Predictive power of social media in US elections

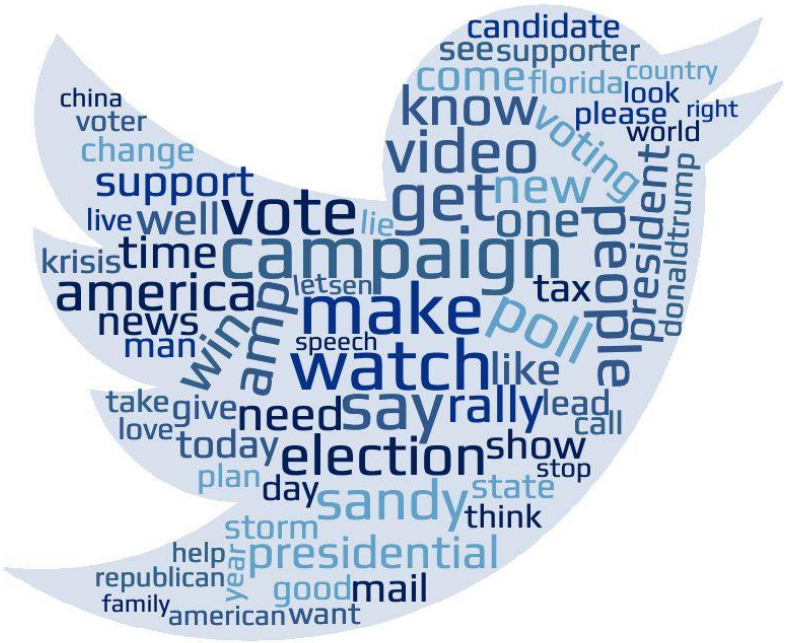
Elizabeth Sames

Social Media era

- >60% world population

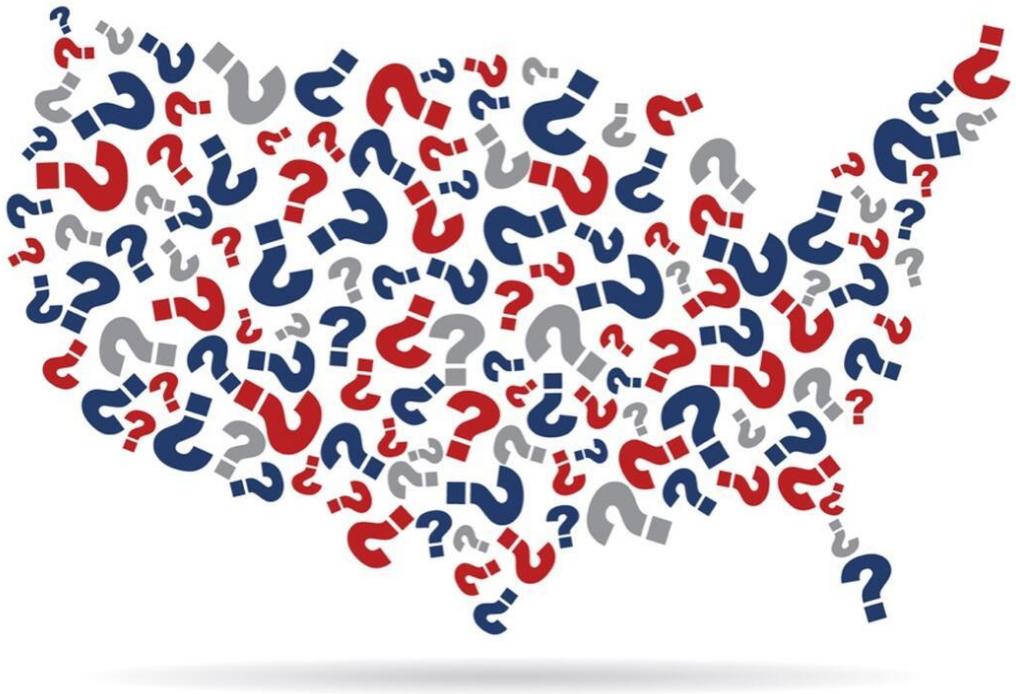
- 2.4 hours on social media daily

Twitter



- 206 million daily active users worldwide
- 200 million tweets per day

- 73 million users in the US
- ~20% of all tweets are political



Can we harness the power of Twitter data to predict the outcome of an election?

