Mr. Steal Yo Elections

Mikias Alemu, Bhaanu Kaul, Divyansh Gupta, Zakk Lefkowitz

Introduction

This project focuses on harnessing the power of a vast microblogging network, Twitter, to gain insight into the political leanings of the American people. Traditionally, polling has been used to achieve the same goal, yet polling has many shortcomings - notably the immense cost to reach a large audience. Twitter bypasses this by providing simple access to billions of users' thoughts, opinions, feelings, etc. By applying modern natural language processing and machine learning techniques to Twitter data, we are able to produce quantitative data about the American people's emotions and how they correlate to voting in presidential elections.

Objectives

This project discusses the correlation, if any, between user sentiment and political affiliation based on Twitter tweets around the United States. We strived to answer the following questions:

- What is the, if any, correlation between a person's emotions and their political leanings?
- How can a correlation between sentiment and political affiliation be used to sway voters toward a certain party or candidate?
- Can this correlation, if any, be used to reproduce the results of previous elections or predict future elections?

Data Type and Acquisition

We got access to about 450 million individual tweets from both the 2012 and 2016 presidential elections. These tweets came from datahub.io (2012 election) and the Harvard Dataverse (2016 election). The tweets from both sources were prefiltered using keywords based on topics pertaining to each election.

Tweets from 2012 were tweet objects directly from the Twitter API. Tweets from 2016 were unique identifiers for individual tweet objects. To retrieve the tweets from 2016, we would have had to access the Twitter API which is heavily rate limited. This caused us narrow the scope to the 2012 election data.

We used the following fields for our analysis:

Tweet	User
Text	Follower Count
HashTags	Friend Count (Users a person follows)
Created At Date	Location
Retweets	Tweet Count
	Verified Status

Sentiment Analysis

We used Python's TextBlob Library.This library is built off of NLTK, a popular Python Natural Language library.

This library was trained extensively on IMDB movie reviews.

Sentiment Classifications: *Positive, Negative, Neutral*

Political Classification System

Training sets were created from known Republican, Democratic Twitter accounts, including verified key republican, democrat and neutral spokespersons and news anchors.

Decision Tree Classifier

Accuracy: 50% (9,000 Tweet Training Set)

Output classification: Republican, Democrat, Third Party

NLTK Naive Bayes Classifier:

Accuracy: 70-75% (10,000 - 15,000 Tweet Training Set)

Output Classification: republican probability, democrat probability, third party probability.

Scikit Logistical Regression (In Progress):

Accuracy: 80 - 85% (9,000 Tweet Training Set)

Output classification: Republican, Democrat (no neutral)

Infrastructure

Classification required massive compute power / memory.

First laptop Design - 100 tweets per minute.

