Interaction between Political Contributions and Political Speech

Introduction

Politicians frequently accept donations (directly or indirectly) from individuals or entities associated with particular industries. This allows donors access (again, directly or indirectly) to politicians’ ears, and potentially gives priority to their agendas. Organizations like OpenSecrets have collected information about these donations. Unfortunately, though, many individuals remain unaware of the ways in which donations affect the policies and speech of their representatives. Applications such as Greenhouse attempt to make this information more accessible while browsing online articles and the like. I think there is room, however, for an approach that provides this information in an even more context-dependent way.

Proposal

I propose studying whether it is possible to use natural language processing (NLP) techniques to identify political speech that potentially evinces conflicts of interest due to industry donations.

Accomplishing this involves constructing models that classify text as being related to any number of given industries—essentially N models for N industries, each of which produces a probability that the text addresses topics that concern that industry. Then, if an instance of political speech scores high for a particular industry, and that politician has received money from that industry above a certain threshold, there is evidence of a potential conflict of interest.

1 https://www.opensecrets.org/industries/methodology.php
2 https://allaregreen.us/
This project would be far from the first to study conflicts of interest, but to my knowledge, would be novel in its automation of the process of discovering them.

**Data Sources**

Information about the donations can come from the OpenSecrets API[^3], which I believe is what the aforementioned Greenhouse extension uses.

Speech from politicians can come from either individual tweets (Congressional accounts should be available through “the @unitedstates project”[^4], and the Twitter Search API makes it relatively easy to gather tweets[^5]) or from the Congressional Record (available through Propublica[^6]). Political speech relating to an industry could also take the form of cosponsoring (or perhaps even just voting on) a bill that concerns an industry. Some information on bills cosponsorship and category is available through the U.S. Government Publishing Office[^8] and Propublica.[^9] I probably will focus on either one or two of these sources of data, since different forms of “speech” presumably require different models to classify accurately.

**Modeling**

There are multiple approaches that could be used to classify speech. One approach would be find companies associated with particular U.S. “Standard Industrial Classification code[s]”[^10] and build a dictionary for each relevant industry based on words that are frequently mentioned near or in the same documents on the web as those companies are. This dictionary could then be

[^3]: https://www.opensecrets.org/resources/create/apis.php
[^4]: https://theunitedstates.io/
[^5]: https://dev.twitter.com/rest/public/search
[^6]: https://projects.propublica.org/represent/
[^7]: https://github.com/propublica/Capitol-Words
[^8]: https://www.gpo.gov/fdsys/bulkdata
[^9]: https://projects.propublica.org/represent/
used to produce probabilities that political speech concerns a particular industry. It might also be possible to less explicitly construct a dictionary; that is, the documents concerning the industry companies could be used as training data for supervised model with the industries as labels. Because the training data would be from a different domain (articles about companies, primarily) than the testing data (political speech), this would be an example of transfer learning. One major advantage of these two approaches is that they do not require manually labeling training data.

A second approach would be to gather and label instances of political speech that is known to relate to particular industries, and then use a supervised approach to train a model to classify the text based on the ground truth labels. The primary failing of this technique is that it would require a large amount of labeling to be done. If I chose to go this route, I could do part of the labeling myself, and perhaps also employ Amazon Mechanical Turk.¹¹

Both of these approaches require validation on a test data set, to ensure that the models learned are not spurious. Thus, some labeled data will be required. However, much less labeled data should be necessary for a test set than for a training set, so labeling can be accomplished by hand—again, with support from Mechanical Turk if necessary.

**Deliverables**

The primary deliverable of this project would be a paper describing the dataset collection, methods, and results. This would presumably also serve as the “final project report” described on the CPSC 490 website, and would be of a format such that it could be submitted to the Association for Computational Linguistics (ACL) next year.

All code used for the project will be made available on Github.

¹¹ https://www.mturk.com/mturk/welcome
I think this topic also lends itself naturally to an actual prototype of a tool. Such a tool could take one of several forms. First, if work was done on Twitter speech by politicians, the tool could be a Twitter bot that assesses new tweets by politicians for potential disclosure of conflicts of interest, and replies to those tweets with a message notifying users. Another form the tool could take would be a website that displays conflicts of interest that have come up in a live way (that is, new bills, congressional speeches, tweets, etc.) that illustrate disclosure conflicts of interest. A final form would a (free) subscription service; users could sign up to “follow” particular politicians, and would be notified when potential conflicts of interest occurred. The creation of one of these tools is not guaranteed during the course of this project—time constraints, or lack of success with the modeling component may make it infeasible.

Future Work

At the conclusion of this semester, there will likely be either more successful classification rates that could be achieved, more domains of speech to which the techniques could be applied, or a tool that could be made (as described above) for publishing the modeling results (or improvements on such a tool, if part is implemented during the semester). Future work could proceed in any of these directions.

Relevant Academic Work


