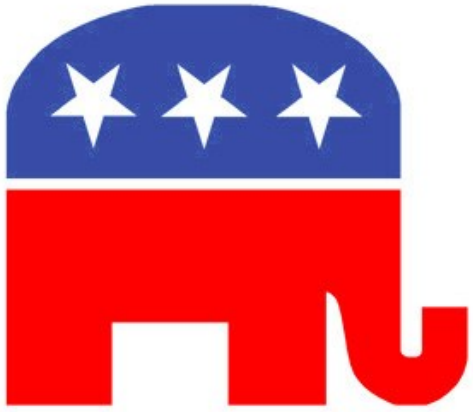
Elephant vs Donkey



— 2020 President Election Prediction

Data Samurai
Annan CHEN, Tianqi SHEN, Bingsu MO, Xiaoya XU

Dec 4, 2019

Overview

COLUMBIA | ENGINEERING
The Fu Foundation School of Engineering and Applied Science

Gradually Losing Smiles?

"President Donald Trump had **abused the power of the presidency** by withholding both military aid and a White House meeting as a means of pressuring **Ukraine**'s newly elected president, Volodymyr Zelensky "

----Impeachment inquiry against Donald Trump

"The report, which states that the president "sought to undermine the integrity of the U.S. presidential election process, and **endangered U.S. national security**," was approved on a party-line vote. "

----Washington Post on Dec 3rd, 9:12 p.m.

Overview

COLUMBIA | ENGINEERING
The Fu Foundation School of Engineering and Applied Science

Census

- **Data Cleaning**
- **Data Visualization**

Algorithms

- Gradient Boosted Tree
- XG Boost
- SVR
- Linear Regression with Regularization

Modelling & Validation

Prediction

Result

Twitter

Different Candidates

Text Mining

- Sentiment Analysis
- Candidate Popularity Comparison
- WordCloud

Verification

Prelim — President Election

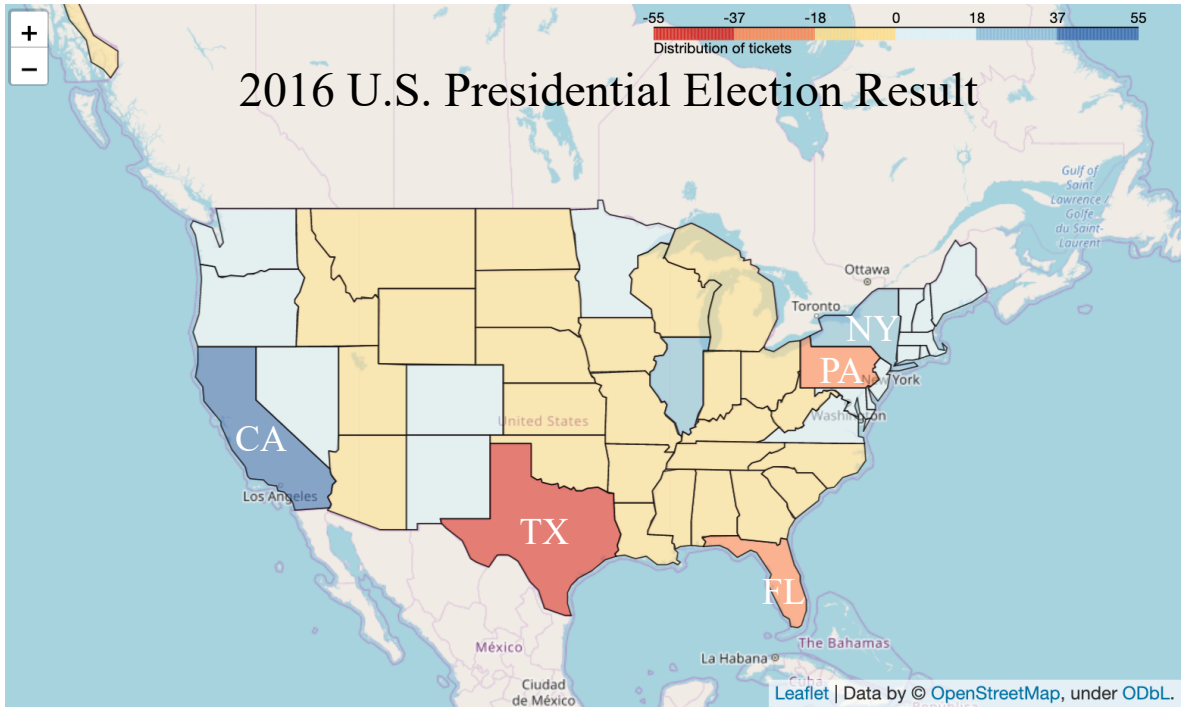
COLUMBIA | ENGINEERING
The Fu Foundation School of Engineering and Applied Science

- Basic Rules for each state:

"Winner Takes All!"

- California (CA): 55
- Texas (TX): 38
- Florida (FL): 29
- New York (NY): 29
- Pennsylvania (PA): 20

538 Votes In Total



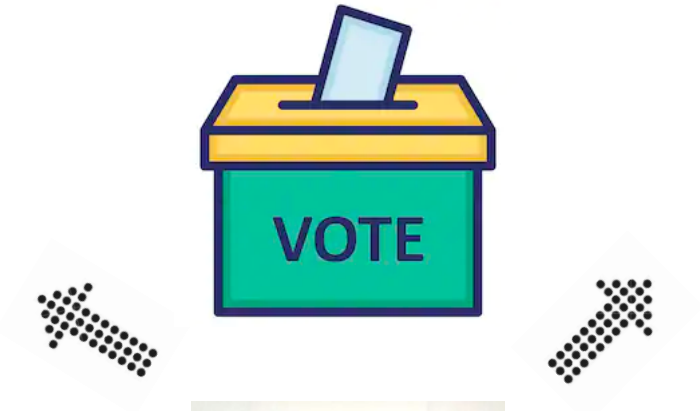
"I got 304"

"I had 332"

Dataset — part 1

COLUMBIA | ENGINEERING
The Fu Foundation School of Engineering and Applied Science

- Data Composition:



An Indecisive Voter

Dataset — part 1

COLUMBIA | ENGINEERING
The Fu Foundation School of Engineering and Applied Science

- Data Composition:

EMPLOYMENT

- Unemployed Rate
- Labor Force
- Working Agriculture

- Health Insurance Coverage

HEALTH CARE

RACE

- White People

- Turnout
- Previous vote

ELECTION

EDUCATION

- Bachelor Holders

ECONOMY

- Gini Coefficient

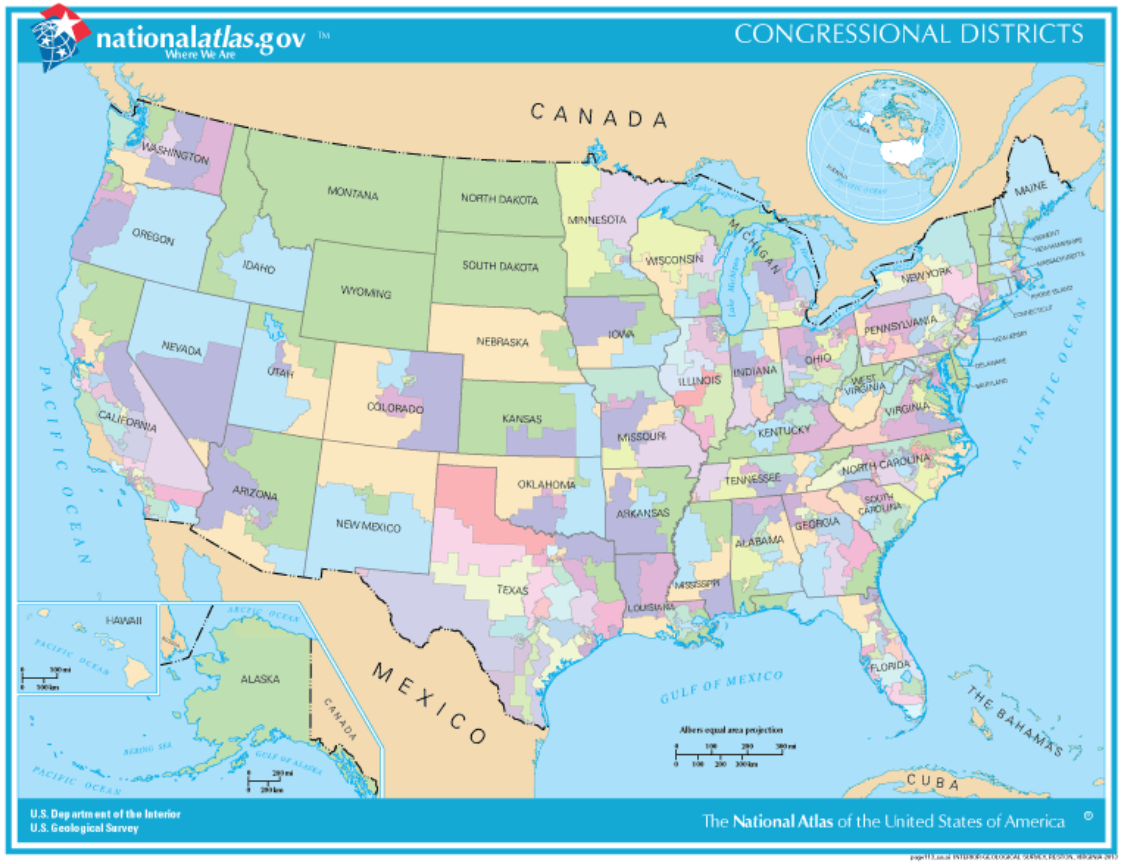
Dataset — part 1

COLUMBIA | ENGINEERING
The Fu Foundation School of Engineering and Applied Science

- Data Source:
  - United States Census Bureau (Census API)
  - American Community Survey (ACS) – 1 year

