# Political Party Predictor

Karina Sanchez, Elijah Freeman, Alexandra Camargo

## Motivation

- 2020 Presidential Election
- Polarization between Democrats and Republicans in U.S. political climate
- Twitter as a critical part of politician's political campaigns and method of connecting to the community they serve
- Differentiated social media sentiment between political parties as shown through the results of a classifier

## Problem Definition

- Given a tweet from a politician, determine whether the politician is **Republican** or **Democratic**
- Politicians in our dataset: current senators, 2020 presidential candidates, current members of congress, current representatives


## Challenges

- Dataset generation using web scraping and the Twitter API and creating our own tokenizer
- Model selection: Naïve Bayes, Logistic Regression, SVM, and Unigram RandomForest
- Deciding the best combination of NLP techniques for model improvement: removing stop words, adding negation tags, n-grams

## Approaches

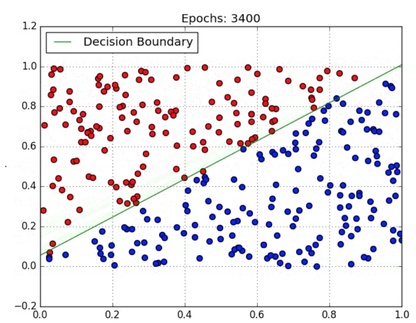
**Dataset:** a list of tokenized tweets with all punctuation except hashtags and @'s. We found that while most NLP methods remove all punctuation except delimiters, #'s and @'s tend to be a strong indicator of political party.

**Model Selection:** we used five fold validation in order to select the best classifier, n-gram, whether or not to choose stop words, and whether to use smoothing in our model.

**Stop Words:** words such as 'and', 'the', 'at', 'in', etc. are very common in the English languages and therefore tend to add noise to sentiment classifiers. Removing stop words improved average accuracy in all classifiers.

**N-Grams:** we tested out using the presence of individual (unigram), pairs (bigram), and triples (trigram) of words in each tweet as input features to our various classifiers for cross validation.

**Logistic Regression:** a classification method under 'Supervised' ML that uses a logistic function to model a binary dependent variable. We used Scikit-learn to model our logistic regression.




## Results

**The final training / test set included:**

- 37,798 / 9,345 Democratic Tweets
- 29,370 / 7,343 Republican Tweets

**Filtered Unigram Logistic Regression Results::**

- **Precision:** 0.7924827188940092
- **Recall:** 0.7687901648505169
- **Accuracy:** 0.8144774688398849

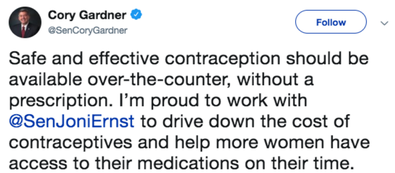
| | Classified as Democrat | Classified as Republican |
|---|---|---|
| **Actual Democrat** | 8,089 | 1,655 |
| **Actual Republican** | 1,441 | 5,503 |

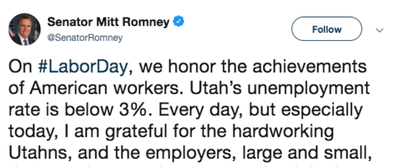
## Analysis

### Correctly Classified



### Incorrectly Classified




Large dataset and political sentiment's heavy dependence on current events lead to our training set, which consisted of a politician's entire twitter history, to generate underfitted results.

## Acknowledgements & References

We would like acknowledge the staff of CS 221, particularly our mentor Marcus Pålsson.

https://towardsdatascience.com/random-forest-and-its-implementation-71824ced454f
https://towardsdatascience.com/introduction-to-logistic-regression-66248243c148
**Scikit, NLKT, Tweepy, BeautifulSoup**