# Real-time Prediction of 2020 US House Elections Using Twitter Data

Andrea Jaba (adjaba) and Soumya Ram (soumyar)

## Abstract

Even though research has been conducted on predicting US House election results using social media, the conclusions are mixed. Our objective was to develop a data science pipeline - from data gathering, cleaning, integration, to using a machine learning model for predictions - to determine results for tossup district elections using Twitter data. Furthermore, a clear, easy to use visualization was developed, that could be updated real-time with limited manual help.

## Introduction

Predicting elections is a perennial problem in U.S. politics. Developing a good predictor can yield insights about how elections are won, which can also yield insights on campaigns. Furthermore, developing a real-time visualization can help people see these predicted results easier.

After exploring different types of US elections, a decision was made to focus on predicting for US district elections. Since district elections are numerous, they provide a data-rich field to train the model. Such an advantage is not offered with presidential and state elections. We chose to look at consecutive elections as the influence of social media would be most consistent between them.

However, most districts maintain the same party affiliation after every election. Therefore, the chosen focus was on predicting the results of tossup district elections.

There has been some research into using social media data to predict elections around the world. There are many convenient aspects to using social media data. It can provide a real-time signal on how a candidate is performing, and unlike polling, it is much easier to collect.

However, the literature on the efficacy of social media to predict elections is mixed. Previous work used aggregated facebook polling to accurately predict 74% of US House elections and 81% of US Senate races[1]. Ironically, though, the district predictions would be more accurate if the model had simply outputted the winning party in the previous election. Over 90% of districts consistently maintain the same party affiliation. Other research found a strong correlation between the number of tweets about a candidate and pre-electoral polling, and report a very low Mean Average Error of 1.65%[3], but a paper replicating this approach did not succeed[2]. Yet another paper finds that there is not a significant correlation between the sentiment analysis of Twitter data and pre-electoral polls[4].

This question was chosen for further investigation. Because of the real-time convenience of social media, predictions were incorporated into an interactive visualization where users can see the probability of each party winning across different districts. In addition, they can view the probability a candidate will win over time.Thus, the focus of the project became: Is it possible to generate and visualize real-time predictions of tossup US House elections using Twitter data?

## Dataset information

The project used a dataset from Harvard Dataverse, which contains the names and total votes of the candidates of all House midterm elections from 1976-2018. From this dataset, the 2014 data became the training set, 2016 data became the validation set, and 2018 data was the test set. To identify the swing districts for each of the three elections, Cook Political Report predictions were used.

To predict the 2020 midterm elections, a dataset of possible candidates by district was manually compiled. To obtain the list of swing districts, Cook Political Report predictions were used.

