

## Abstract

This project aims to demonstrate an approach for seeking conflicts of interest in political speech—that is, instances where politicians voice opinions (e.g. via Twitter, speech on the congressional floor, or simply voting on a bill) that are related to industries from which they have received donations. This involves training models to recognize text as related to a particular industry, then running those models on political speech, and highlighting instances of where politicians speak about industries that have funded them beyond a certain threshold. The primary goal of this project in this semester is to investigate whether it is possible to train these models to be effective, and to document the methodology of this investigation.

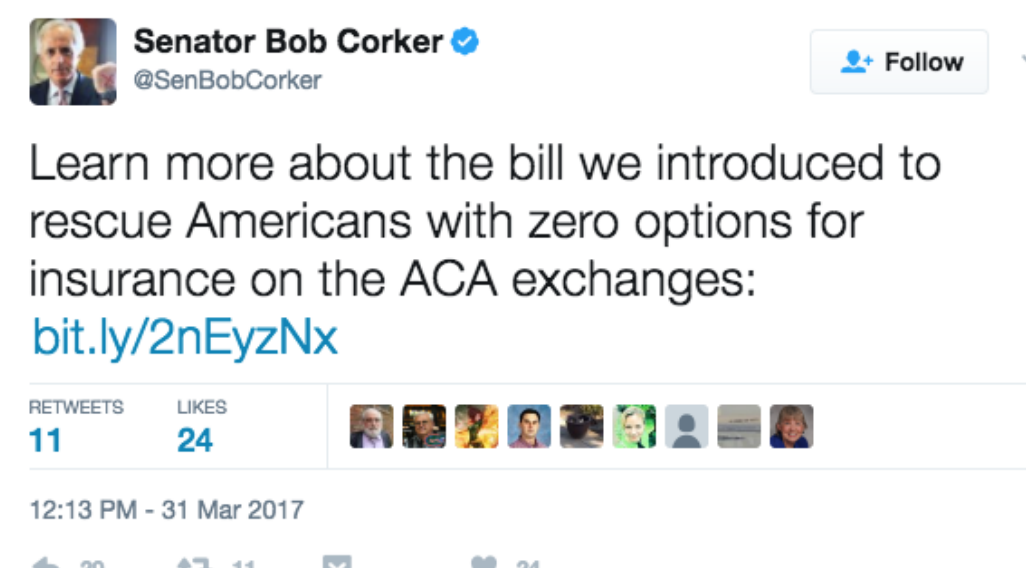
## Data Collected

- Partial tweet history for 526 current members of the 115<sup>th</sup> U.S. Congress
- Top ten industries donating to each member of congress, as well as totals by sector
- Company descriptions for ~3,400 publicly traded companies labeled with OpenSecrets industry classifications

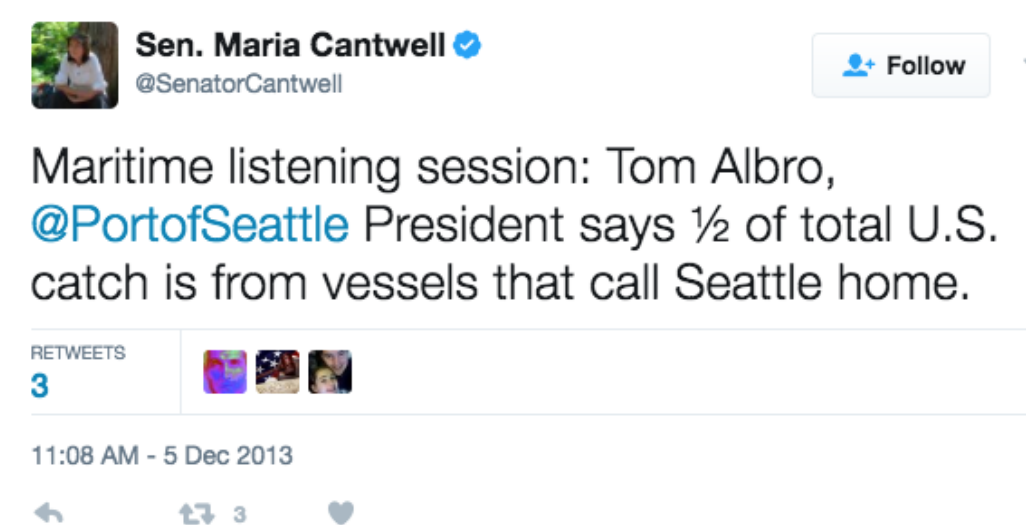
## Data Sources



## Domain Adaptation Results



**Figure 1.** Tweet labeled industry “Insurance” with probability 1.0. “Insurance” one of top ten industry donors: potential conflict of interest. Model used: bagging.



**Figure 2.** Tweet labeled industry “Sea Transport” with probability 0.8. The “Sea Transport” Label has only 39 samples. Note: not a suspected conflict of interest. Model used: bagging.

## Modeling

Approach:

- Build models that correctly classify company descriptions to OpenSecrets industry and sector labels
- Apply trained model to corpus of political tweets (domain adaptation)

Challenges:

- Unevenly distribution of labels
- Tweets are often totally unrelated to any industry, while training data is all by definition related to an industry

Pipeline:

- Tokenize text and count tokens, removing stopwords (bigrams and lemmatization not useful)
- Term frequency-inverse document frequency (tf-idf) (except for Multinomial Naïve Bayes)
- Classifying model (tested: Multinomial Naïve Bayes, Bagging, Support Vector Machine, Logistic Regression)

## In-Domain Results

Task	MNB	Bagging	SVM	Logistic
Sector	0.852	0.810	0.896	0.904
Industry	0.724	0.712	0.839	0.844s
isPharma	0.971	0.962	0.977	0.981

**Table 1.** Classification rates for several models on several different tasks.

## Tools



## Discussion

- Unsurprising that ngrams not useful: ordering of text less important than vocabulary
- Difficulty of classification:
  - Industry > Sector > Binary
- Logistic regression performs best within domain, but does not adapt as well to tweets as Bagging Classifier

## Conclusion

This project has demonstrated the feasibility of collecting data and training basic models that can help identify conflicts of interest. Though not as much modeling progress was made as would have been ideal, groundwork is laid for a tool that can help enable greater public participation in the political process by informing citizens of conflicts of interest in a context-dependent manner.

## Acknowledgements

I'd like to thank Professor Dragomir Radev for his support and insights while advising this project.