- Data Items:
  - Congressional District (435)

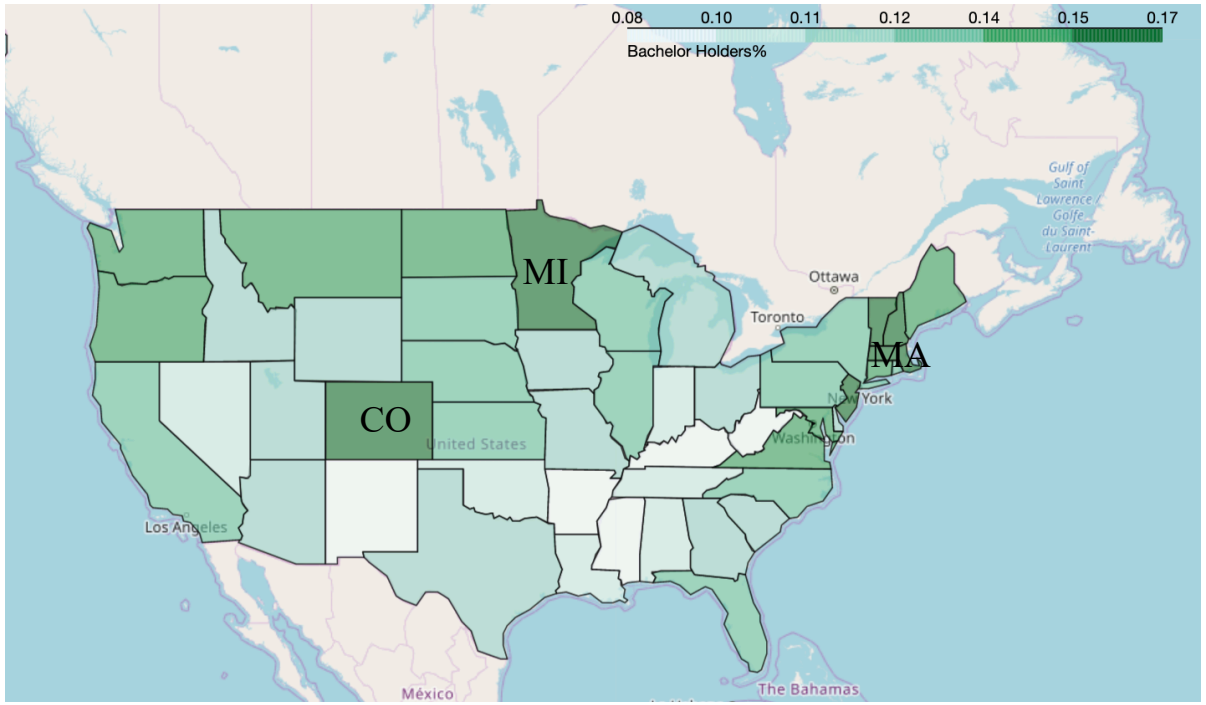
- Data Features (44):
  - Census Data, Turnout, Approval Rate, etc.



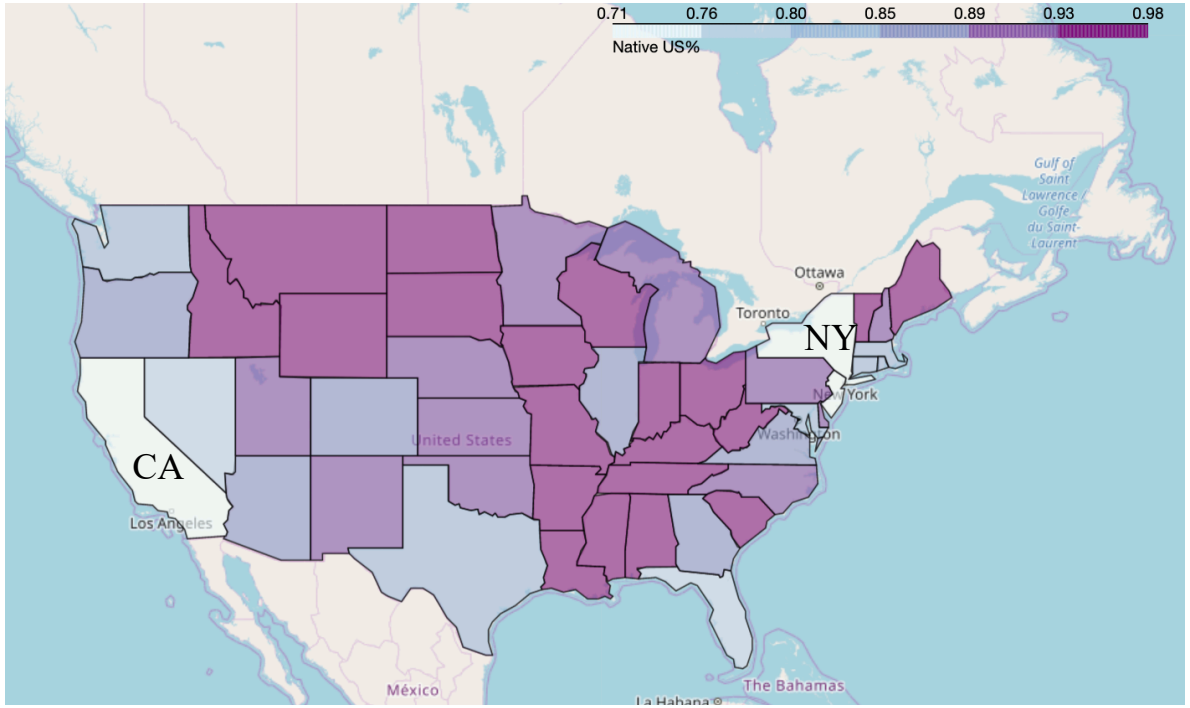
Dataset — part 1

COLUMBIA | ENGINEERING
The Fu Foundation School of Engineering and Applied Science

People Holding Bachelor Degree(%), 2016



Native US people(%) in the U.S., 2016



Feature Transformation

Categorical Variable: State name – Longitude/Latitude

Numerical Variable (Normalization): Population composition

Numerical Variable (Standardization): Mean – 0; Variance - 1

Dependent Variable: The difference between votes for two parties (percent)

```
Index(['P_Democrat', 'P_Republican', 'Pop', 'Females', 'Median_Age',
       'Veterans', 'White_People', 'Afr_Am_People', 'Asian',
       'American_Ind_Alk_Ntv', 'Cuban_Origin', 'Puerto_Rican_Origin',
       'Dominican_Origin', 'Mexican_Origin', 'Native_US', 'Foreign_Born',
       'Children', 'Married_Households', 'Less_Highschool', 'Bachelor_Holders',
       'Labor_Force_Eligible', 'Unemployed', 'Median_Household_Income',
       'Median_Income', 'Below_Poverty_Level_LTM', 'Wealthy_Households',
       'Gini_Index', 'Median_Age_Of_Worker', 'Workers', 'Working_Agricult',
       'Working_Construction', 'Working_Manufacturing', 'Working_Retail',
       'Working_Transportation', 'Working_Finance', 'Working_Education',
       'Working_Health', 'Working_Food', 'Working_Public_Admin',
       'Working_Information', 'Median_Gross_Rent', 'District_Name', 'State_Id',
       'CD', 'State_Full', 'State', 'Health_insurance_coverage', 'Turn_Out',
       'Approval_Rate_Predecessor', 'Vote_2012', 'Vote_2016', 'Vote_2020',
       'Latitude', 'Longitude'],
      dtype='object')
```

Algorithms

COLUMBIA | ENGINEERING
The Fu Foundation School of Engineering and Applied Science

Metrics: The square root of mean squared error based on cross validation on training data

Gradient Boosted Tree	It builds the model in a stage-wise fashion like other boosting methods do, and it generalizes them by allowing optimization of an arbitrary differentiable loss function.
XG Boost	XG Boost is an optimized distributed gradient boosting library designed to be highly efficient, flexible and portable.
SVR	Support-vector machines (SVMs) are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis.
Lasso/Ridge Regression	Lasso/Ridge are regression analysis methods that perform both variable selection and regularization in order to enhance the prediction accuracy. The lowest root MSE: 5.8%

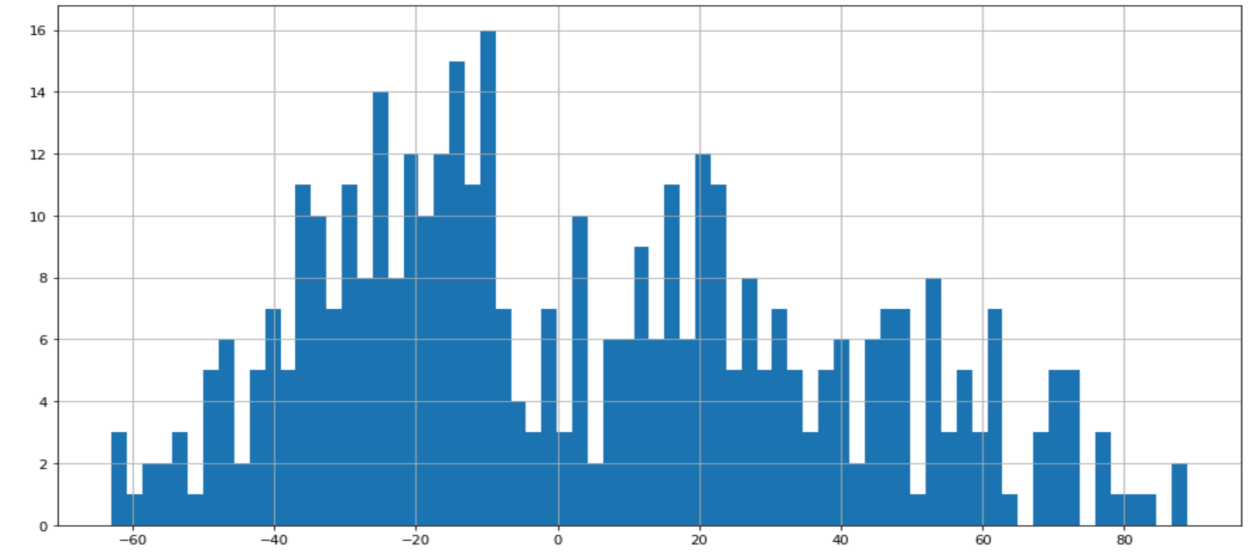
Validation

COLUMBIA | ENGINEERING
The Fu Foundation School of Engineering and Applied Science