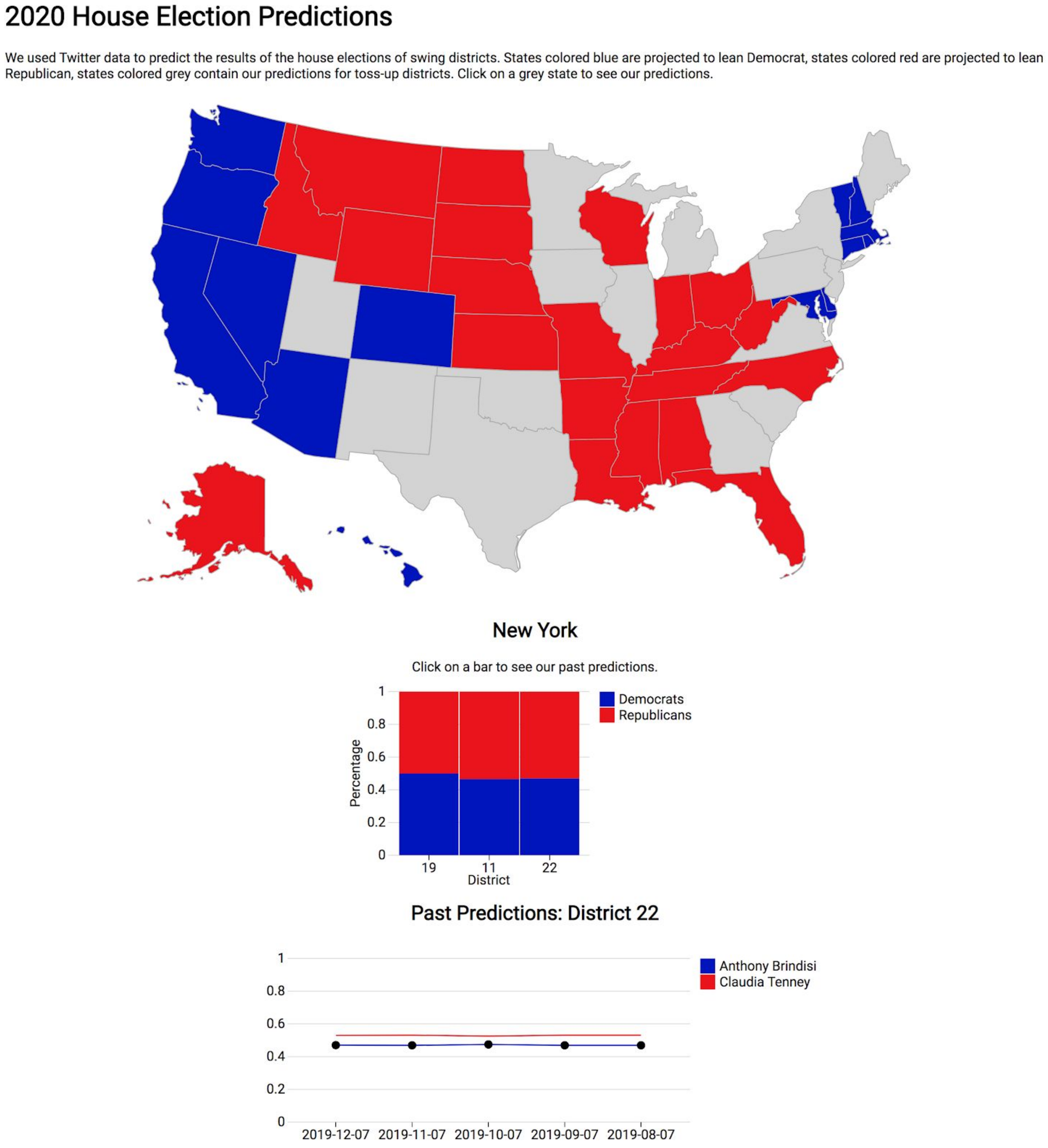
## Approach

### Data Mining

Using the package twitterscraper, all tweets mentioning a candidate over the course of a day were gathered. Attempts to mine tweets over longer periods of time resulted in inefficiencies in memory and a substantial delay due to lack of computational power. Afterwards, Wikipedia tables were scraped automatically to extract close districts. Then, the sentiment of the tweets were analyzed using the NLTK library, a leading library for sentiment analysis. Using the results of the sentiment analysis, feature vectors were constructed per candidate containing the average “positivity” of these tweets, the average “negativity” of those tweets, the average “neutralness” of those tweets, and the number of tweets overall.

These feature vectors were used as inputs into a support vector machine (SVM) model. The model that had the highest accuracy on the validation set (2016) was chosen among the 600 different SVM models that were trained on the training (2014) set.

## Visualization



2020 House Election Predictions

We used Twitter data to predict the results of the house elections of swing districts. States colored blue are projected to lean Democrat, states colored red are projected to lean Republican, states colored grey contain our predictions for toss-up districts. Click on a grey state to see our predictions.

New York

Click on a bar to see our past predictions.

Past Predictions: District 22

### *Visualization Functionality*

The website is divided into three sections. The topmost part of the website is a USA map separated by state, and colored by political inclination. States were colored blue and red based on the inclinations of its districts manually compiled from Wikipedia and grey if the state has swing districts with our predictions.

When the user clicks on a grey state, a bar chart shows up displaying the predictions for each of the swing districts. In particular, the website displays a bar chart containing the top 2 predicted candidates and the probability of each person winning if they battle each other for each swing district. This data is based on the predictions using the latest collected Twitter data. The user can hover over a bar to see the candidate's name and the exact predicted winning probability.

