about the Biden administration trading an arms dealer for Brittney Griner,… https://t.co/I0DxfQgQHq	for those with their knickers twist about the biden administration trading arms dealer for brittney griner
1	RT @NoLieWithBTC: This is US Marine Trevor Reed, who Biden got released from Russia. His release was delayed by Republicans like Marjorie T…	this marine trevor reed who biden got released from russia his release was delayed republicans like marjorie
2	RT @Blake_Allen13: Biden right now https://t.co/YV5sdYdzg5	biden right now
3	RT @ChuckCallesto: BREAKING REPORT: Felony Arrest Warrant Issued For Trans nonbinary Biden Official Sam Brinton FOR ANOTHER ALLEGED THEFT…	breaking report felony arrest warrant issued for trans nonbinary biden official sam brinton for another alleged theft
4	RT @EndWokeness: Jill Biden urges Americans to get another booster before Christmas\n\n“Your president and I care about you and want to make…	jill biden urges americans get another booster before christmas your president and care about you and want make

You can see from here that the new column is the same, but without useless words. After that, I generated word clouds for people tweets without the stop words, which is a library function that have more insignificant words, and here are how they look like.









Words Matching

When we compare these word clouds with the word clouds that are from articles that I collected we can see that a lot of the words look similar. However, we need to know the ratio of similarity between people’s tweets and articles, so we can guess who has higher popularity. To do this, I used token_sort_ratio() function from thefuzz library, this function calculates the Levenshtein distance similarity ratio (in percent) between the two strings. Here is what I got.

```
Similarity ratio for Biden:  53
Similarity ratio for Whitmer:  45
Similarity ratio for Newsom:  51
Similarity ratio for DeSantis:  44
Similarity ratio for Pence:  46
Similarity ratio for Haley:  37
```

We can see from this figure that Biden has the highest ratio, which is expected as he is the current president. We can conclude from these ratio that Joe Biden has the highest popularity among them.

Ranking tweets

The similarity ratio shows us that Joe Biden has the highest popularity among the 6 candidates. However, you cannot guess based on that that Biden will win the election in 2024 because people in twitter may tweet in a negative or positive way. Luckily, there is a way to find how positive or negative something is. To do this analysis, I used sentiment function from textblob library. This function provides the polarity and subjectivity of a tweet. The polarity measures how positive or negative a tweet is (ranges from -1 to 1), and the subjectivity measures how subjective the tweet is (ranges from 0 to 1, 0 means objective and 1 means subjective). I added a column into the data frames, polarity and subjectivity. After I added these columns for each candidate, I measured the mean, minimum, maximum, and the median. Here is what I got,

'For Newsom'

	polarity	subjectivity
mean	-0.009394	0.335152
amax	0.800000	1.000000
amin	-0.800000	0.000000
median	0.000000	0.400000

'For Whitmer'

	polarity	subjectivity
mean	0.109364	0.374802
amax	0.500000	1.000000
amin	-0.400000	0.000000
median	0.000000	0.150000

'For Haley'

	polarity	subjectivity
mean	0.067807	0.136216
amax	0.750000	0.950000
amin	-0.516667	0.000000
median	0.000000	0.000000

'For Pence'

	polarity	subjectivity
mean	0.093003	0.245913
amax	0.525000	1.000000
amin	-0.155556	0.000000
median	0.000000	0.000000

'For Biden'

	polarity	subjectivity
mean	0.027511	0.251798
amax	0.550000	1.000000
amin	-1.000000	0.000000
median	0.000000	0.100000

'For DeSantis'

	polarity	subjectivity
mean	0.015428	0.302152
amax	0.800000	1.000000
amin	-1.000000	0.000000
median	0.000000	0.254167

We can see from these figures that Whitmer has the highest mean polarity, but her polarity is very close to Pence, and Newsom has the lowest polarity. Also, Whitmer has the highest subjectivity, and Haley has the lowest subjectivity.

Challenges

In this project, I struggled a lot with using naïve bayes classification, which is a machine learning classifier that classify texts based on other data. My idea was to collect articles and classify them as negative or positive. Before I classified the tweets using naïve bayes classifier, I collected some articles about each of the candidates about why they should be president in 2024 and why they should not be president. I categorized these articles as positive and negative. Based on these articles, I tried to build classifiers on the positive and negative articles on training data, and test them on the testing data, but I got an error message when I tried to fit them MultinomialNB().fit(). I think the training data cannot be classified as it is not big enough. I think if I had a bigger classified dataset naïve bayes would work. This might be unnecessary, but I think 12 to 13 hours of my time worth putting in my report.

Conclusion

After generating word clouds, matching words between tweets and articles, and measuring polarity and subjectivity, now it is time to make a conclusion. Similarity ratio shows that Biden has the highest popularity, and I can say that Biden, Pence, DeSantis, Newsom, and Whitmer all have close ratios to each other. Also, Whitmer has the highest polarity and subjectivity, which means that people positive about her. Based on these data, it is predicted that Whitmer is going to win among them.