Sentiment Analysis of Political Statements Using Cross-Field Training Corpora

Krissy Gianforte, UC Berkeley MIDS
W266, Fall 2018

Abstract

*Political language can be difficult to understand. Opinions are masked behind idioms, and campaign editors carefully review all communications for potentially contentious phrases; even when politicians **do** express strong sentiments, criticisms are often stated as outright facts. It is a real challenge to understand a politician's true sentiments and stance on a particular issue. Natural Language Processing could ideally help with such deciphering, but currently no large corpus of political statement texts exists to train sophisticated sentiment models. (Most previous work focuses on Twitter, which has a distinctly different communication style than formal statements.) This work explores three corpora from other fields– VADER (social media), Opinion Lexicon (product reviews), and SentiWordNet (WordNet vocabulary) – and explores whether any may be a suitable training set for analyzing political statements. Ultimately, this work proves that politics has a language all its own, and a specific corpus of political statement data will be necessary if future works wish to develop accurate models for comprehending the language of government.*

The project code and supplemental files are available as a Github repository at https://github.com/KrissyDG/PoliticalSentimentAnalysis.

I. Introduction

The language of politics can be difficult to understand. Politicians' statements are often convoluted, full of claims that are somehow firm yet noncommittal. Given such evasiveness, it can be extremely difficult for citizens to understand the positioning of their own government representatives; yet citizens do have a real interest in knowing how their representatives are likely to vote on the issues. This project applies machine learning and language processing to that problem, in order to decode political language and anticipate representatives' voting based solely on their self-professed positions on the given issues.

"Stance analysis" has been attempted before by many research groups, however previous projects have almost universally focused on detecting stance in Tweets. This sort of work is discussed further in Section II, but in general it is facilitated by the magnitude of data available on Twitter and the expressiveness of Twitter as a medium. This project, in contrast, focuses on the issue statements published on congressmen's official campaign websites, which use more formal language[1]. This focus on formally released statements introduces constraints, largely in the amount of data available. Many classic high-performing models (ex. neural networks) are out-of-reach given the small amount of training data.

Within that constraint, the work presented here focuses on preprocessing the political-language texts, then implements models trained on three corpora from other (non-political) fields and evaluates their performance. Ultimately, this work answers the question: given that labeled, formal, political-language data isn't available in large quantities, is there an adequate training dataset available from another field? Or is the language of politics nuanced enough to demand its own, specific corpus?

[1] Throughout this paper, members of the US House of Representatives will be referred to generally as "congressmen" rather than "congresspeople". While there *are* female members of the house, the traditional term "congressman" is used here for ease of reading.

II. Related Work

Many research teams have attempted to extract political stance from politicians' words, though those projects focus almost exclusively on tweets for a few reasons. First, Twitter is an ideal source for collecting large multitudes of data. For example, a recent paper by Johnson et al collected 99,161 tweets by monitoring the accounts of just 32 politicians. Second, the expressive communication style of social media such as twitter makes sentiment analysis much easier. Delivery patterns such as capitalization, repeated punctuation, hashtags, and emoji make the author's sentiment much clearer than it would be in other written formats.

2016 work by Mohammad et al [1] was able to detecting stance and sentiment in tweets concerning political topics: atheism, climate change, feminism, Hilary Clinton, and legalization of abortion. Their stance model considered features such as n-grams, sentiment, and encodings to understand the author's position on the target issue. Interestingly, the sentiment segment of their model used preexisting lexicons, rather than training on the set of politically-oriented tweets. *This implementation was a large inspiration for this paper's attempt to use cross-field lexicons for sentiment detection in political texts.* A total of five lexicons were used to determine sentiment, though many (NRC Hashtag Sentiment and NRC Emoticon Sentiment, for example) were focused on social media texts. Only one of the five lexicons seemed generic enough to be explored in this project: the Hu and Liu Opinion Lexicon. Note that the "encodings" feature of the Mohammad et al model is also strongly tailored for social media, as it searches for "hashtags, characters in upper case, elongated words (ex. *sweeeeet*), and punctuations such as exclamation and question marks."

