# Classifying Party Affiliation Based on Campaign Rhetoric

Rory Fitzpatrick (*roryfitz*), Garrett Merz (*gwmerz*), Julia Pakela (*jpakela*)

Machine learning has been vastly applied to natural language processing. We report the performance of several classification methods used to determine political party based on campaign rhetoric. We use corpora of campaign speeches from the 1960, 2008 and 2016 elections cycles. Finally, we investigate the use of classification performance as a metric for quantifying party polarization as a function of time.

## INTRODUCTION

In the era of the 24/7 news cycle, political rhetoric has transformed into the stringing-together of sound bites to be swept up by the media and repeated out of context *ad infinitum*. Political speech is rife with partisan buzzwords, a rhetorical structure that may lend itself to classifying party affiliation based on political discourse. In addition to providing a unique classification scheme, party affiliation classifiers could provide a metric for quantifying political polarization in a given election cycle or administration. We show that we can accurately classify political campaign speech based on the party of the speaker, and in turn use the classification performance over time to quantify party polarization.

## DATASETS AND DATA PROCESSING

We use bipartisan campaign rhetoric curated in [4] from the 1960, 2008, and 2016 election cycles. We selected these years because they each contained a large sample of speech from both political parties and occur at dramatically different points in the country's political history. We excluded, for instance, the 2012 election cycle, because the speeches collected from the Obama campaign are all variations of the same stump speech, and would bias the classifier. Initially, words contained in the Python `nltk` 'stopwords' and 'names' corpora are removed from each document. We acknowledge that there are still many names and proper nouns that are included in our resulting vocabularies. Further study could be done on the effect of removing or keeping all proper nouns (e.g. one might expect "Reagan" to be more common in Republican speeches) but it is outside the scope of this report.

We lemmatize and stem using the `nltk` `WordNetLemmatizer` and `PorterStemmer` to reduce variations in words. No effort is made to correct typos. The resulting texts are converted to the bag-of-words format using three different frequency measures: **(1)** standard term frequency (*tf*), **(2)** presence or absence of a word (*bool*) and **(3)** a frequency measure known as term frequency-inverse document frequency (*tfidf*). The *tfidf* frequency is calculated as follows:

$$w = f_{t,d} \ln(1 + N/n_t), \qquad (1)$$

where $f_{t,d}$ is the frequency of the word in the current speech, $N$ is the total number of documents in the training set, and $n_t$ is the number of documents in the training set that contain the word. We have also chosen to add one to the argument of the log for smoothing. We do not restrict the size of the bag-of-words vocabulary.

| | 1960 | 2008 | 2016 |
|---|---|---|---|
| Democrat | 598 | 362 | 76 |
| Republican | 311 | 222 | 82 |

TABLE I. The number of speeches available for each election cycle.

## CLASSIFICATION METHODS

Similar studies have been done using floor speech from the Senate and House of Representatives in [7] which uses the popular text classification methods, Naive Bayes (NB) and Support Vector Machines (SVM) However, it was shown in [2] and [5] that these methods are sensitive to outliers. We speculate that this is why in [7] it was found that House speeches were better suited than Senate speeches for training party classifiers - the House is considered to be more partisan than the Senate, and documents from this group would presumably contain fewer outlier events. In addition to the naive methods for classification, we implement robust methods for SVM as described in [1] and [6] to investigate whether it is possible to design a classifier which is better suited for training on texts from more moderate speakers. We also test classification methods less traditionally used for text classification: logistic regression (LR) and $k$-nearest neighbors (KNN). We began by training an cross-validating within a single election cycle for each classification method and each word frequency measure. Then, we trained on a single election cycle and tested on the remaining two election cycles. We also report, based on the NB method, the words most indicative and Democratic and Republican speech in each election cycle and across all three cycles.

| Method | Freq. | 1960 | 2008 | 2016 |
|---|---|---|---|---|
| NB | *tf* | 92.0 | 81.1 | 91.7 |
| | *bool* | 92.7 | 83.0 | 92.9 |
| | *tfidf* | 91.1 | 83.9 | 93.0 |
| SVM | *tf* | 91.4 | 87.1 | 85.6 |
| | *bool* | 93.3 | 91.2 | 91.2 |
| | *tfidf* | 91.8 | 86.0 | 90.5 |
| LR | *tf* | ZZ | ZZ | ZZ |
| | *bool* | ZZ | ZZ | ZZ |
| | *tfidf* | ZZ | ZZ | ZZ |
| KNN | *tf* | ZZ | ZZ | ZZ |
| | *bool* | ZZ | ZZ | ZZ |
| | *tfidf* | ZZ | ZZ | ZZ |
| RSVM | *tf* | ZZ | ZZ | ZZ |
| | *bool* | ZZ | ZZ | ZZ |
| | *tfidf* | ZZ | ZZ | ZZ |

TABLE II. Cross validation scores by method and word frequency metric.