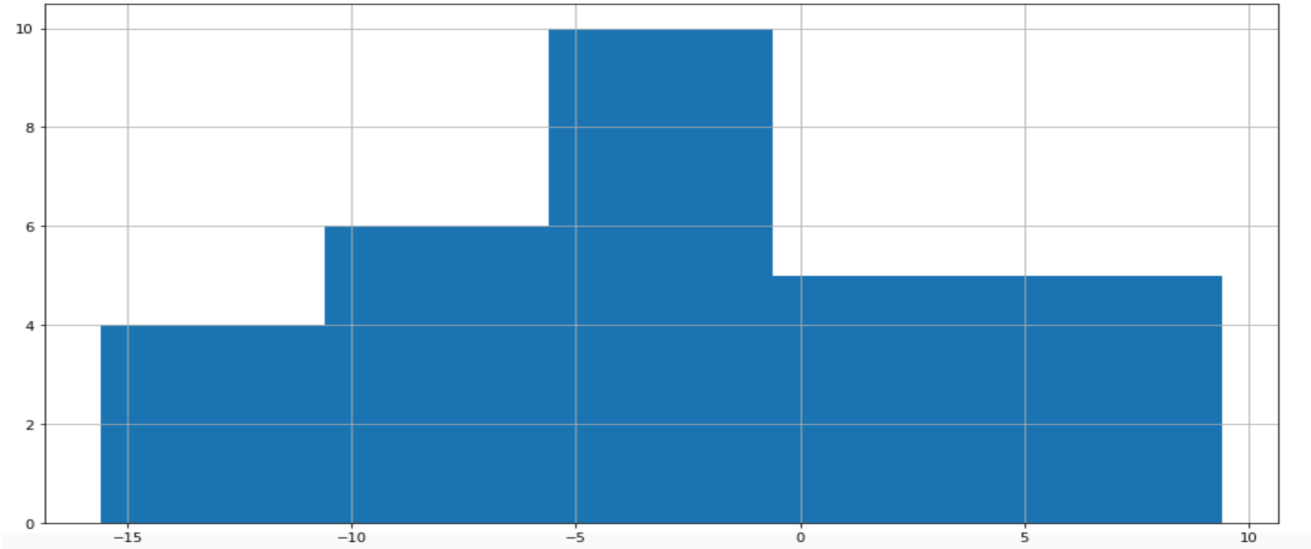
Validation on 2016 testing data: Root MSE is 8.9%

93% of validation data has correct winner prediction at CD level

```
result[result.sign_correct].actual.hist(bins=70, figsize=(15,8))
<matplotlib.axes._subplots.AxesSubplot at 0x1a1f4f40b8>
```




```
result[~result.sign_correct].actual.hist(bins=5, figsize=(15,8))
<matplotlib.axes._subplots.AxesSubplot at 0x1a1f4dc320>
```

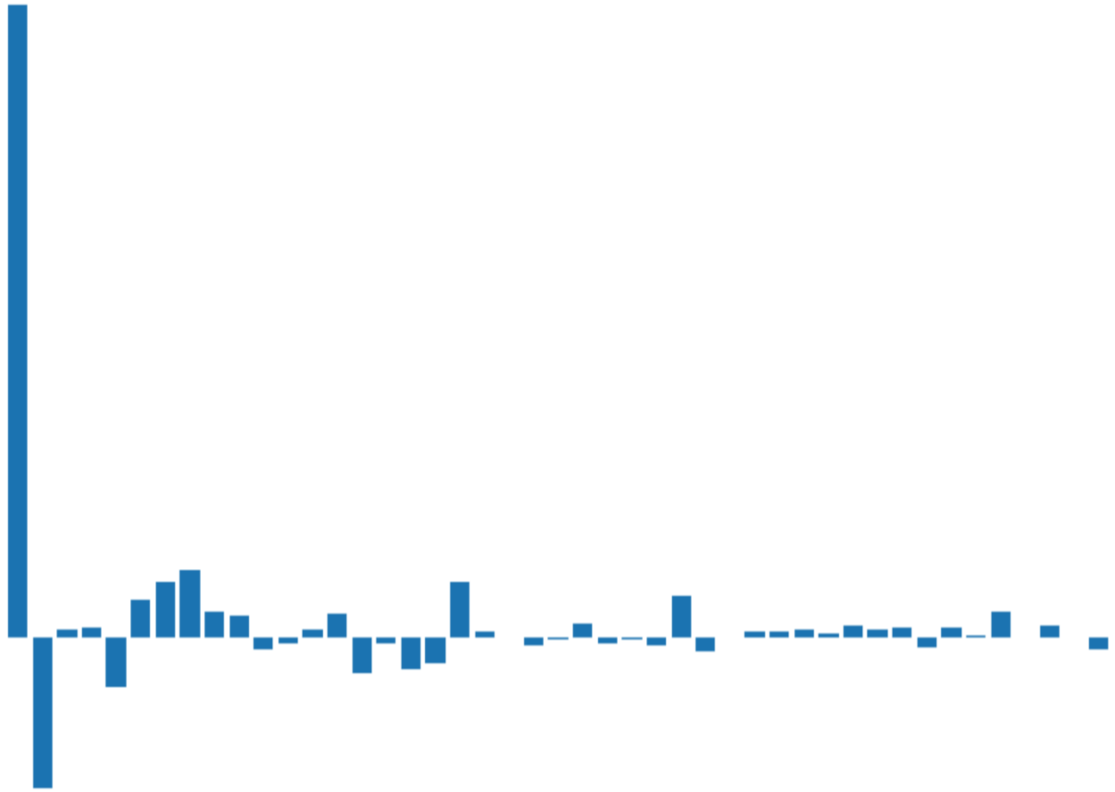


Feature Importance

COLUMBIA | ENGINEERING
The Fu Foundation School of Engineering and Applied Science

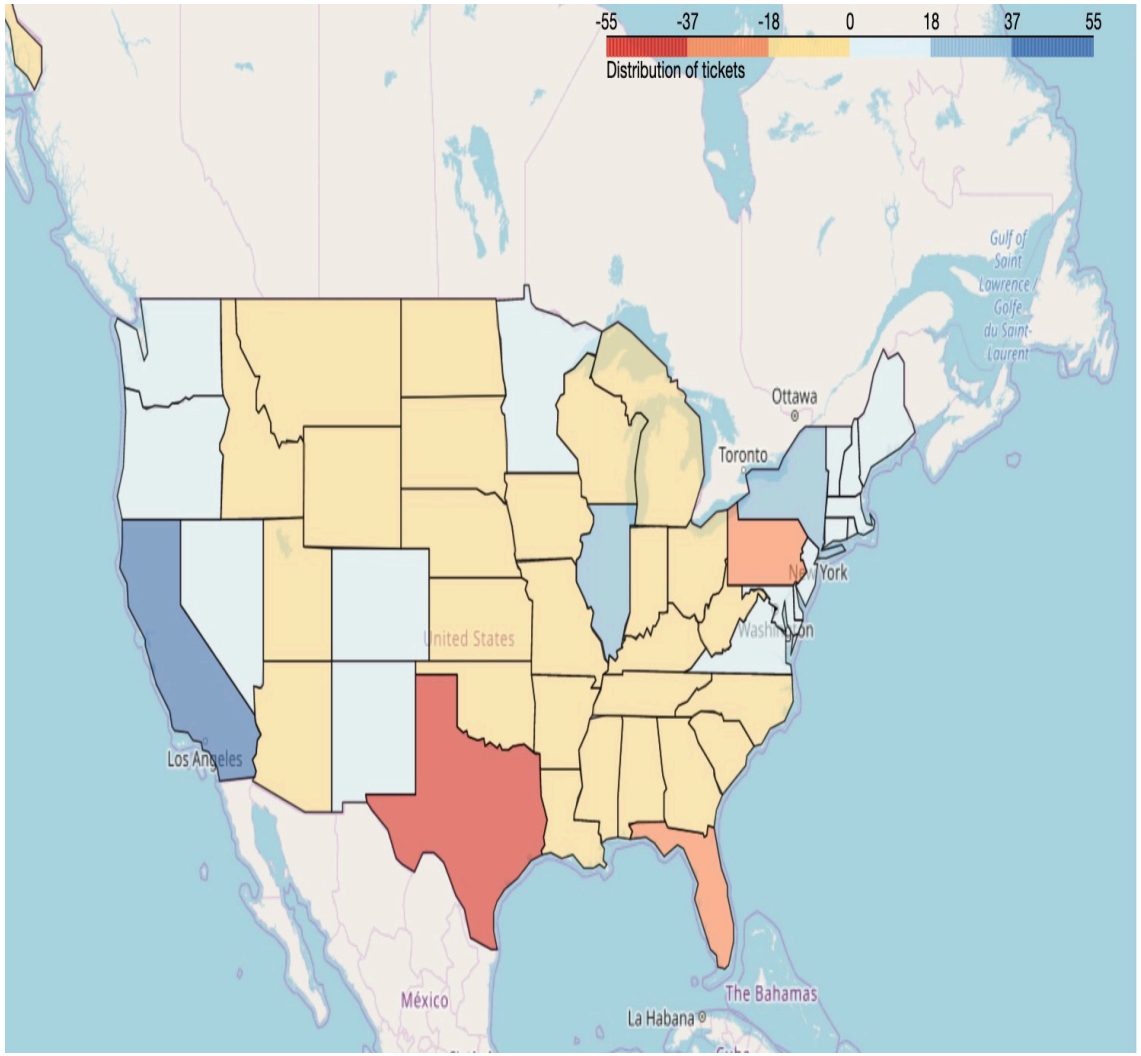
National wide: **Coefficient ranking of lasso regression**