Another 2016 work by Johnson et al [2] attempted to predict politicians' stances using Twitter activity over time. Their model considered the entire body of a politician's tweets – and, interestingly, times when the politician did *not* tweet about a trending topic – to understand stance. Keyword filtering was used to sort a dataset of 99,161 tweets into 16 topic areas such as ACA, immigration, and social security. *This filtering technique was incorporated into this paper's work and proved useful for refining generic Healthcare statements towards the specific issue of Obamacare.* The Johnson project used OpinionFinder2.0 to detect the sentiment of tweets, labeling each as positive, negative, or neutral [3]. The OpinionFinder tool filters texts for subjectivity in order to reveal "internal mental or emotional states, including speculations, beliefs, emotions, evaluations, goals, and judgments", then attempts to detect sentiment in these more emotional statements. *The subjectivity-filtering concept was incorporated into this project, as discussed in Section V*.

Both of the works discussed above successfully detected sentiment and stance in politicians' tweets. However, more formal political issue statements present an entirely new challenge, given the less expressive language and formatting, and a much smaller data set of texts.

III. Specific Application

a. Issue: Healthcare

This project focuses on the conflict surrounding the Patient Protection and Affordable Care Act (PPACA, or "Obamacare"). Healthcare is a very polarizing issue; Democrats are strongly in favor and proud of passing the ACA, whereas Republicans are intensely critical and adamant that the act should be repealed. The Obamacare debate provides a uniquely perfect opportunity for this project. Most congressmen have published statements on their campaign websites declaring their position regarding Obamacare, so free-text statement data is available. Recent voting on a bill to repeal Obamacare provides ideal "label" data, as it should reflect each congressman's actual position on the issue; the recency of the vote makes it even more appropriate to pair with current statement data.

b. Data Collection

Votes:

In May of 2017, the United States House of Representatives voted on the American Healthcare Act of 2017 (HR 1628) [4]. This bill was the first attempt to repeal and replace Obamacare, spearheaded largely by the Republican segment of the US Congress [5]. 410 sitting members of the House voted on the bill, ultimately passing it by a vote of 217 to 213. The voting was largely split along party lines, though some number of Republicans did vote *against* the bill (that is, they voted to maintain Obamacare).

The voting record for HR 1628 is available publicly through the US House of Representatives Clerk's Office [6]. The voting results table was extracted and imported into a data table and associated with congressmen data by name and (in the case of duplicate name) state.

Statements:

Each United States Congressman publishes a .gov website which contains fairly standard information: a brief biography, voting history, issue statements, contact information, and any other public relations materials the congressman wishes to release. The United States Library of Congress provides a convenient index page for locating the .gov website of each representative [7]; in this case, the index was filtered for the 115th congress (2017-2018) to select the members sitting in office when HR 1628 passed through the house. This index page also provides basic information on each congressman, including first and last name, state of representation, and political party. These fields were transferred into the data table for reference, should the need arise.

The .gov website of each congressman was then visited and the "Issues" page was manually located. Any available statement on Healthcare was copied into the data table exactly as published on the website. In some cases, no statement was available: most commonly, these websites simply said "For more information concerning work and views related to health care, please contact our office", rather than offering a specific statement. In such a case (or in any other case where a statement could not be found), the *statement* field was left blank in the data table and these congressmen were ultimately left out of the study.

IV. Approach

The data used for this project was strictly limited to Healthcare statements given by the 410 congressmen who voted on HR 1628. As explained in previous work by Mohammad et al, "the words and concepts used … are not expected to generalize across the targets [topics]", so statements concerning other issues (such as education, immigration, etc) would not be appropriate supplements to expand the dataset. Even for a single issue, though, extracting statements from congressmen's websites proved time consuming given the manual data collection methods discussed above. In contrast to Twitter, which has a friendly api and is easily combed, government websites are arbitrarily designed and perhaps even clumsy; each statement had to be manually located and copied.

Given those two hinderances, the dataset of relevant political stance texts was severely limited in size. Therefore, the goal of this project (as explained in Section I) was to determine whether corpora from other (non-political) fields could provide adequate training data, rather than having to divide the collected data into training and test sets.

