Predicting Politician NRA Grades from Congressional Speeches

Anqi Ji Allison Koenecke Julia Olivieri

{anqi koenecke jolivier} @stanford.edu

Introduction

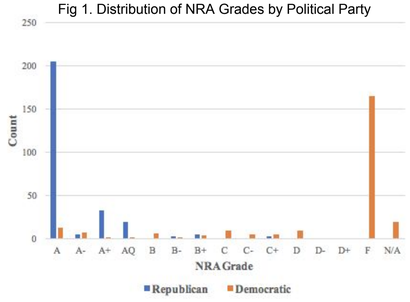
Gun control has consistently been at the forefront of American political discourse; it is generally accepted that Democrats lean pro-gun control, while Republicans are more likely to be in the National Rifle Association's good graces. However, it is quantitatively unclear to what extent political party alone can predict a congress person's NRA rating. Further, we pose the question: do politician speeches correspond to how they vote on guns?

We built multinomial logistic regression models that classify the output as one of four NRA rating categories: A, B, C/D, and F. The features used as inputs include politician data (party, age, state, gender) and featurized congressional speech texts.

Data

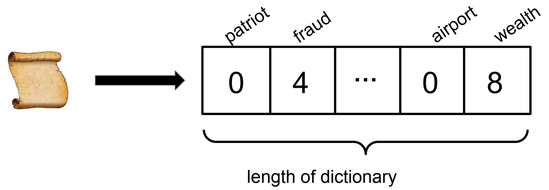
- Speeches scraped from Congressional Record [3] (22,363 speeches for 331 candidates from 2011-2016)
- NRA grades as of 2013 (A+ to F) and politician information scraped from Propublica database (15 ratings for 536 politicians)
- Merging both datasets yields 21,785 speeches for 316 politicians

Fig 1. Distribution of NRA Grades by Political Party



Features

Frequency-based:



Congressperson features of political party, state, age, and gender are used as a baseline metric for predicting NRA rating. Derived features include text classification word vectors, and ensuing model predictions. A two-level model (text classification, then multinomial regression) uses all features to reasonably weight speeches against personal data. The output is also derived: NRA ratings are bucketed such that similar ratings (e.g. B+, B, and B-) into one category (e.g. B).

Models

Text classification is done via four methods: three variants of Naive Bayes, and also using Support Vector Machines.

Naive Bayes is given by $\hat{y} = argmax_{k\in\{1,\dots,K\}}\, p(C_k)\prod_{i=1}^{n} p(x_i|C_k)$ [4]
Naive Bayes Assumption: All features are independent with each other, i.e.

$$p(C_k|x_1,\dots,x_n) = p(C_k)\prod_{i=1}^{n} p(x_i|C_k)$$

where x_i^l indicates i^{th} feature of l^{th}example, classes are $\{C_1, C_2, C_3, C_4\}$, m = # of training examples, n = # of features

(1) Gaussian Naïve Bayes: $p(x_i|c_k) = Gaussian\,(\mu_{x_i,c_k}, \Sigma_{x_i,c_k})$
where $\mu_{x_i,c_k} = \frac{1}{m}\sum_{l:y_l=c_k} x_i^l$, $\Sigma_{x_i,c_k} = \frac{1}{m}\sum_{l:y_l=c_k}(x_i-\mu)^2$.

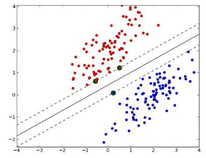
(2) Bernoulli Naive Bayes: $p(x_i|c_k) = P(i|c_k)x_i + (1-P(i|c_k))(1-x_i)$

(3) Multinomial Naïve Bayes: $p(x_i|c_k) = \frac{N_{c_k i}+1}{N_{c_k}+n}$.
where $N_{c_k i}$ is the number of times the i^{th} feature appears in class c_k, and N_{c_k} is the total count of all features for class c_k.

Support vector machines separate different classes by the largest margin possible.