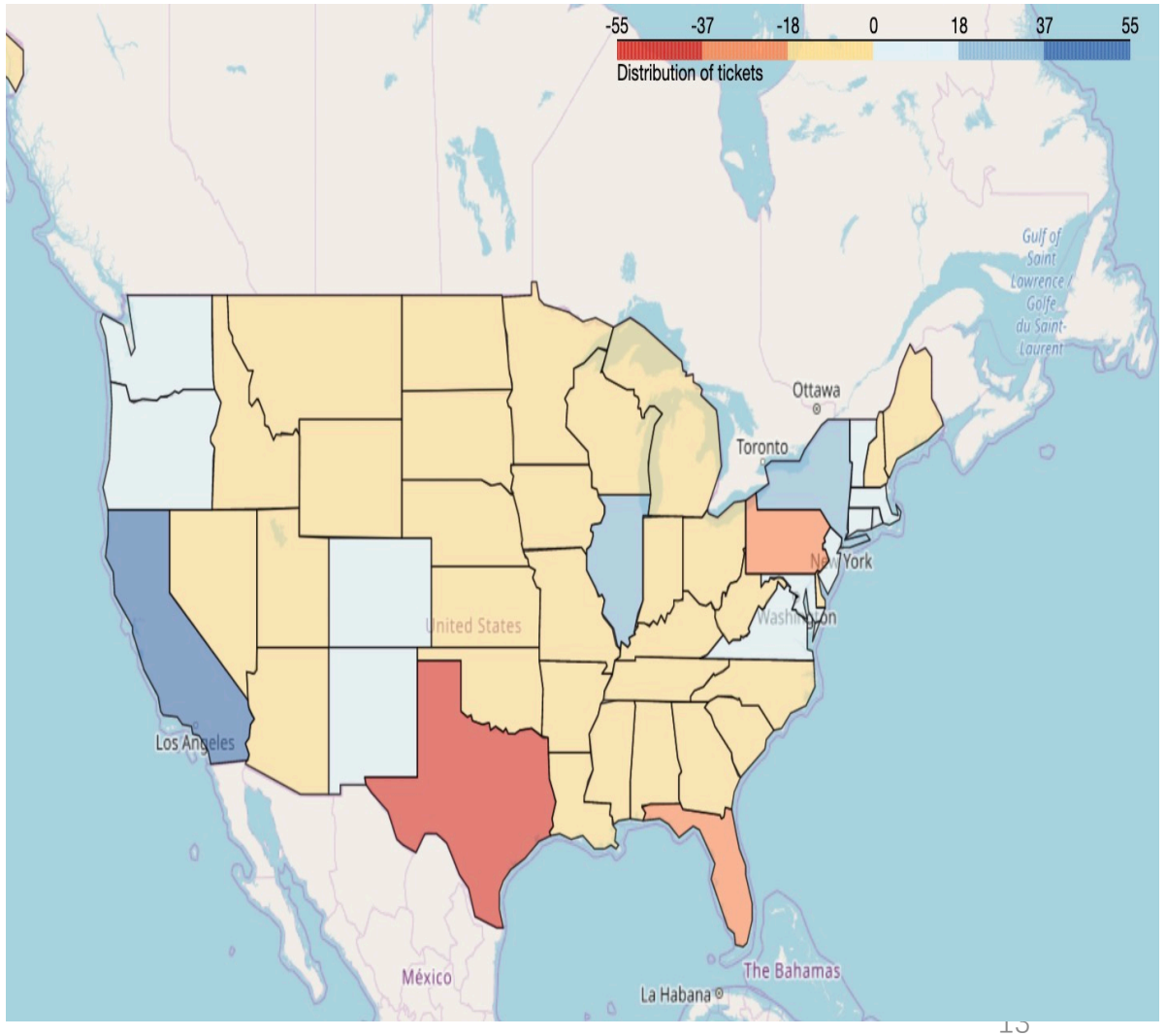
	column_name	coefficient	coef_magnitude
0	P_Democrat	24.303052	24.303052
1	P_Republican	-5.825934	5.825934
7	Afr_Am_People	2.532814	2.532814
6	White_People	2.113056	2.113056
18	Less_Highschool	2.094419	2.094419



Vote Comparison(2016, 2020)

COLUMBIA | ENGINEERING
The Fu Foundation School of Engineering and Applied Science





Why a different party?

COLUMBIA | ENGINEERING
The Fu Foundation School of Engineering and Applied Science

SHAP, explain output for every data in the sample

2016 for California:

2020 for California:

However, is it too naive to make prediction without considering candidates' personal characteristics?

Dataset — part 2

COLUMBIA | ENGINEERING
The Fu Foundation School of Engineering and Applied Science

- Candidates:

Democratic

Elizabeth Warren

Bernie Sanders

Cory Booker

Julián Castro

Kamala Harris

Kirsten Gillibrand

Beto O'Rourke

Republic

Donald J. Trump

Dataset — part 2

COLUMBIA | ENGINEERING
The Fu Foundation School of Engineering and Applied Science

- Data Source:
  - Twitter (Through Twitter API and tweepy package)
  - 1.We get each candidates' tweets and related likes and retweets through web scraping. For each candidates, we get their most recent 3000 tweets due to Twitter API limits.
  - 2. We also derive the most recent tweets that mentioned our target candidate

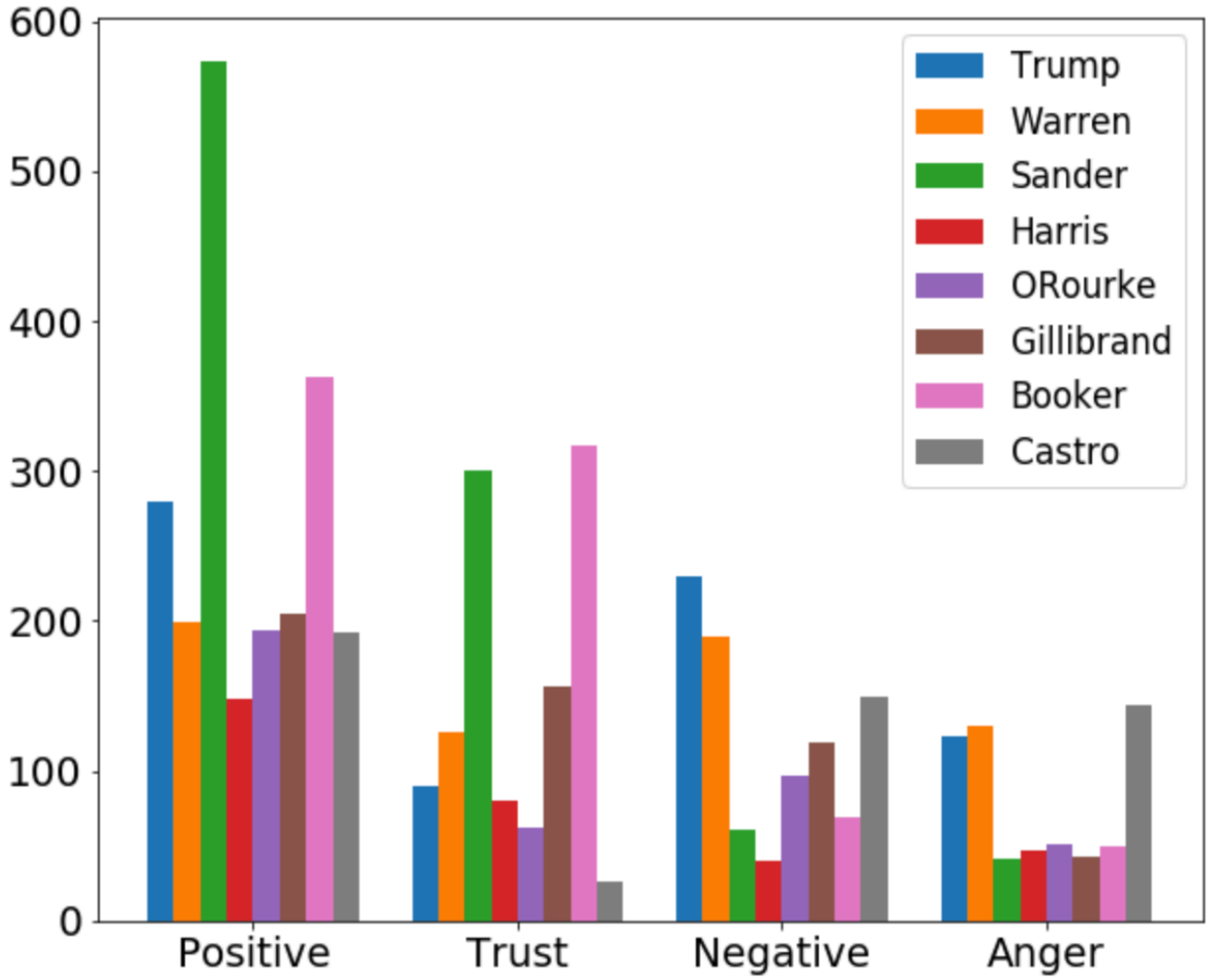
- Fun Fact:
  - For president Trump, 3000 tweets only takes us back to Aug 2019, which means president Trump tweets approximately 20 times a day on average. (including retweets)