Three corpora from different fields were chosen from the Natural Language Toolkit library [8]:

The **VADER Sentiment Lexicon** evaluates a full text as positive or negative, and returns an intensity measure of the sentiment. VADER was designed specifically for social media text; it accounts for capitalization and punctuation (which intensify the words they accompany) and can interpret common emoji [9].

The **Opinion Lexicon** dataset (Liu and Hu) contains a list of 6800 positive and negative words; the scoring system searches for each word in a passage in the positive or negative list, then returns the wordcount difference (ie. number of positive words minus number of negative words). This lexicon is based on customer product reviews, so it contains the sort of language used on review sites such as Amazon [10].

SentiWordNet provides a sentiment score for each word and sense in the WordNet synsets; scores may be positive or negative values. A text is scored using SentiWordNet by summing the sentiment scores for each word in the passage [11].

This project used each of the corpora in turn to conduct sentiment analysis on the congressmen's statements, then used the resulting sentiment values to predict how each congressman voted on the 2017 bill to repeal Obamacare.

Note that the models use *only* the statement text to make the vote prediction. Political party is surely the strongest predictor of voting on this sort of polarized issue [12], but the goal of this project is not simply to develop an accurate vote-prediction model. The desire is to explore the nuances of political language.

V. Methods

a. Initial Data Cleansing

The initial data table contained 430 rows: one for each congressman who voted on HR 1628. As noted above, however, statements were not available for every representative. Removing the rows that did not contain statement text reduced the size of the data set to 296 rows, each containing the following information:

- Representative's name and state
- Political party (Democrat or Republican)
- Statement text
- Vote (1/YES on 1628/*against* Obamacare; 0/NO on 1628/*for* Obamacare)
- isRep (convenient Boolean field for whether the congressman is Republican)
- crossVote (convenient Boolean field for votes across party lines, ie. a Republican who voted against the bill)

The 135 congressmen who did not include Healthcare statements on their website represented 31.4% of all voting members; a disproportionate 61.9% of cross-voters did not publish statements.

b. Preprocessing Statements

A large amount of careful preprocessing was used to prepare the statement texts for sentiment analysis. First, the text was reformatted to be entirely lowercase (for more universal phrase matching). Both line breaks and punctuation were interpreted as sentence separators; this helped differentiate section headings from the sentences that followed.

Next, the statements were filtered for relevant sentences according to the model set by Johnson et al [2]. In that work, the team filtered tweets for relevance using a list of approximately 7 keywords; they advocate for keeping the list short in order to avoid overselection "(ex. avoiding tweets about praying for a friend's health but keeping tweets discussing health care)". This project used the following list of four keywords to filter statements: ["obamacare", "obama care", "affordable care", "aca"]. Some experimentation was done including the terms "insurance", "coverage", "health plan", and "repeal", however such expansion pulled in

sentences related to Medicare and other issues unrelated to HR 1628, so ultimately the more refined list was used for filtering.

A small number of statements did not contain any of the four keywords; those rows were removed from the dataset after filtering. The final data frame fed into the modeling phase contained 250 rows.

c. Sentiment Models

Three sentiment analysis models were developed, each trained using one of the lexicons discussed in Section IV. All modeling employed the simple bag-of-words structure, to facilitate rapid and diverse exploration of different training options. (Of course, future work should apply more advanced modeling techniques once an appropriate training dataset is solidified.)

The VADER model used the NLTK Sentiment Intensity Analyzer module to generate a pair of positive and negative sentiment scores for each text. The Opinion Lexicon (Liu/Hu) model totaled the number of positive and negative words in each statement and returned the net positive count (such that negative statements returned a negative LH_sentiment value). Finally, the model based on SentiWordNet lemmatized each word in the statement, then totaled the positive and negative sentiment values for all words to create an overall positive/negative score pair.

The model predicting vote from sentiment used intuitive dividing lines based on the training lexicon, rather than trained division lines based on the statement/vote dataset. This was a necessary approach, given the restriction to use only freely available lexicons and *not* statement data for all training tasks (recall Section IV). Simply put, votes were predicted based on the zero-sentiment line; positive-sentiment statements were corelated with a 0 vote (that is, against repeal and *for* Obamacare) and negative-sentiment statements with a 1 vote (that is, for repeal and *against* Obamacare).

d. Subjectivity Analysis

In an attempt to additionally refine the statement text being fed into the model, subjectivity analysis was applied to extract only subjective sentences. This concept was based on OpinionFinder2.0, which filters texts for subjectivity in order to reveal "internal mental or emotional states, including speculations, beliefs, emotions, evaluations, goals, and judgments" [3]. It was hoped that these subjective sentences may contain stronger sentimental language, thereby assisting the sentiment analysis models.