(4) Linear SVM: Primal formulation is:
$\min P(w,b) = \frac{1}{2}|w|^2 + C\sum_i H_1[y_i\, f(x_i)]$ where hinge loss $H_1(z) = \max(0, 1-z)$ and $f(x) = w\cdot x + b$. [6]

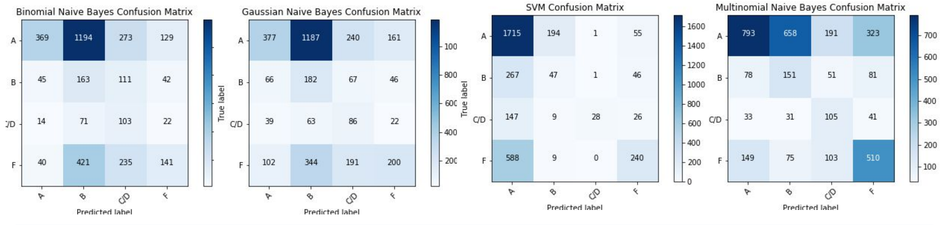
Fig 2. Illustration for SVM [5]



These models are compared to a baseline **multinomial logistic regression** to predict classifications based on a politician's attributes: political party, state, age, and gender. Iterations were also run to include test set text classification outputs as a feature.

(5) Multinomial Logistic Regression:
The model is given by $\ln\left(\frac{\pi_j}{\pi_J}\right) = \alpha_j + \beta_j X$ where there are J = 4 categories of the outcome indexed by j, and $\pi_j = \frac{1}{1+\sum e^{\alpha_j+\beta_j X}}$. [6]

All models are trained using 5-fold cross-validation, and then run on the held-out test set (20% of the full dataset size).



Results

Among text models, the Multinomial Naive Bayes and SVM models perform best on classifying the four NRA rating categories. However, politician-specific features are far better predictors of NRA rating; in fact, including the best Naive Bayes predictions as features in the multinomial regression reduces model accuracy.

Models	Features	Training Set Size	Test Set Size	Training Error	Test Error
Multinomial Regression	Party, State, Age, Gender	17,428	4,357	3.84%	3.72%
Gaussian Naive Bayes	Frequency Featurized Text	18,412	3,373	60.2%	74.9%
Bernoulli Naive Bayes	Frequency Featurized Text	18,412	3,373	64.3%	77%
Multinomial Naive Bayes	Frequency Featurized Text	18,412	3,373	38.9%	53.8%
Support Vector Machine	Frequency Featurized Text	18,412	3,373	48.9%	39.8%
Multinomial Regression, Multinomial Naive Bayes	Party, State, Age, Gender, Frequency Featurized Text	2,698	675	4.64%	6.22%

Discussion

The MNB and SVM algorithms perform the best on our data, which makes sense because MNB takes into account the number of occurrences of words, and SVMs do not assume independence of features. The low accuracy in text classification can be attributed to the fact that few speeches mention gun rights specifically. Multinomial regressions on politician features alone are over 95% accurate, even excluding speech data, because party is a high predictor of NRA rating (only using party and state as features to avoid overfitting also yields over 90% accuracy). Using speech data we can predict NRA ratings with 60.2% accuracy for up-and-coming politicians who have made speeches. For example, based on our MNB model, we predict Doug Jones, an unrated candidate for Senator of Alabama, to have rating C/D from the NRA [7]. Results such as this can be extrapolated to intuit how much campaign funding future politicians might receive from the NRA.

Future Work

To improve our performance we plan on using word embeddings to featurize the speeches, which will our models to incorporate some sentiment analysis. We will also see if categorizing ratings differently (for example, making the bins A/B, C/D, and F) gives us more predictive power.

References

[1] A nation divided: classifying presidential speeches. http://cs229.stanford.edu/proj2016/report/AcharyaCrawfordMaduabum-ClassifyingPresidentialSpeeches-report.pdf. Accessed:2017-11-21
[2] Where congress stands on guns: https://projects.propublica.org/guns/#nra Accessed:2017-12-09
[3] Discover u.s. Government information. www.govinfo.gov. Accessed:2017-11-21
[4] Gaussian Naive Bayes.http://scikit-learn.org/stable/modules/naive_bayes.html#gaussian-naive-bayes Accessed:2017-12-09
[5] Support vector machine https://goo.gl/images/yrPjWv Accessed:2017-12-09
[6] CS 229 Lecture Notes http://cs229.stanford.edu/syllabus.html Accessed:2017-12-09
[7] Doug Jones, US senate. https://dougjonesforsenate.com/priorities/ Accessed:2017-12-09