Yale

Text Mining to Find Interactions between Political Contributions and Political Speech

Lincoln Swaine-Moore (advised by Dragomir Radev)[1]

[1]Computer Science, Yale College, New Haven, CT

Motivation

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

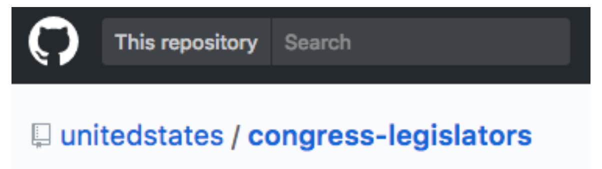
- Partial tweet history for 526 current members of the 115th U.S. Congress
- Top ten industries donating to each member of congress, as well as totals by sector (from OpenSecrets.org)
- Descriptions for ~3,400 publicly traded companies labeled with OpenSecrets industry classifications

Primary Data Sources

REUTERS

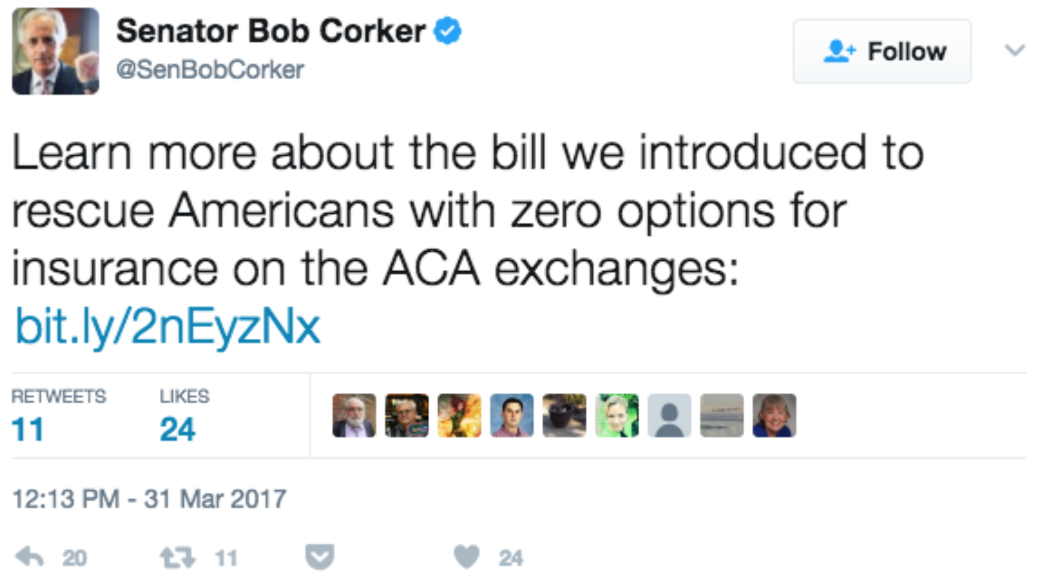
Nasdaq

OpenSecrets.org
Center for Responsive Politics


This repository Search

unitedstates / congress-legislators

Domain Adaptation Results

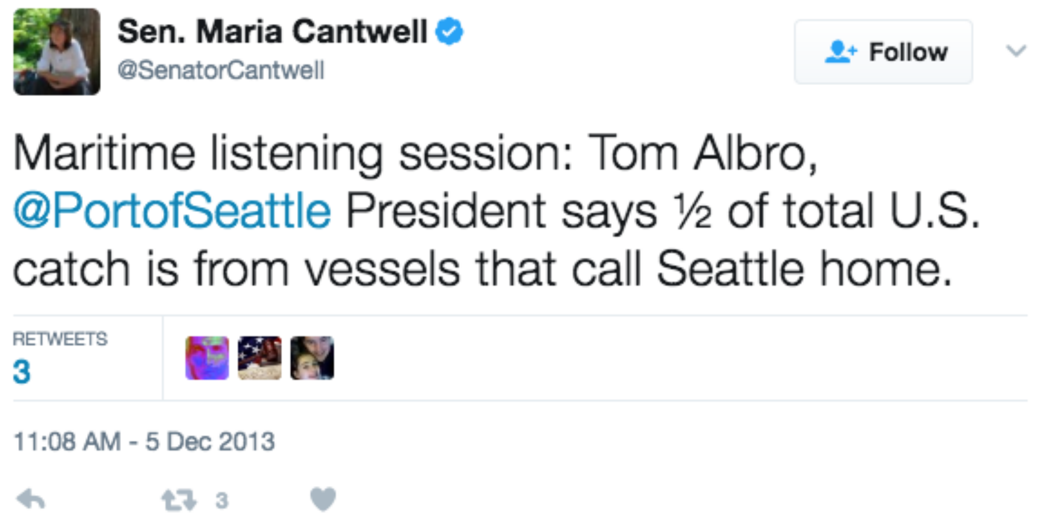

Senator Bob Corker
@SenBobCorker

Learn more about the bill we introduced to rescue Americans with zero options for insurance on the ACA exchanges: bit.ly/2nEyzNx

RETWEETS 11 LIKES 24

12:13 PM - 31 Mar 2017

Figure 1. Tweet labeled industry "Insurance" with probability 1.0. "Insurance" one of top ten industry donors: potential conflict of interest. Model used: bagging with decision trees..


Sen. Maria Cantwell
@SenatorCantwell

Maritime listening session: Tom Albro, @PortofSeattle President says ½ of total U.S. catch is from vessels that call Seattle home.

RETWEETS 3

11:08 AM - 5 Dec 2013

Figure 2. Tweet labeled industry "Sea Transport" with probability 1.0. The "Sea Transport" Label has only 39 samples. Note: not a suspected conflict of interest. Model used: bagging with decision trees.

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 with Decision Trees, Support Vector Machine, Logistic Regression)

In-Domain Results

Task	*MNB*	*BDT*	*SVM*	*Logistic*
Sector	0.861	0.813	0.893	0.892
Industry	0.708	0.703	0.814	0.795
isPharma	0.985	0.961	0.984	0.980

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

Tools

jupyter

scikit learn

Natural Language Toolkit (NLTK)

Discussion

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

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. See report for full acknowledgements.