Texts from the NLTK Subjectivity Corpus [13] were used to train a Naïve Bayes model to identify subjective versus objective sentences. Each congressmen's statement was then re-filtered, with each sentence fed into the subjectivity model and only the subjective sentences retained. The re-filtered statements were fed into the same sentiment models discussed above to once again predict voting.

> *i promised the people of the 36th district of texas that i would repeal obamacare and replace it with a workable, affordable solution which lowers premiums while protecting those with preexisting conditions. while republicans alone provided the votes to pass this bill, it was still the product of finding common ground to provide the american people with relief from obamacare. unfortunately, the senate defeated this bill by a single vote, and the fight to repeal obamacare continues*

Subjectivity Analysis Example. The subjectivity model evaluates each sentence in the statement text and extracts only the subjective sentences for further modeling. Here, sentences labeled as 'subjective' are shown in blue.

VI. Results and Analysis

a. Baseline Models

As expected, party affiliation is an intensely strong predictor of voting pattern. The model *Vote = isRep* (that is, predicting that all Republicans vote YES and Democrats vote NO) results in 97.6% accuracy. However, this baseline model is not based on the statement texts at all, so it is perhaps not a fair comparator for the natural language modeling in this project. A more appropriate baseline model for comparison is to predict that every congressman voted with the majority. With 129 YES votes in the final dataset and 121 NO votes, the accuracy of such a model would be 51.6%.

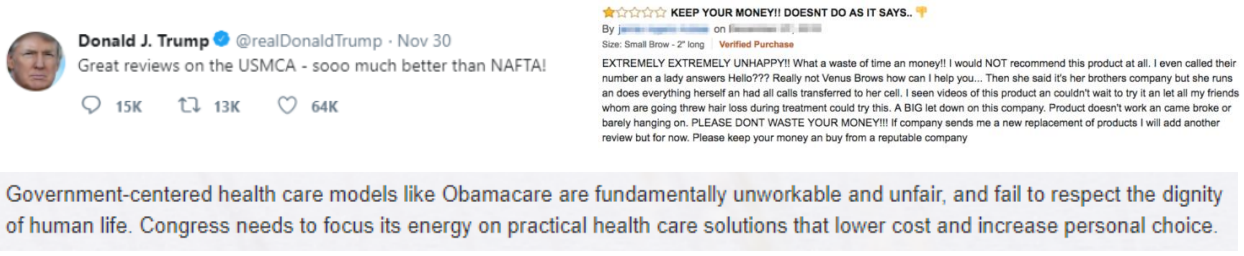
b. Sentiment Models – Accuracy Metrics

In general, the Sentiment Analysis models were barely able to outperform even the more forgiving baseline model. Confusion tables and plots for each model are given below; Section c discusses error patterns and likely reasons for the model inaccuracies.

c. Error Patterns and Issues

All three sentiment models dramatically overestimated the positivity in congressmen's statements, sensing positivity in nearly all statements. When negative sentiment *was* detected (and the model predicted a YES vote, against Obamacare), the congressman in question did in fact vote YES over 90% of the time. However, most often the model predicted a NO vote (that is, positive sentiment regarding Obamacare) for both congressmen on both sides of the issue.

It seems that political language hides negativity behind formality. Often, accusations and insults are masked by false claims given as facts, rather than being outright expressed. Office editors clean all statements for any typeface, punctuation, or language issues before release, so strong sentiment is not expressed through dramatic delivery. This is in strong contrast to the social media and product review texts that the training lexicons are based on, where capitalization, punctuation, and emoji are used to express feelings. It makes sense, then, that models searching for such dramatic formatting and language would interpret composed, "politically correct" language as positive.