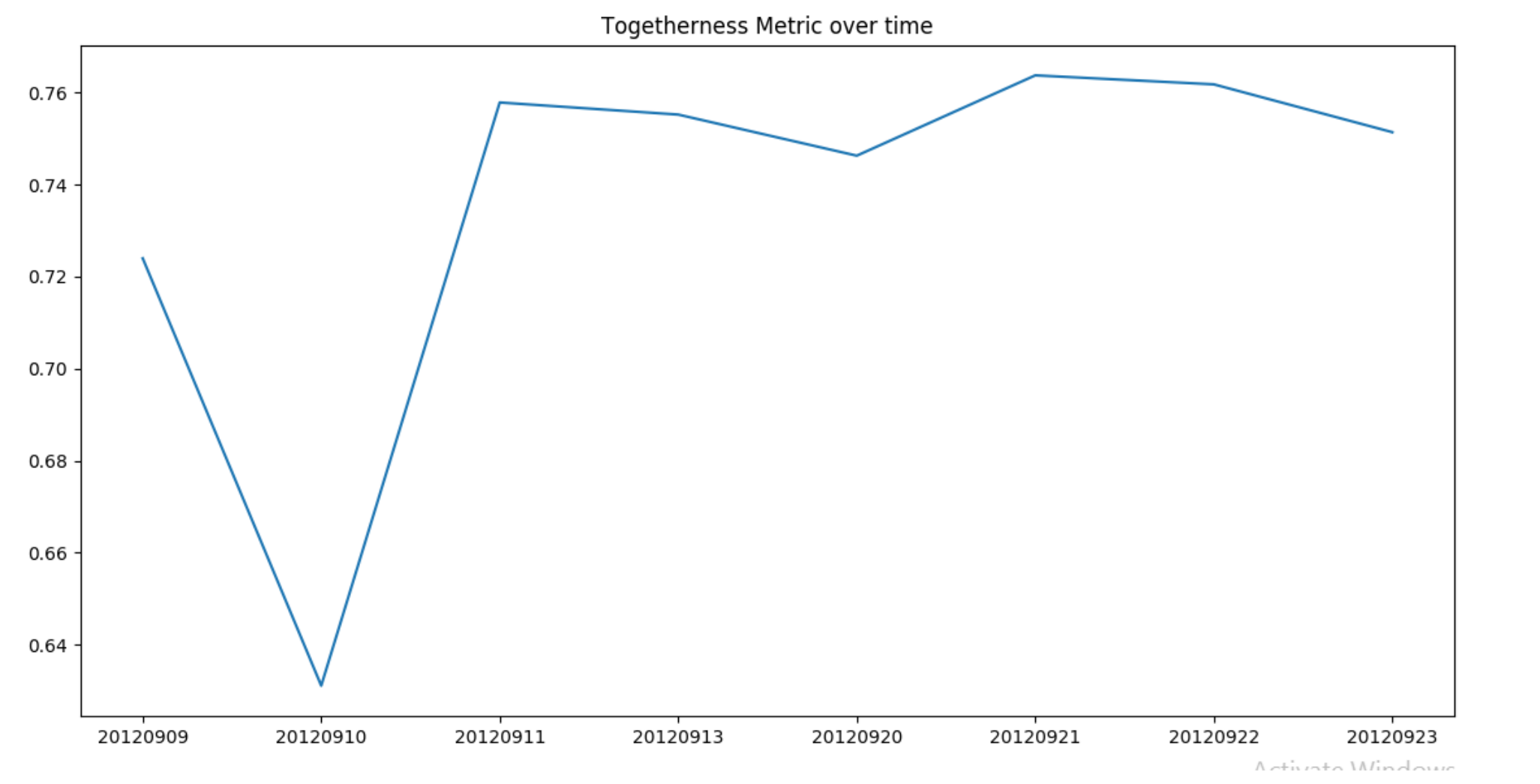
Second Multithreaded design - 400 tweets per minute.

AWS Tweet Processing Pipeline - Nearly infinite throughput.

Utilized AWS to create an auto-scaling queue service that requires simple HTTP requests to process tweets into MySQL.

Processed over 5 million tweets in 3 days (limiting factor: $).