Sentiment Analysis

COLUMBIA | ENGINEERING
The Fu Foundation School of Engineering and Applied Science

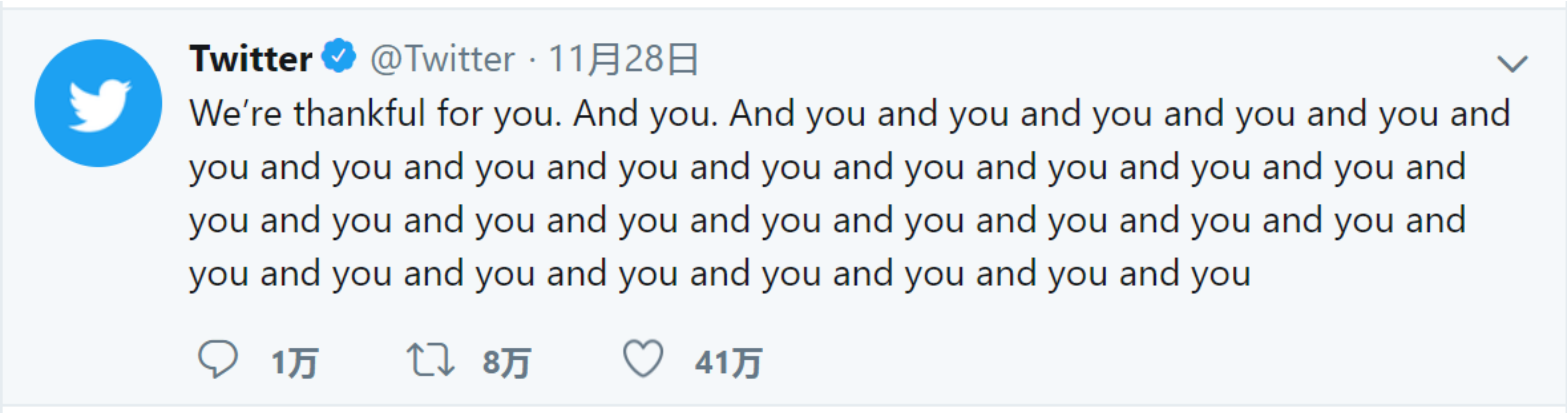
- We search different candidates' name on Twitter
- NRC data
- We choose 2 sentiments and 2 emotions as comparison basis
  - Positive
  - Negative
  - Trust
  - Anger

- Result: **Bernie Sander** and **Cory Booker** are highly competitive



Comparison Regarding Likes/Retweets

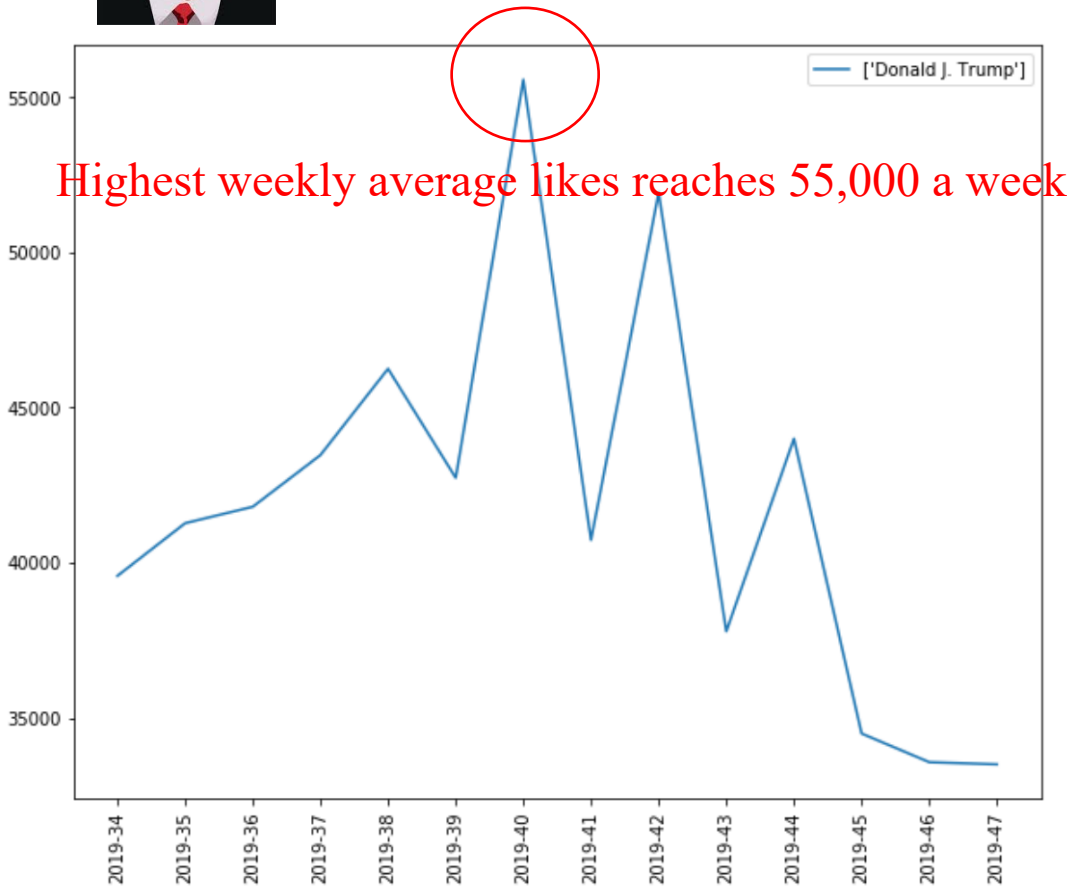
- We scrape each candidate's most recently 3000 tweets due to twitter API limits
- We analyze their weekly average likes / reweets



Monthly Average Likes

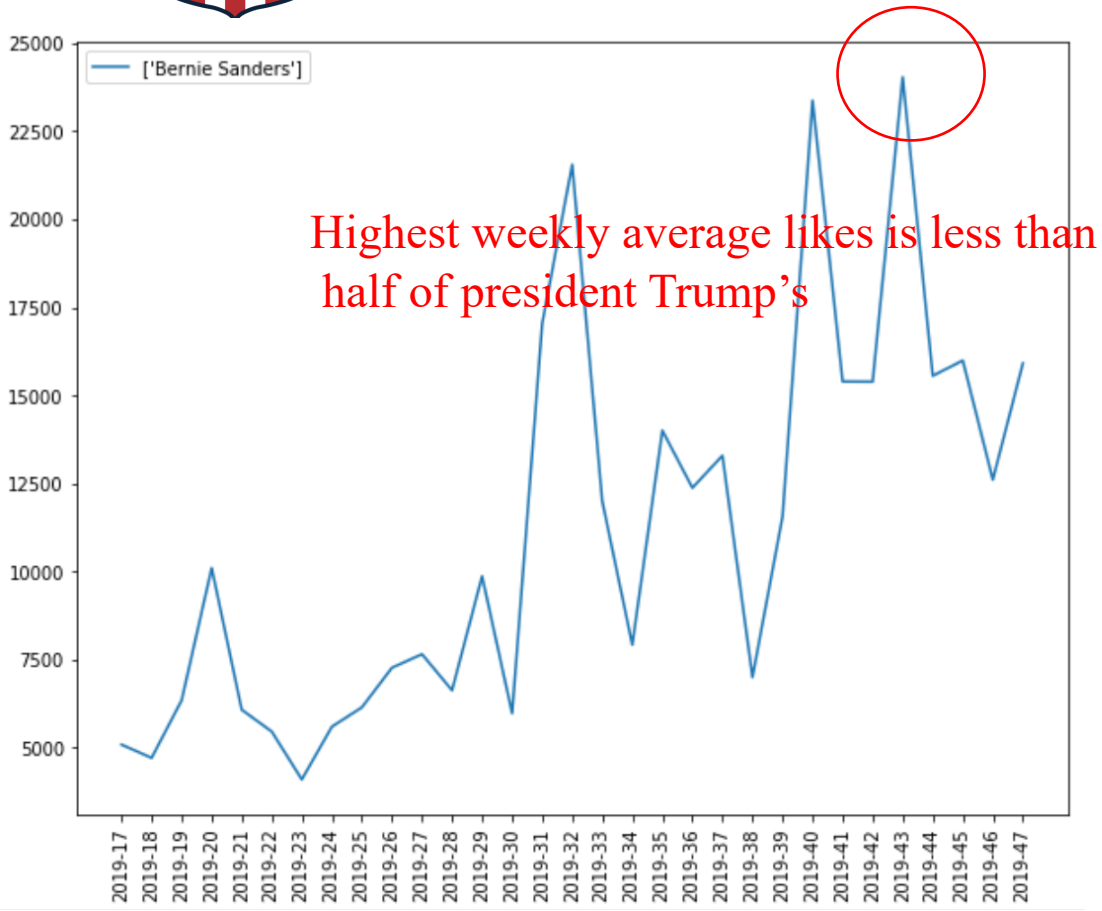
COLUMBIA | ENGINEERING
The Fu Foundation School of Engineering and Applied Science