Expressiveness of Political Statement Texts Versus other Fields. Twitter (a) and review sites (b) tend to contain much more expressive texts, incorporating capitalization, punctuation, and even emoji to convey strong sentiments. In contrast, political statements (c) hide strong emotion behind formal, seemingly-factual language.

Similarly, the subjectivity detection model struggled to accurately extract subjective sentences from the congressmen's issue statements because of the way opinions are presented by politicians. The subjectivity corpus used to train the classifier was built on movie review texts, so the model looked for subjective words typical in movie discussions: "charm", "engaging", "performances", and "interesting" are among the top subjective features. Politicians do not use such overtly judgmental language (even when making subjective claims) which makes it difficult for the subjectivity model to accurately label opinions as such.

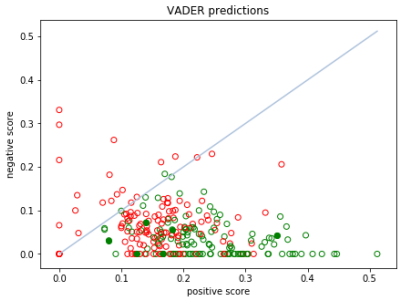
	Baseline Majority Class	VADER	Opinion Lexicon (Liu-Hu)	SentiWordNet
Voting Data	51.6%	---	---	---
Full Topic Statements	---	54.8%	57.6%	58.4%
Subjective Statements	---	55.3%	58.5%	61.3%

Legend:

- YES vote / anti-Obamacare
- NO vote / pro-Obamacare
- Cross-party NO vote (ie. Republican voting NO)

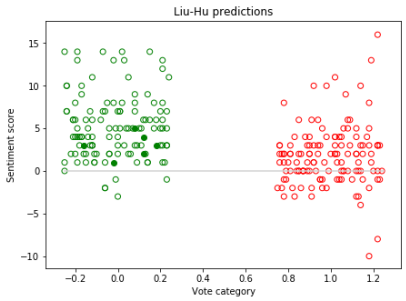
Full Topic Statements

VADER



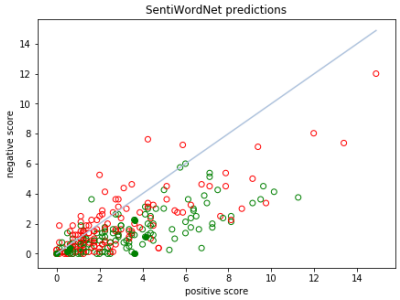
	Predicted Vote	
Actual Vote	**NO**	**YES**
NO	120	1
YES	112	17

Opinion Lexicon



	Predicted Vote	
Actual Vote	**NO**	**YES**
NO	116	5
YES	101	28

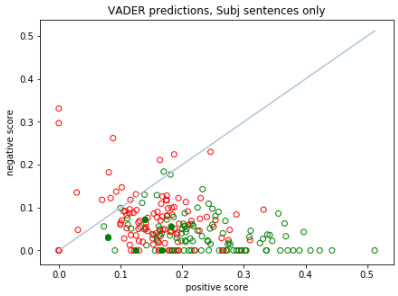
SentiWordNet



	Predicted Vote	
Actual Vote	**NO**	**YES**
NO	113	8
YES	96	33

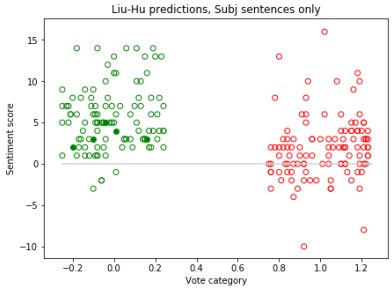
Subjective Statements

VADER



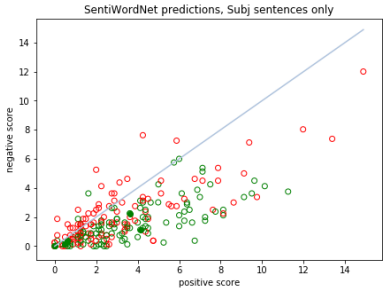
	Predicted Vote	
Actual Vote	**NO**	**YES**
NO	106	1
YES	96	14

Opinion Lexicon



	Predicted Vote	
Actual Vote	**NO**	**YES**
NO	103	4
YES	86	24

SentiWordNet



	Predicted Vote	
Actual Vote	**NO**	**YES**
NO	103	4
YES	80	30

Performance of Sentiment Models. The VADER, Opinion Lexicon, and SentiWordNet models barely outperformed the simple majority-classifier baseline model. Introducing subjectivity filtering did improve the model accuracies slightly, though not a significant amount. Confusion tables and plots show that all models dramatically overestimated the positivity in most statements.