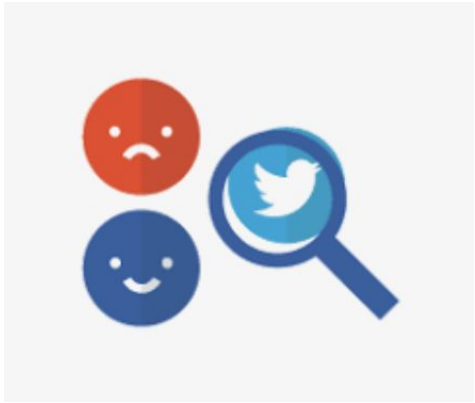
THE DATA

- US Presidential elections 2008-2020 (4 elections)

- US Gubernatorial elections 2009-2020 (156 elections)

Webscraping election data and tweets.

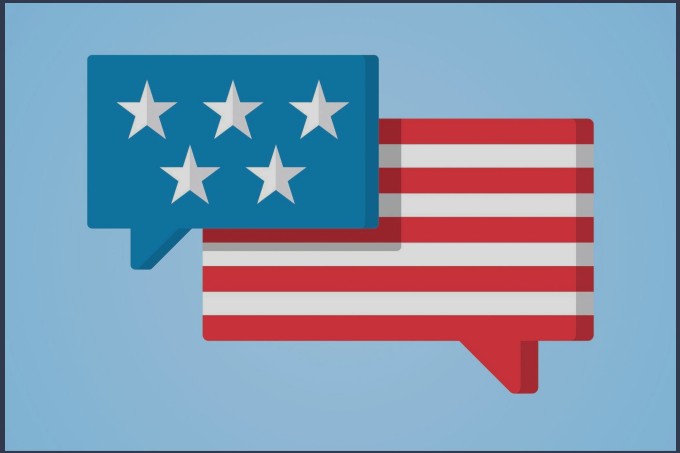
Tweet sentiment classification using NLTK.