Furthermore, the user can also click on each bar to see past predictions for each district. As of now, these predictions were obtained by gathering data and running the model on data from one day each month from the past months.

### *Visualization Structure*

Javascript, HTML and CSS was used to build the frontend, and Python Flask to build the backend. For graphs, the D3 library was used to make the US maps, bar charts and line graphs. Existing code online was used to get a visualization of the US state map, but the coloring of US map based on political inclinations, and creation of bar charts and line graphs were done from scratch. Bar charts were chosen so that the user can see the split in each district, and line graphs were chosen to display past predictions to emphasize on how predictions on each candidate has changed over time.

**Challenges**

The first challenge was acquiring a Twitter developer account. Since this process was not progressing, the package twitterscraper was chosen to scrape tweets.

The second challenge was data compilation. Code was written to transform Wikipedia tables into a usable format. In addition, for the 2020 congressional elections, the total list of candidates running was too long to collect data for. Thus, manual screening occurred across  all districts to find the most promising candidates.

For the visualization, D3 ended up being surprisingly more difficult to learn than expected. As a team, we were not very familiar with Javascript, CSS, and HTML, so we had a steep learning curve in the beginning.

**Results**

The SVM model had a 57% accuracy on the test set (2018), a 45% accuracy on the val set (2016), and a 97% accuracy on the training dataset (2014). Clearly, the model had overfitted on the training data. When analyzing the different components of the model, the conclusion was that the sentiment analysis worked well on tweets: it returned an answer that concurred with human judgement. However, the majority of the predictions were very close to 0.5. This shows that the training process had simply selected a separator through the dataset. But because the districts were tossup districts, these results did not generalize well. Thus, the results show that relying on Twitter alone to predict tossup districts is a difficult task: these districts may be too close to call.

## Discussion

### Benchmarking

Existing social media based methods predict the Congressional elections with 74% accuracy, and 538 predicts Congressional elections with 96% accuracy. However, these focus on all elections while the current focus was only on the closest elections. Thus, their accuracies are not representative.

In addition, the papers that focused on election prediction using Twitter had collected several thousand tweets per candidate. However, we scraped tweets for the candidate at several randomly chosen days before the election. Overall, we had approximately ~200 tweets per a candidate compared to the ~20,000 described in the papers.

This design decision also affected the visualization. If no tweets about a candidate occur on the chosen day, the candidate is assigned the zero feature vector. This leads them to have a zero probability of winning in the visualization. Thus, the visualization can have sharp contrasts over time points as the candidate's feature vector changes from zero to nonzero.

## Further Work

There are two avenues identified for future work in prediction. Firstly, the model can be trained on all US House Districts. This accuracy can provide a benchmark against the past work on predicting district elections through Facebook polling.

In addition, it is possible to explore what aspects of an election Twitter can predict. It may not be able to accurately predict the final result, but it could predict attributes such as the degree of polarization of the candidate, or the degree of engagement of their supporters. These attributes can serve as components to a more expansive model.

For future work in the visualization, more information can be displayed such as additional information on the sentiments (ex: volume of tweets, sentiment breakdown), the breakdown of district inclinations for each state (i.e. exactly how many districts lean towards each party), as well as an overall summary of House election results if the predictions were accurate (i.e. how many seats represent each party).

## Citations

1)A. Carr, “Facebook, twitter election results prove remarkably accurate,”Fast Company, 2010, http://bit.ly/dW5gxo.

2)B. O’Connor, R. Balasubramanyan, B. R. Routledge, and N. A. Smith,“From tweets to polls: Linking text sentiment to public opinion time series,” in Proc. of 4th ICWSM. AAAI Press, 2010, pp. 122–129.

3)A. Tumasjan, T. Sprenger, P. G. Sandner, and I. M. Welpe, “Predicting elections with twitter: What 140 characters reveal about political sentiment,” in Proc. of 4th ICWSM. AAAI Press, 2010, pp. 178–185

4)Metaxas, Panagiotis T., Eni Mustafaraj, and Dani Gayo-Avello. "How (not) to predict elections." 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing. IEEE, 2011.