Donald J. Trump



Highest weekly average likes reaches 55,000 a week

Bernie Sanders

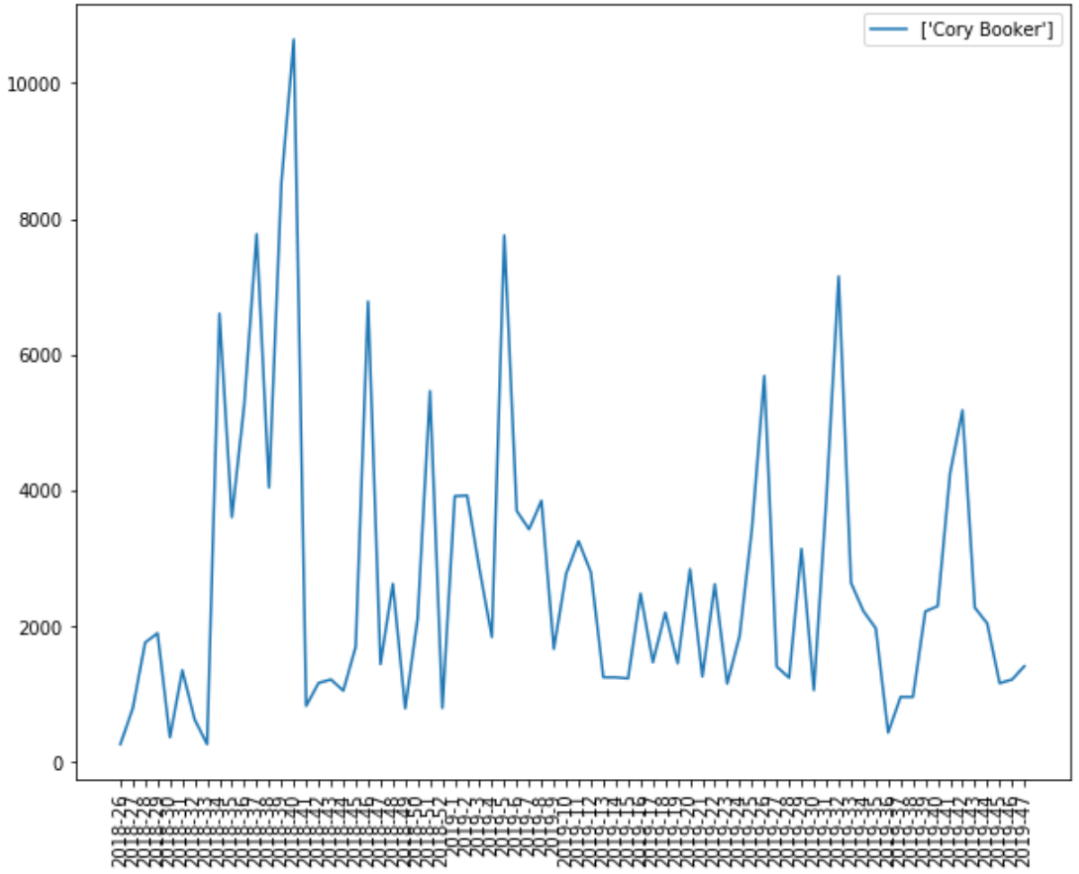


Highest weekly average likes is less than half of president Trump's

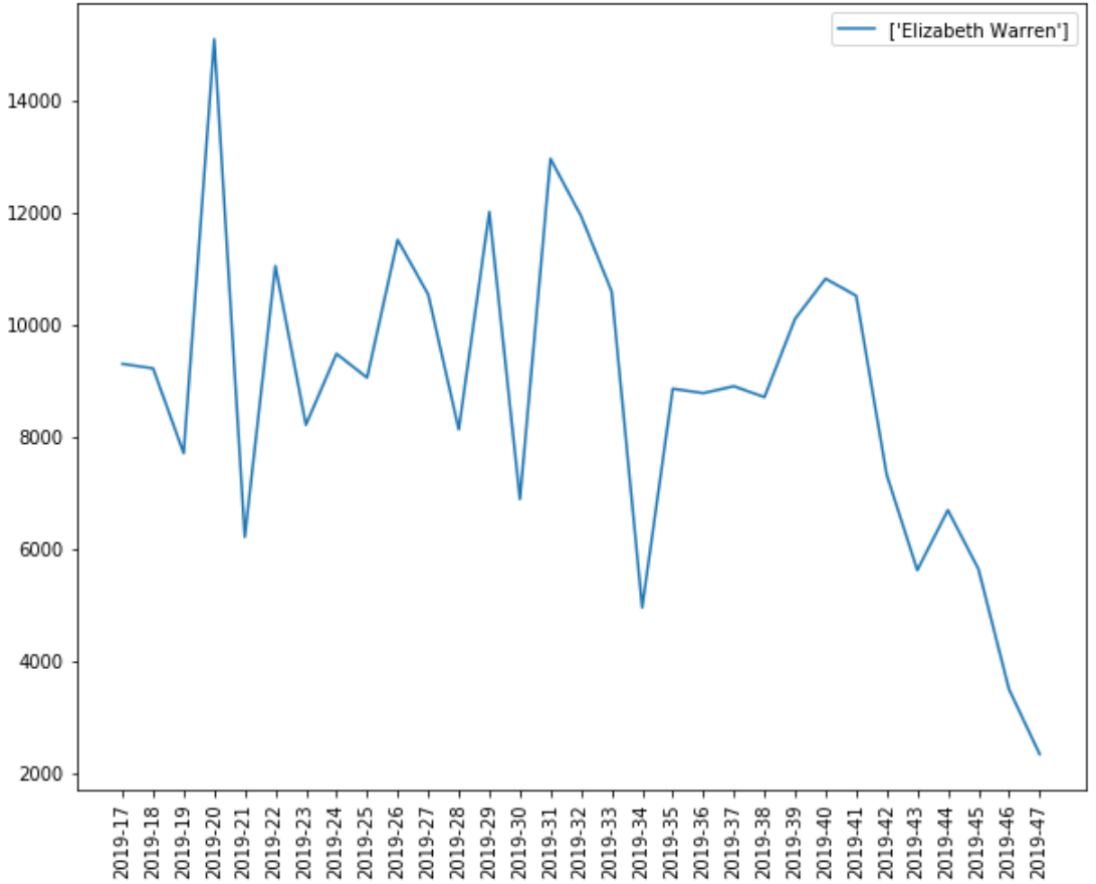
Monthly Average Likes

COLUMBIA | ENGINEERING
The Fu Foundation School of Engineering and Applied Science