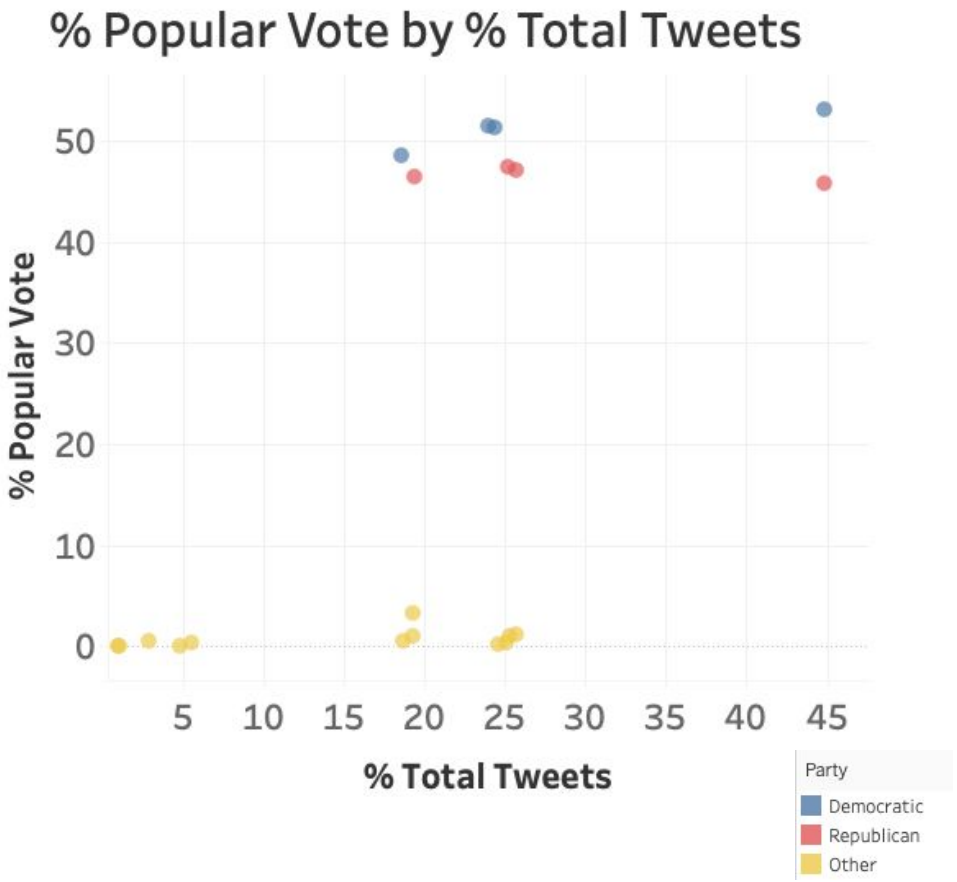


Predict election outcomes with a regression model.

