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.

Primary Data Sources

- 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

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)

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.