Cory Booker



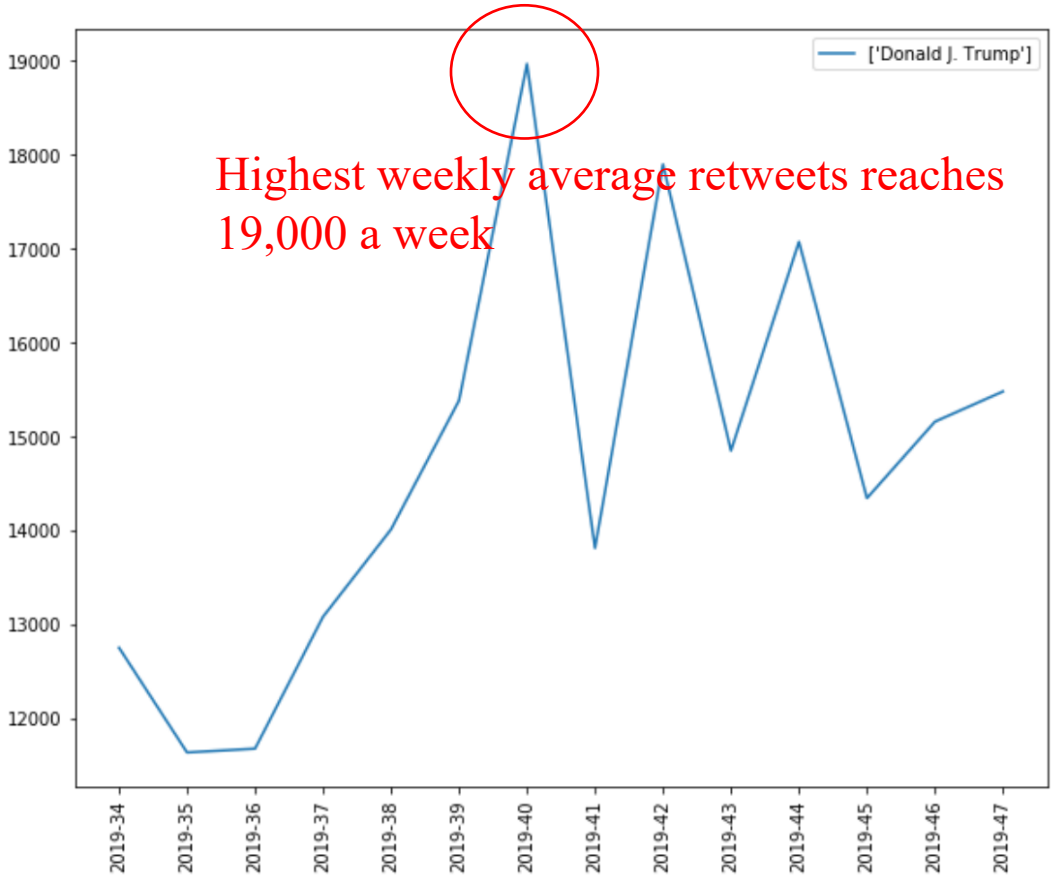
Elizabeth Warren



Monthly Average Retweets

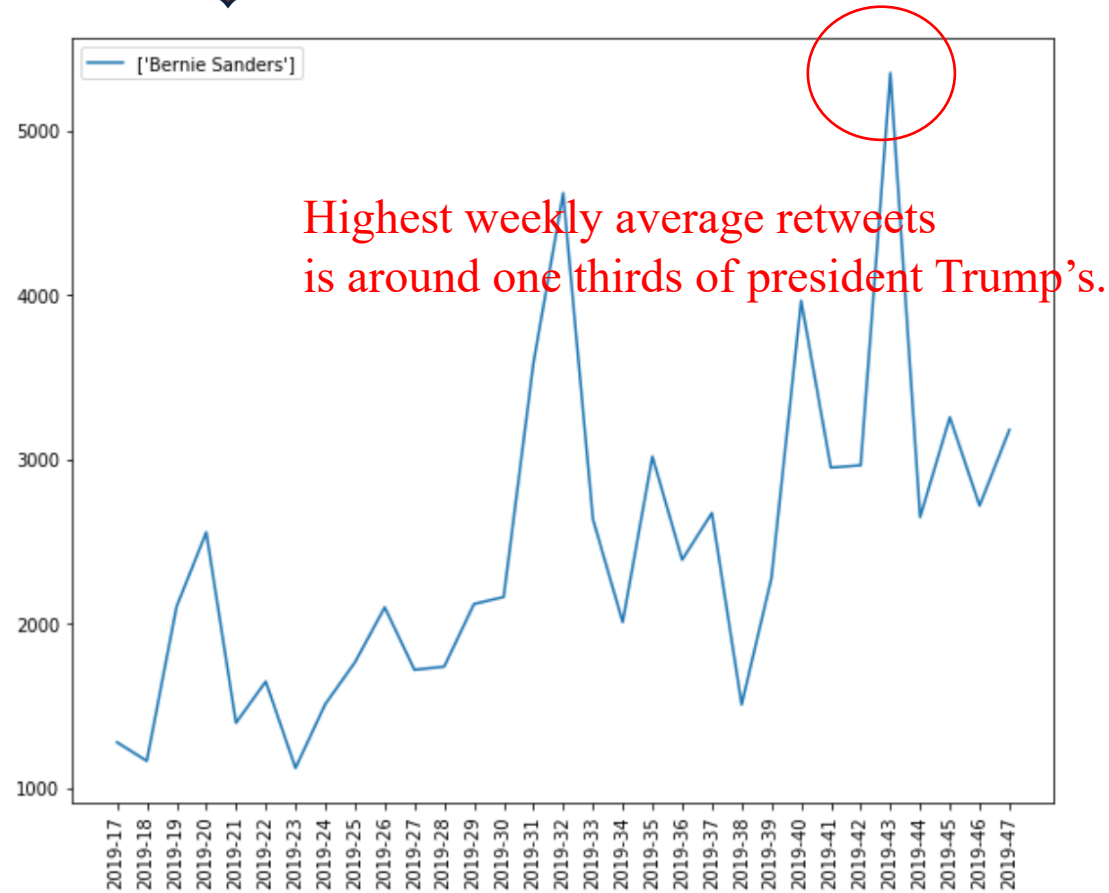
COLUMBIA | ENGINEERING
The Fu Foundation School of Engineering and Applied Science

Donald J. Trump



Highest weekly average retweets reaches 19,000 a week

Bernie Sanders



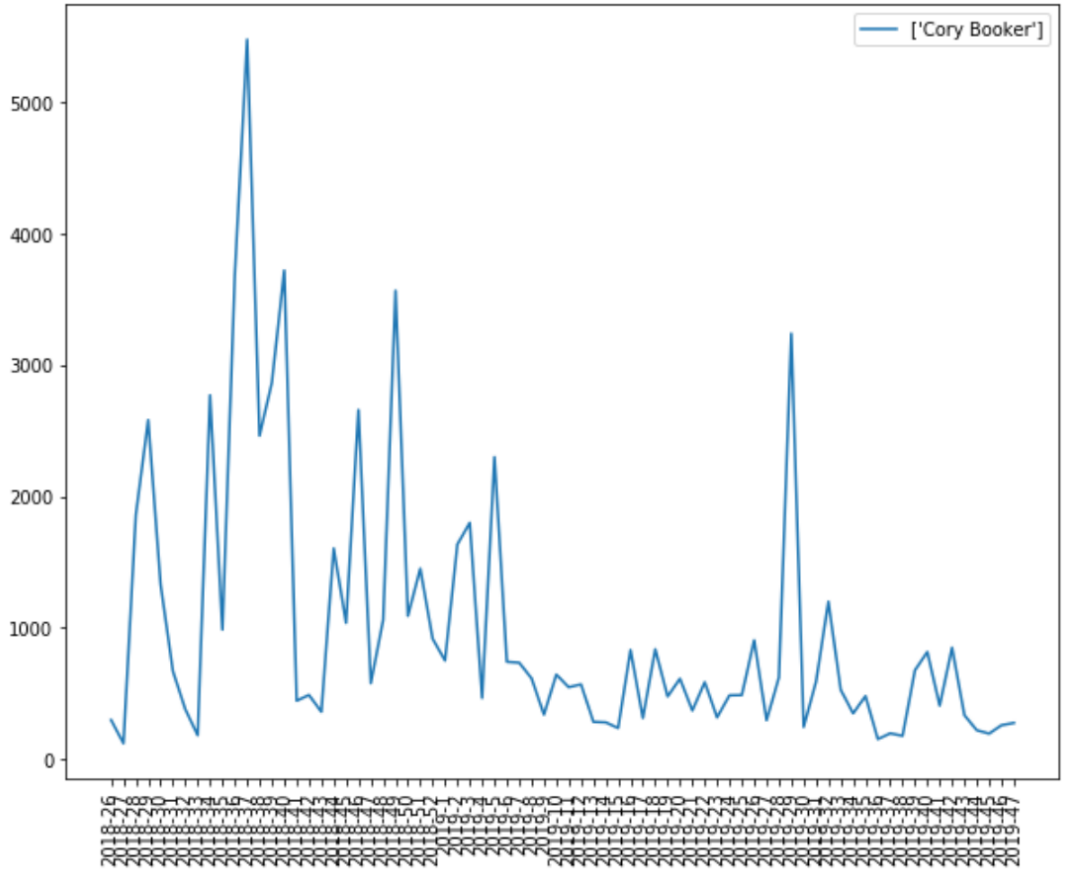
Highest weekly average retweets is around one thirds of president Trump's.

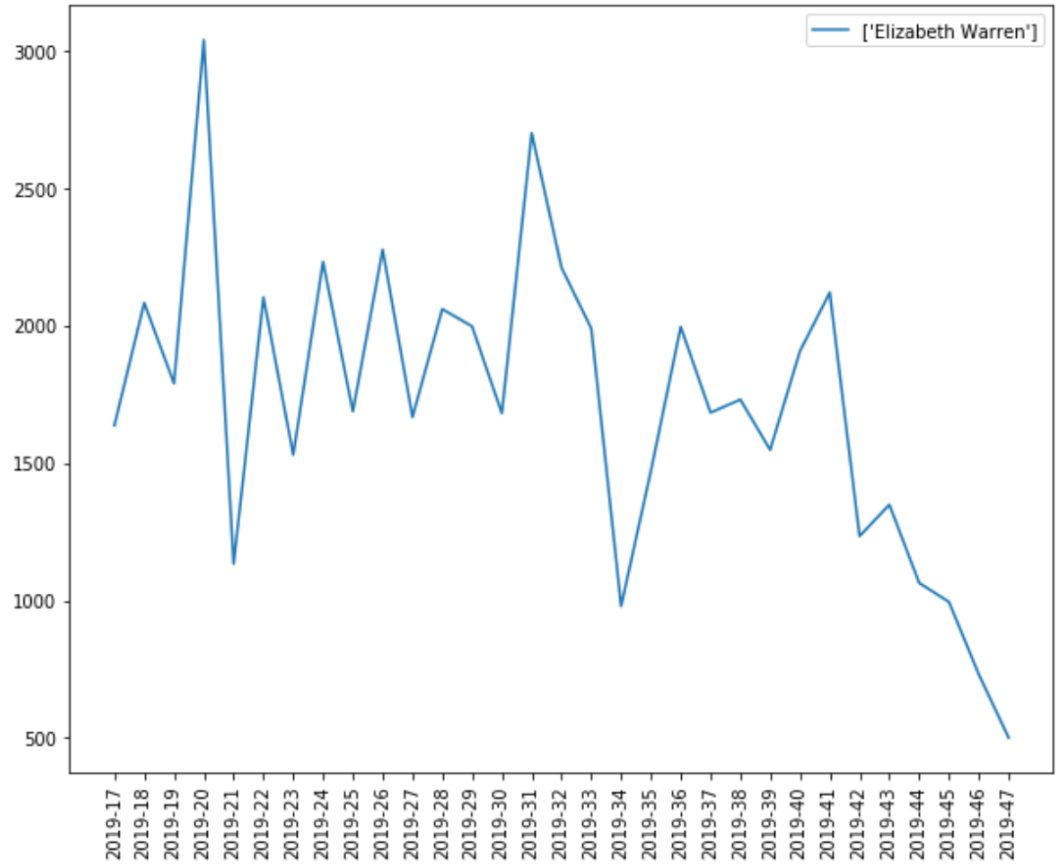
Monthly Average Reweets

COLUMBIA | ENGINEERING
The Fu Foundation School of Engineering and Applied Science