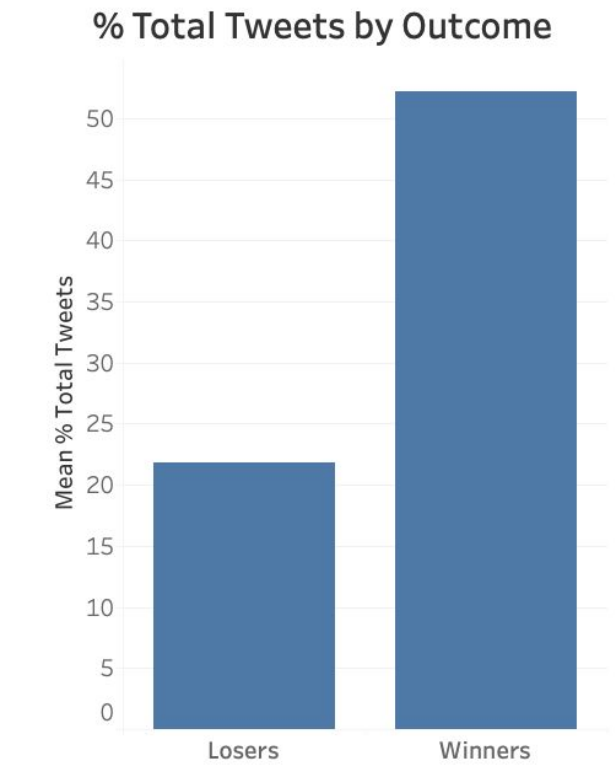
THE PROCESS

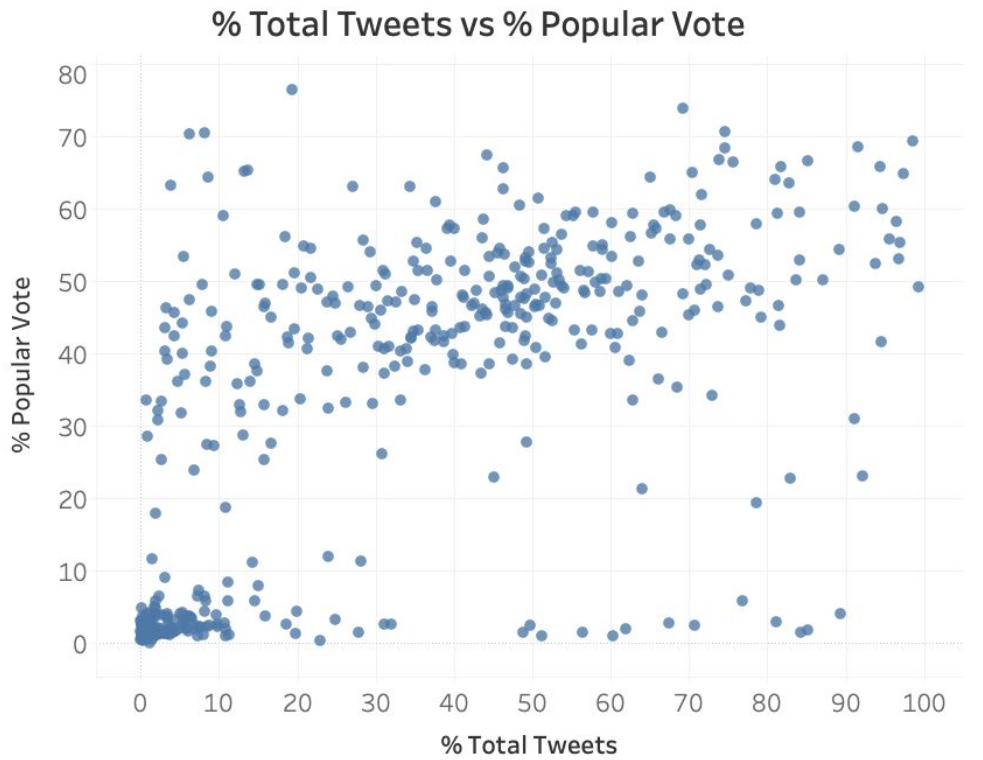
Presidential trends



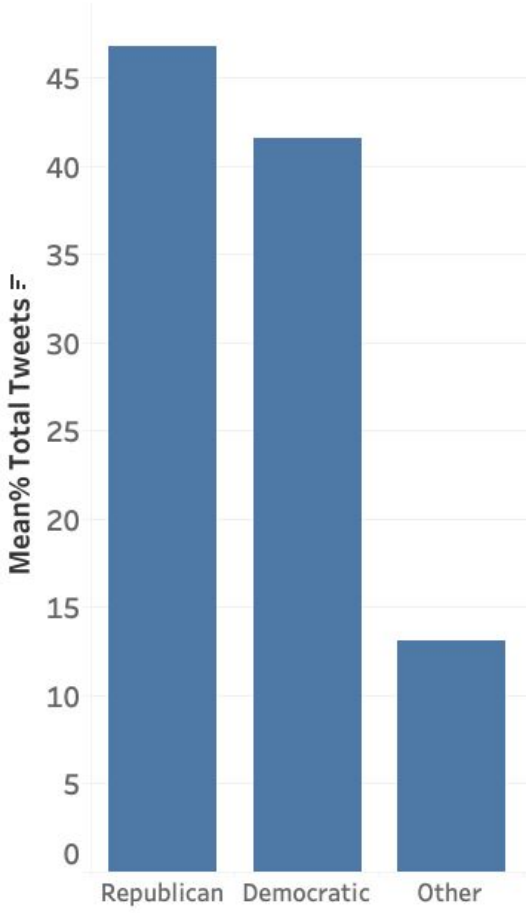