## RESULTS

### Cross validation

Results shown in Table II.

### Predictions between years

Are certain years more or less predictive of other years?

Need to make a table of results.

### Predictive words

Results shown in Table IV.

## CONCLUSIONS

Further study could be done to expand the dataset to examine the change of rhetoric from year to year and explore how long a particular election cycle might be predictive for future election cycles.

Another confounding factor may be the presence or absence of an incumbent running in the election for a particular year. This candidate would have no opponent within their own party and may speak differently as a result.

[1] Chandra, B. *et. al.* (2007). Robust Approach for Estimating Probabilities in Naive-Bayes Classifier. *International Conference on Pattern Recognition and Machine Intelligence*, 11-16.

[2] Kwon, N. *et. al.* (2006). Identifying and classifying subjective claims. *Proceedings of the 8th Annual International Digital Government Research Conference*, 7681.

[3] McCallum, A. & Nigam, K. (1998). A comparison of event models for naive Bayes text classification. *Proceedings of the 1998 Association for the Advancement of Artificial Intelligence Workshop on Learning for Text Categorization (AAAI98)*, 4148.

[4] Peters, G.,The American Presidency Project [online]. Santa Barbara, CA: University of California (hosted), Gerhard Peters (database). Available from World Wide Web: http://www.presidency.ucsb.edu/

[5] Thomas, M. *et. al.* (2006). Get out the vote: Determining support or opposition from Congressional floor-debate transcripts. *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP06)*, 327335.

[6] Xu, L. *et. Al.* (2006). Robust Support Vector Machine Training via Convex Outlier Ablation. *Proceedings of the 21st national conference on Artificial intelligence*, Vol 1, 536-542.

[7] Yu, B. *et. al.* (2008). Classifying party affiliation from political speech. *Journal of Information Technology & Politics*, 5:1, 33-48, DOI: 10.1080/19331680802149608

| Method | Training Year | 1960 (*tf*/*bool*/*tfidf*) | 2008 (*tf*/*bool*/*tfidf*) | 2016 (*tf*/*bool*/*tfidf*) |
|---|---|---|---|---|
| NB | 1960 | — | ZZ/ZZ/ZZ | ZZ/ZZ/ZZ |
| | 2008 | ZZ/ZZ/ZZ | — | ZZ/ZZ/ZZ |
| | 2016 | ZZ/ZZ/ZZ | ZZ/ZZ/ZZ | — |
| SVM | 1960 | — | ZZ/ZZ/ZZ | ZZ/ZZ/ZZ |
| | 2008 | ZZ/ZZ/ZZ | — | ZZ/ZZ/ZZ |
| | 2016 | ZZ/ZZ/ZZ | ZZ/ZZ/ZZ | — |
| LR | 1960 | — | ZZ/ZZ/ZZ | ZZ/ZZ/ZZ |
| | 2008 | ZZ/ZZ/ZZ | — | ZZ/ZZ/ZZ |
| | 2016 | ZZ/ZZ/ZZ | ZZ/ZZ/ZZ | — |
| KNN | 1960 | — | ZZ/ZZ/ZZ | ZZ/ZZ/ZZ |
| | 2008 | ZZ/ZZ/ZZ | — | ZZ/ZZ/ZZ |
| | 2016 | ZZ/ZZ/ZZ | ZZ/ZZ/ZZ | — |
| RSVM | 1960 | — | ZZ/ZZ/ZZ | ZZ/ZZ/ZZ |
| | 2008 | ZZ/ZZ/ZZ | — | ZZ/ZZ/ZZ |
| | 2016 | ZZ/ZZ/ZZ | ZZ/ZZ/ZZ | — |

TABLE III. CAPTION results for training on one year and testing on another

| Year | Freq. | Democrat | Republican |
|---|---|---|---|
| 1960 | *tf* | ZZ | coolidg, |
| | *bool* | ZZ | ZZ |
| | *tfidf* | ZZ | ZZ |
| 2008 | *tf* | ZZ | ZZ |
| | *bool* | ZZ | ZZ |
| | *tfidf* | ZZ | ZZ |
| 2016 | *tf* | ZZ | ZZ |
| | *bool* | ZZ | ZZ |
| | *tfidf* | ZZ | ZZ |

TABLE IV. The words most indicative of both political parties based on election year and word frequency metric.