Cory Booker

Elizabeth Warren

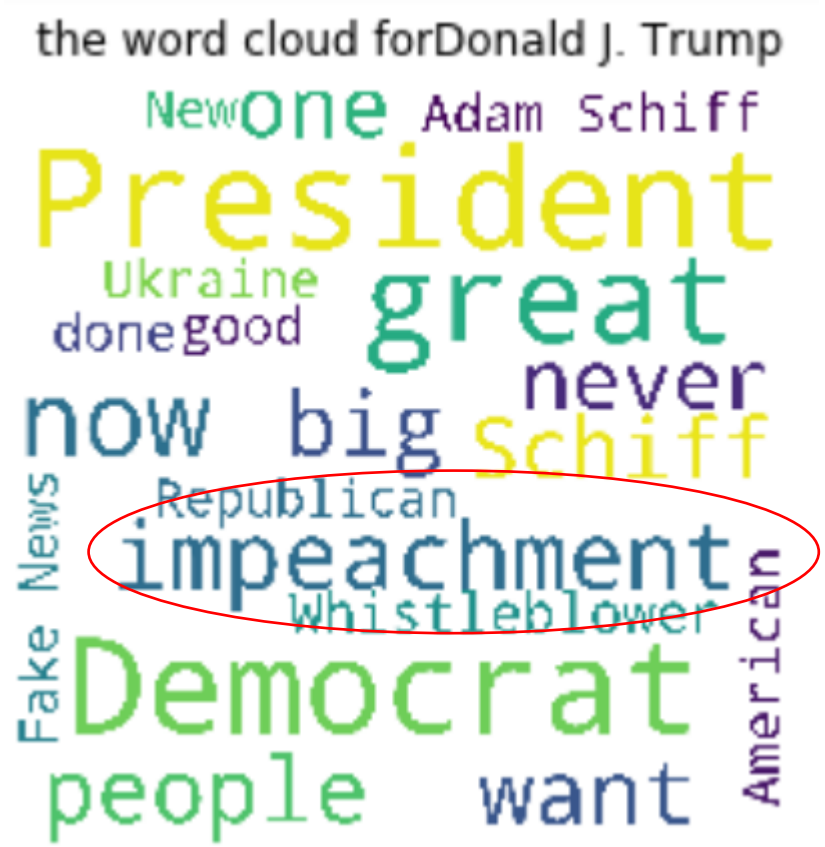




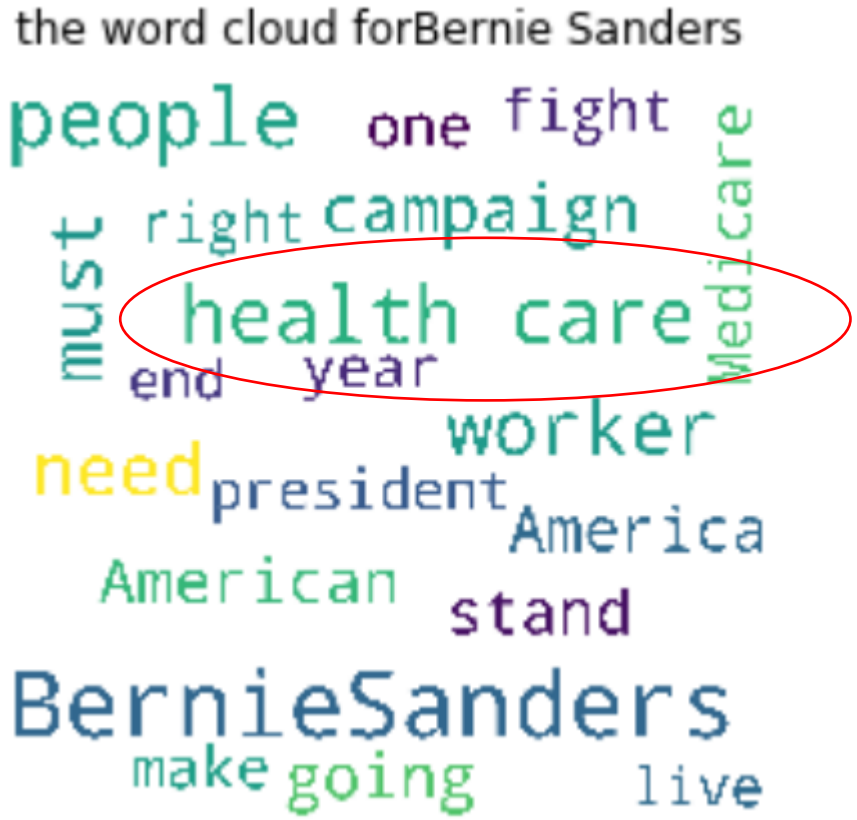
Comparison Regarding WordCloud

COLUMBIA | ENGINEERING
The Fu Foundation School of Engineering and Applied Science

Donald J. Trump



Bernie Sanders

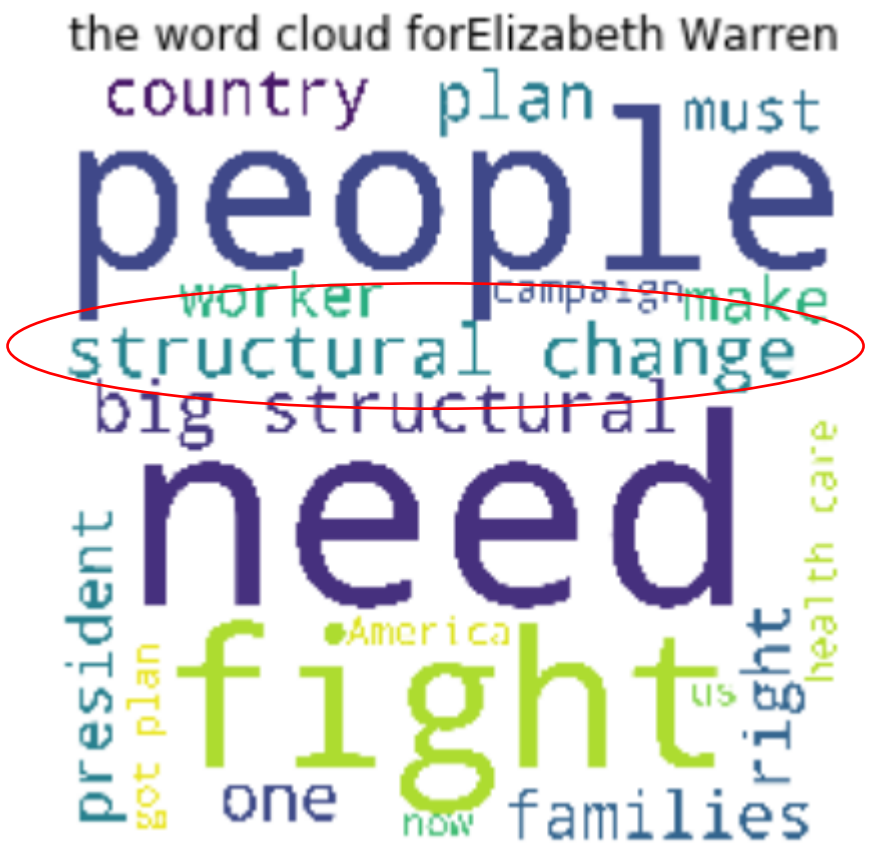


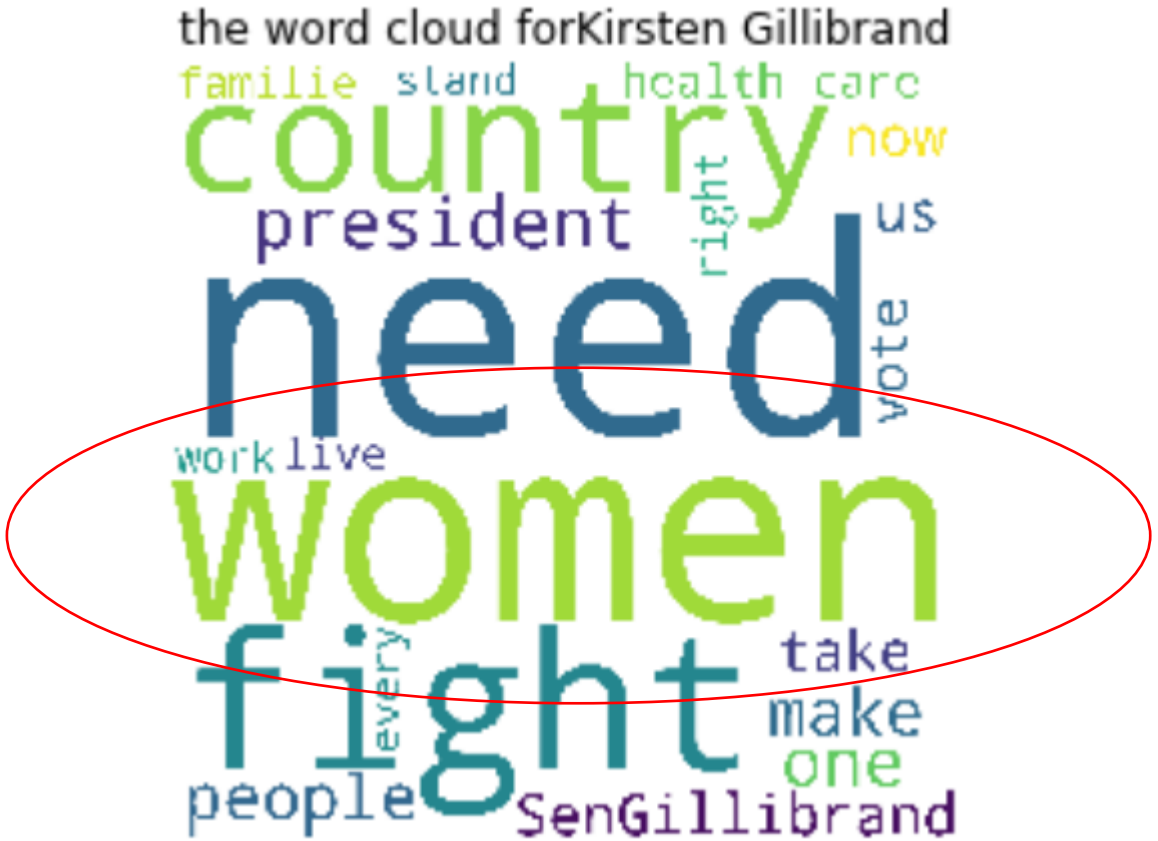
Comparison Regarding WordCloud

COLUMBIA | ENGINEERING
The Fu Foundation School of Engineering and Applied Science

Elizabeth Warren

kirsten Gillibrand





Conclusion

COLUMBIA | ENGINEERING
The Fu Foundation School of Engineering and Applied Science

- Although the attitudes towards Mr. Trump in general are not very positive, he is still considered as the most popular candidate based on data from twitter, which provides him with a large potential in gaining citizens' attentions as well their votes.

- This further verifies the prediction result of our models.


Q
A