Gubernatorial trends

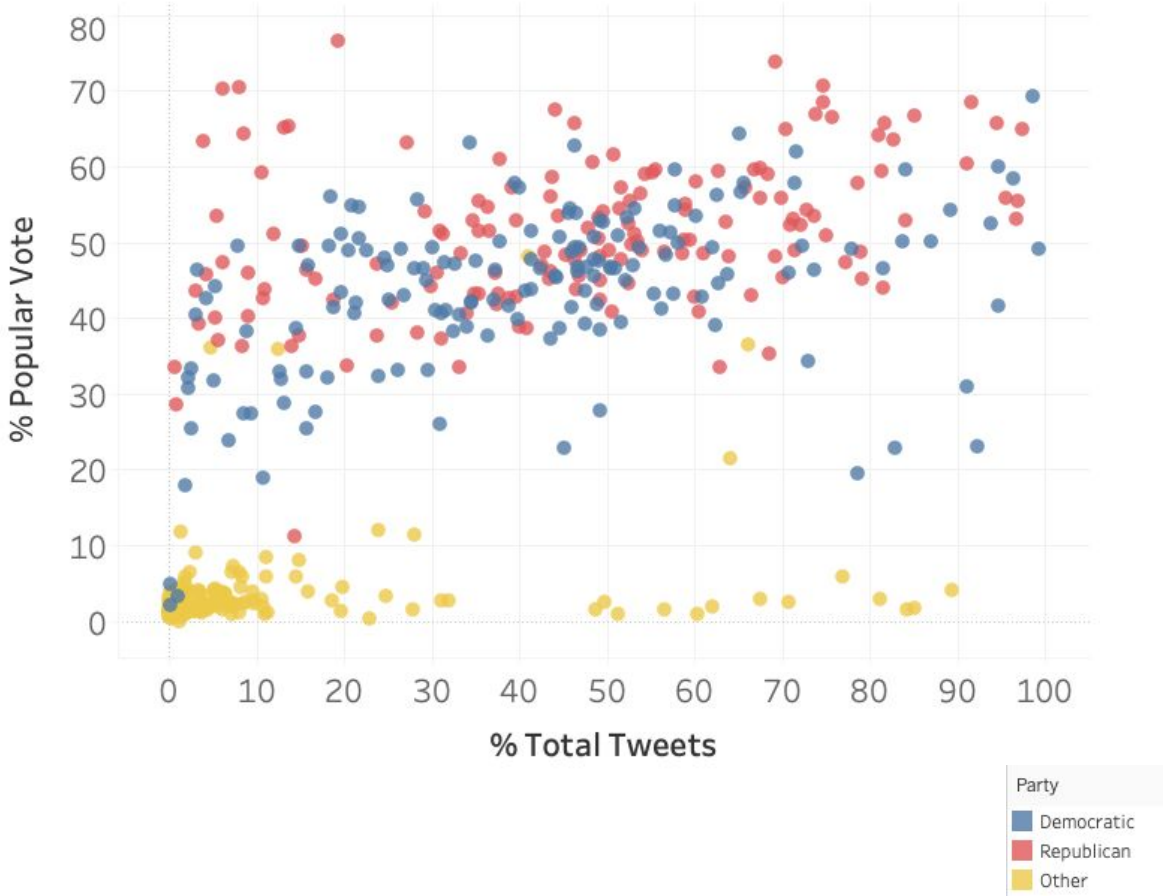




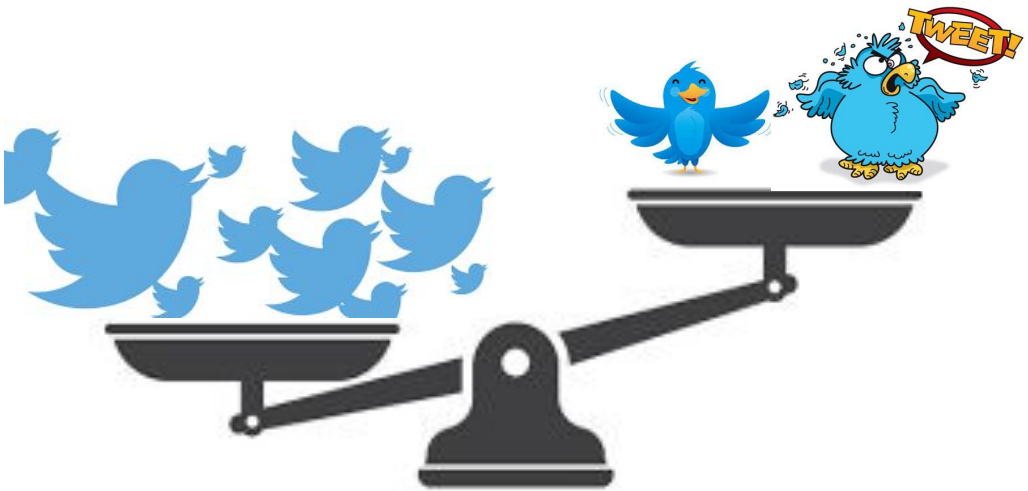
Mean % Tweets by Party