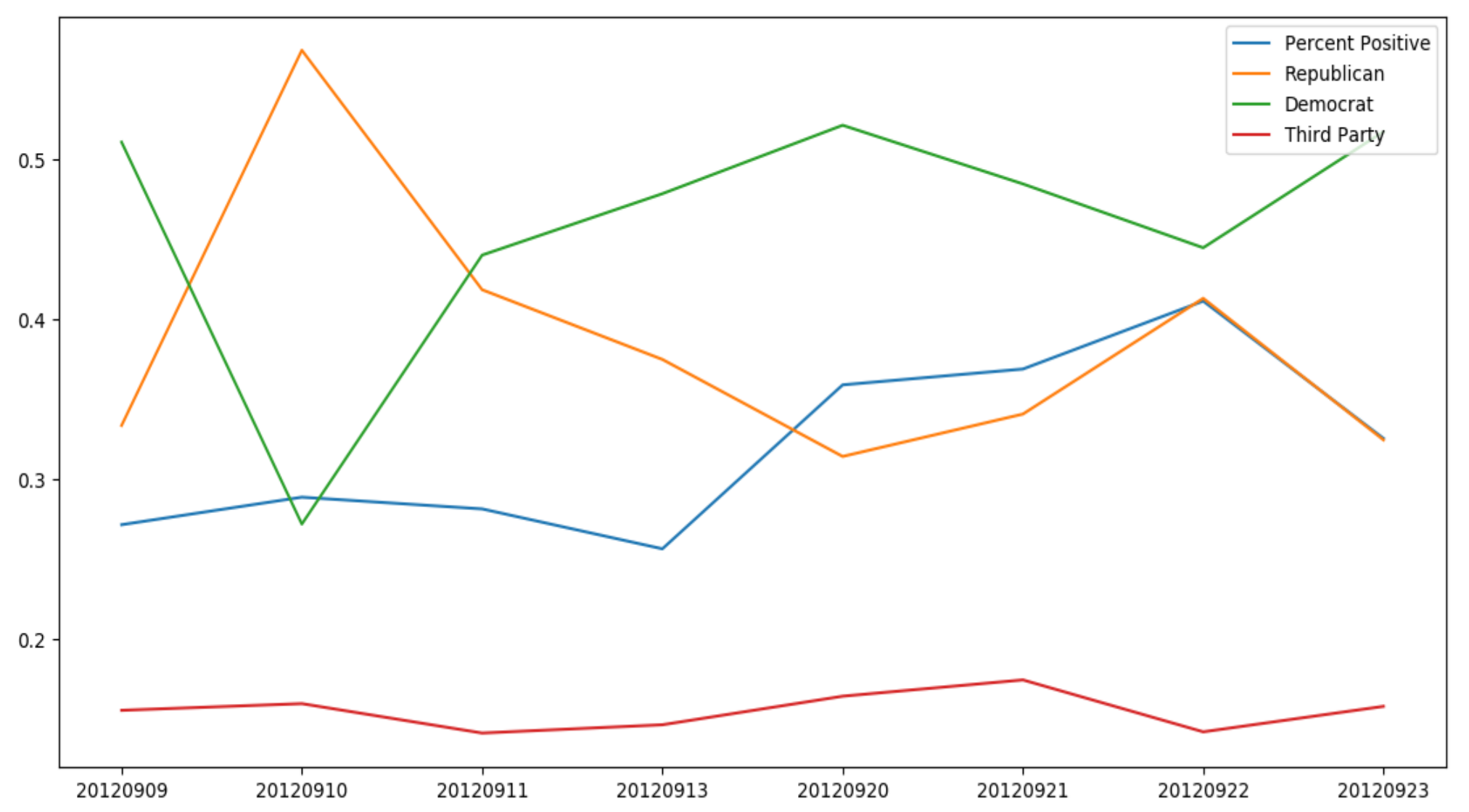
Models and Metrics



Togetherness Metric:

1 - max difference of probabilities = Togetherness

How close together are the people on Twitter in their Political opinions?



The plot above is a simple example of how political leaning can be correlated with overall sentiment. We can see that as percent positive (sentiment) decreases toward the end, there is a decrease in the number of republican tweets and an increase in both third party and democratic tweets.

Compared to the middle of the graph, however, as sentiment increases, democratic tweets increase and republican tweets decrease. Since this graph is from initial data, we cannot effectively correlate sentiment and political affiliation.

We found the average tweet in our dataset to have these averages (we suspect a liberal bias because of Twitter's users):

Republican Prob.	Democratic Prob.	Third Party Prob.
0.357	0.446	0.196

2012 Presidential Electoral College Model

We combined Twitter location data and our tweet political classification data to determine how certain states were likely to vote during the 2012 election. We found that the model grew more accurate as more tweets were processed, and our end result using the Naive Bayes Classifier was as follows:

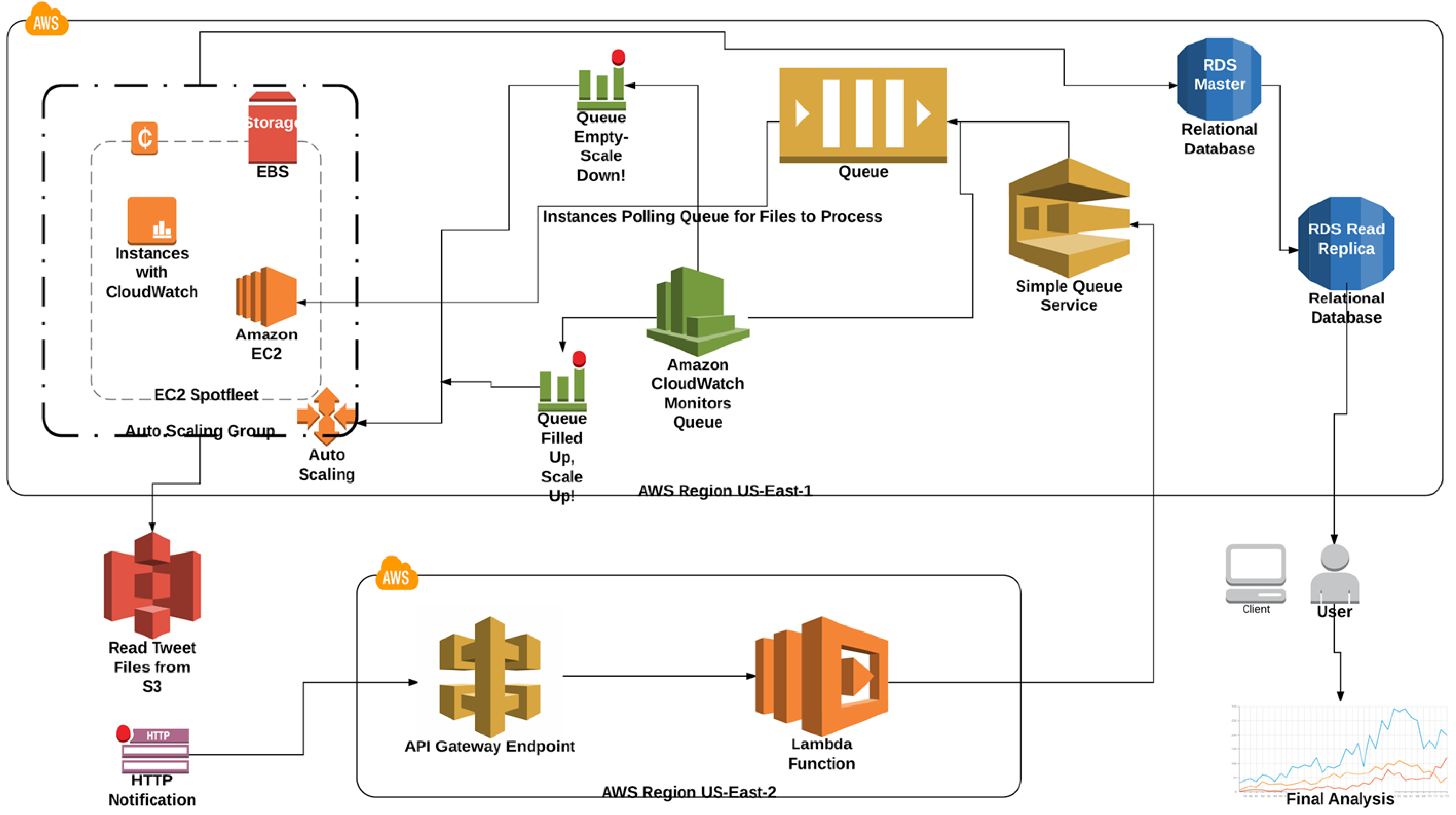
335+/538 Electoral Votes Correctly Predicted
31/51 States/Districts Correctly Predicted

Conclusion & Lessons Learned

Conclusions based on our initial data were not very promising. We initially trained a Naive Bayes classifier on relevant tweets from 2017. This proved to be ineffective as 2012 political topics and tweets were significantly different than todays. We retrained the Naive Bayes classifier with 2012 data, but only achieved minimal increase in accuracy. We believe the Naive Bayes Classifier is inaccurate because it looks at each keyword in the tweet, instead of the tweet in context.

We are continuing our analysis of the twitter data; however, instead of using the Naive Bayes classifier we are switching to the logistical regression classifier. From initial testing on a small dataset, we saw an increase of about 10% accuracy. We are in the process of rerunning our stored tweets through this new classifier.

Overall, we could have done more research on the accuracy of various classifiers and chosen an appropriate one from the start. Along with a new classifier, analyzing a larger timeline of activity. Our current timeline consists of 38 days of activity.



Works Cited

Yaniv Altshuler, Wei Pan, Alex (Sandy) Pentland (2012) Trends Prediction Using Social Diffusion Models [PDF]. *MIT Media Lab*. Retrieved from http://web.media.mit.edu/~yanival/SBP-Behavior-shaping.pdf

Jungherr, A. (2015). Twitter as Political Communication Space: Publics, Prominent Users, and Politicians. *Analyzing Political Communication with Digital Trace Data Contributions to Political Science,* 69-106. doi:10.1007/978-3-319-20319-5_4

Ringsquandl, M., & Petković, D. (2013). *Analyzing Political Sentiment on Twitter* [Scholarly project]. In *University of Applied Sciences Rosenheim*.

"Twitter 2012 Presidential Election." *The Datahub*. Kingmolnar, 08 July 2015. Web. 1 Apr. 2017. <https://datahub.io/dataset/twitter-2012-presidential-election>.

Littman, Justin; Wrubel, Laura; Kerchner, Daniel, 2016, "2016 United States Presidential Election Tweet Ids", doi:10.7910/DVN/PDI7IN, Harvard Dataverse, V3