% Total Tweets vs % Popular Vote by Party

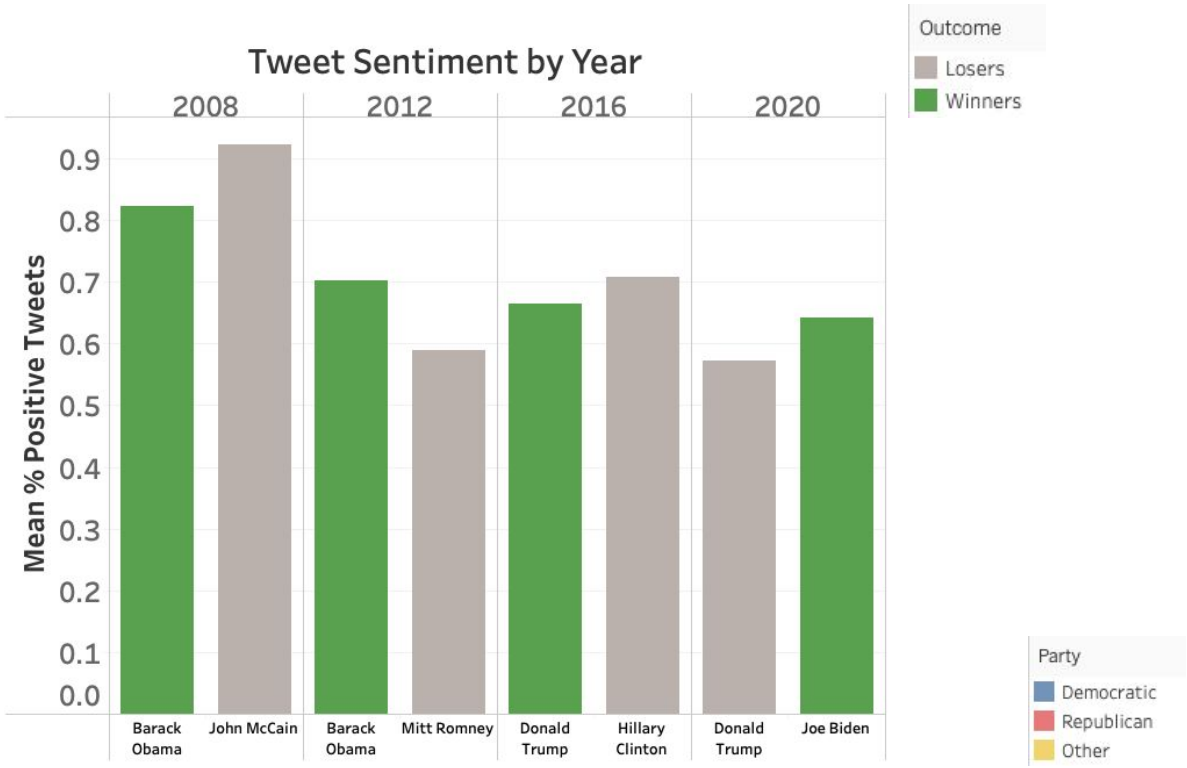


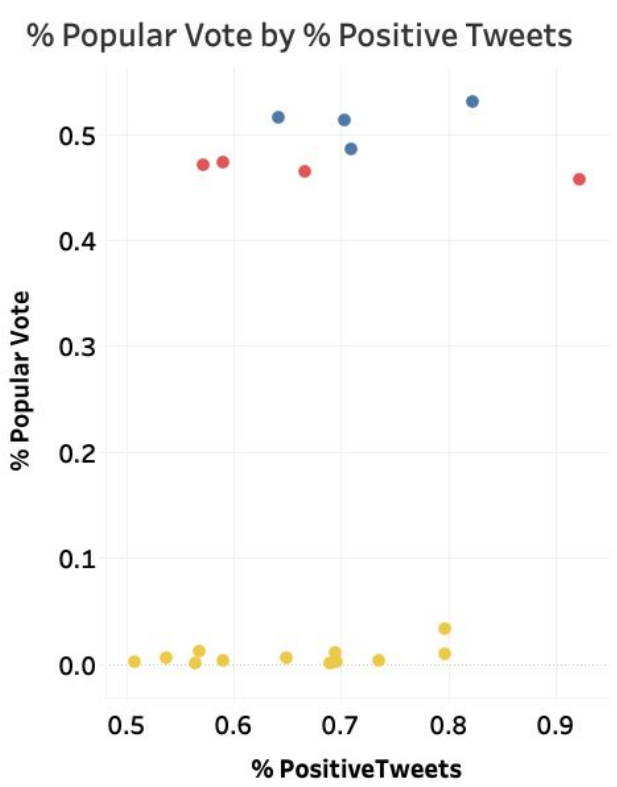
What about sentiment?



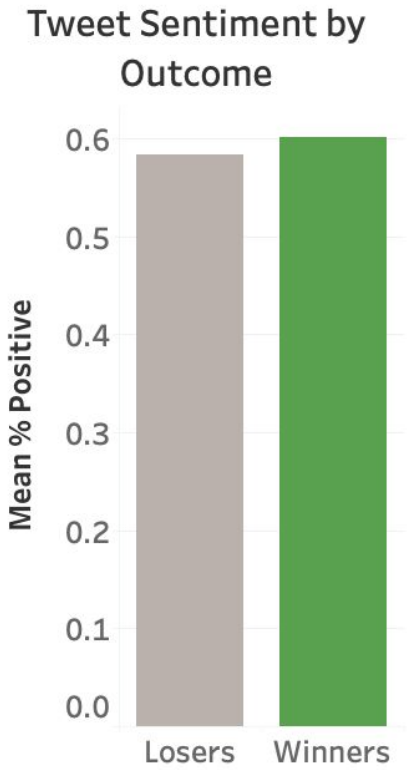
Which has stronger predictive power?

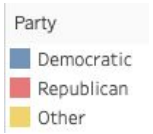
Sentiment analysis (Presidential)

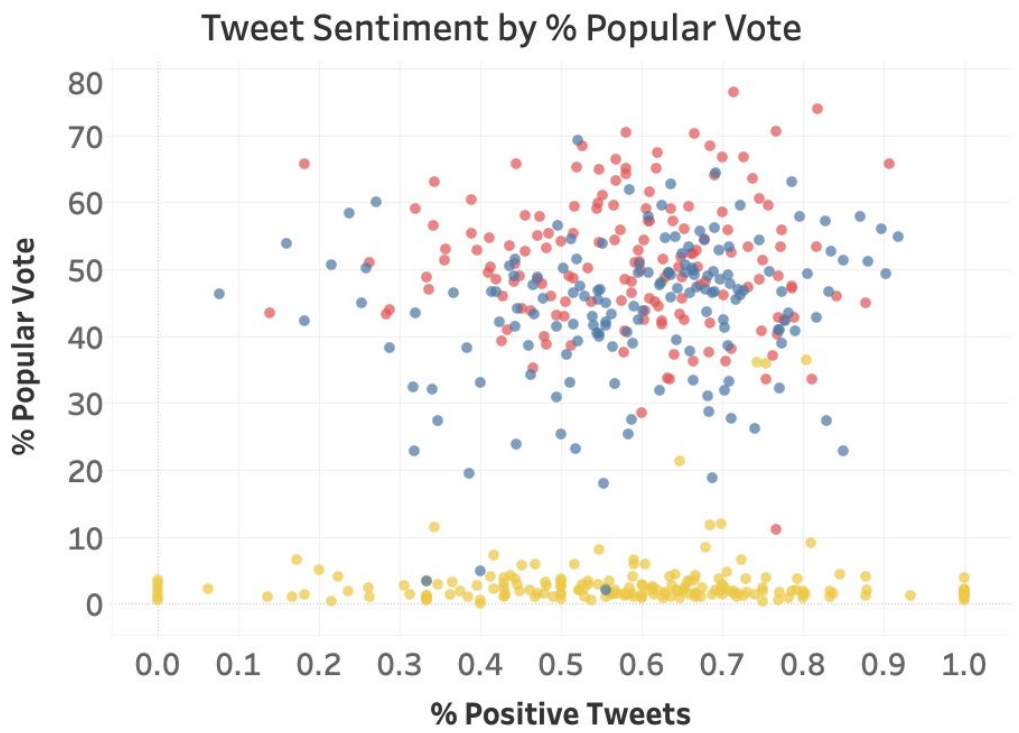


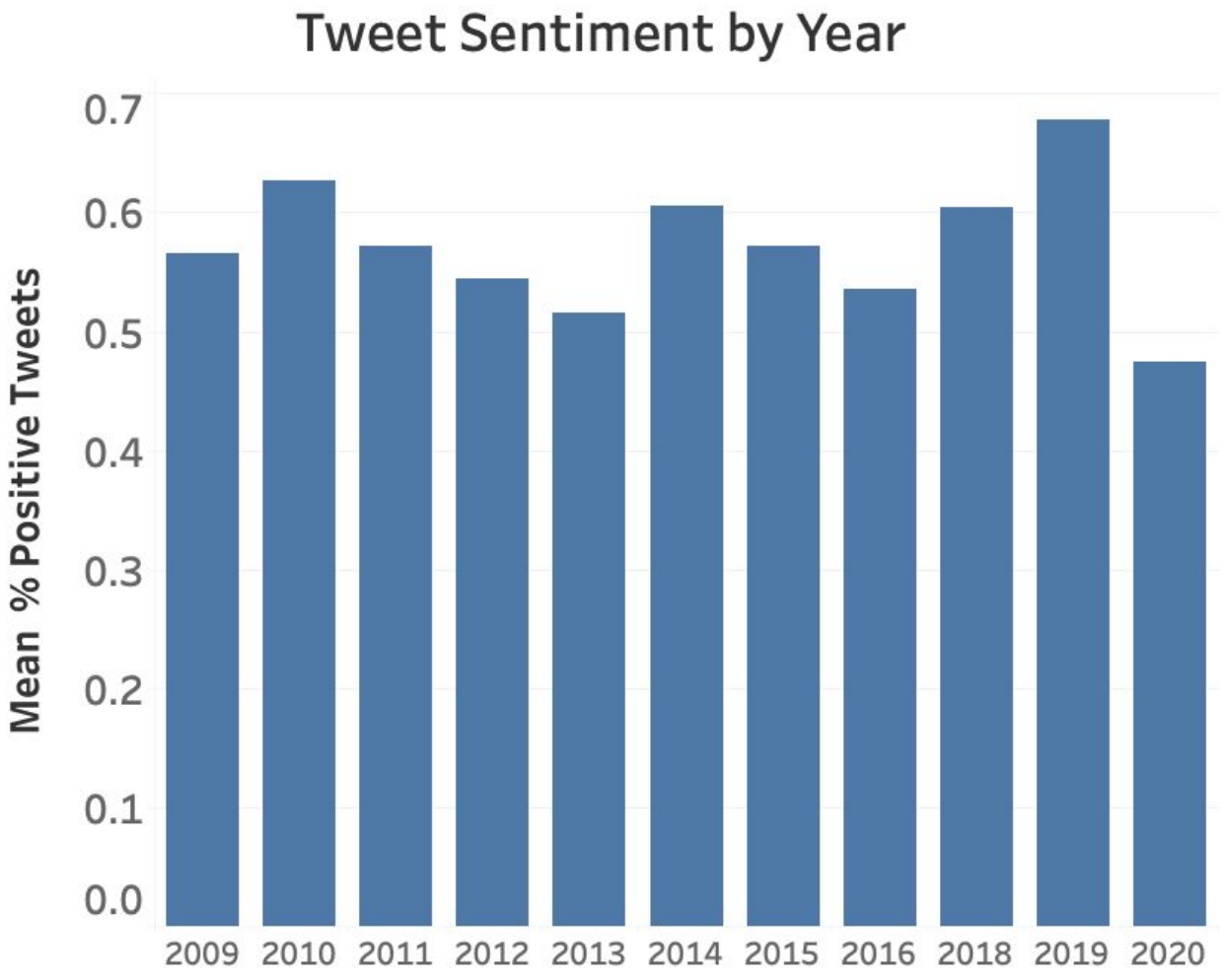


Sentiment analysis results (Gubernatorial)








Tweet Sentiment by Year
Mean % Positive Tweets
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
2009
2010
2011
2012
2013
2014
2015
2016
2018
2019
2020

Building a model to predict Gubernatorial election outcomes

- Linear Regression
- Decision tree
- KNN
- Random Forest

Which model performed best?

Predicting % Popular Vote:

Random Forest:

- Cross validation score: .92
- Rmse: 6.57

Predicting correct outcomes:

Decision Tree:

- Predicted correct winner in 29/36 elections

Most important feature: % total tweets

Conclusions

Problems:

- overlap of demographics of political tweeters and demographics of voters
- vocal minority