… obamacare failed to accomplish real reform, and instead harmed health care, job creation, and the federal deficit at a time when our country could not afford such government inflicted damage. obamacare caused premiums to skyrocket, forcing millions of americans off of their current coverage and putting unelected washington bureaucrats between patients and their doctors. obamacare's equation is simple: higher costs, more debt, fewer doctors, reduced access, fewer jobs, and increased dependency on failed federal programs.

Subjectivity Analysis Issues. The subjectivity labeled only the blue sentences in the above text as 'subjective', though certainly the other sentences also express subjective opinions.

In summary, the known fact once again is proven: relevant, abundant training data is absolutely crucial for accurate modeling! Training on political language texts, rather than trying to pull from other genres, would likely dramatically improve sentiment analysis models in this space.

As a demonstration, consider a simple Naïve Bayes model that predicts votes based on the sentiment values determined in Section V.c. Without adjusting the text corpora or sentiment values, the model simply better-fits the sentiment feature values to the 0/1 voting outcome for each congressman by training on the Healthcare statement dataset. This helps adjust the strong bias towards positivity discussed above. Using 5-fold cross validation (to accommodate the lack of data), such a model is able to achieve 66% prediction accuracy: an improvement of nearly 10%!

VII. Final Conclusion and Implications

This project has demonstrated that natural language text corpora from other fields (such as social media and product reviews) simply do not generalize into the political arena. "Politically correct" language is vastly different from the free expression seen elsewhere online. Having a specific corpus of political language text will be vital to any future language analysis of candidates' issue statements. To enable such work, researchers should begin collecting statement/vote data on an ongoing basis.

Unfortunately, one likely unsolvable issue is that a congressman's statement sometimes simply does not correspond to his vote. For example, Congressman Biggs voted *against* HR 1628 (that is, to maintain Obamacare). However, his .gov website offers the following statement on Healthcare:

Obamacare has become a pariah for all who helped with its passage, and has led to unaffordable premiums and fewer options. For years, Republicans have promised to repeal Obamacare. With the House, Senate, and White House under Republican control, we now have an opportunity to fulfill this promise to the American people. Since taking office at the start of the 115th Congress, I have advocated for the immediate passage of a bill that completely repeals Obamacare and all the regulations associated with it. Many agree that this action would quickly lower premiums for every American… I am committed to the challenge of eliminating Obamacare and to re-establishing our nation's healthcare system to once again be the envy of the world. [14]

Even a human coder would not predict that Biggs would vote NO on a bill to repeal Obamacare, given how strongly his statement criticizes the program! Ultimately, politics is a strange beast and votes are cast for a multitude of reasons, not always related to ideology.

References

(All cited papers are available in the Reference Papers folder of the Github repository)

[1] S. M. P. S. a. S. K. Mohammad, "Stance and Sentiment in Tweets," ACM Transactions on Embedded Computing Systems, University of Ottawa, 2016.

[2] K. a. D. G. Johnson, ""All I know about politics I is what I read in Twitter": Weakly Supervised Models for Extracting Politicians' Stances From Twitter," *International Conference on Computational Linguistics,* vol. 26, pp. 2966-2977, 2016.

[3] University of Pittsburgh, "OpinionFinder 2.x Release Page," [Online]. Available: http://mpqa.cs.pitt.edu/opinionfinder/opinionfinder_2/. [Accessed 7 Dec 2018].

[4] United States Library of Congress, "H.R.1628 - American Healthcare Act of 2017," [Online]. Available: https://www.congress.gov/bill/115th-congress/house-bill/1628?r=357. [Accessed 7 Dec 2018].

[5] Henry J Kaiser Family Foundation, "Summary of the American Health Care Act," 2017.

[6] Office of the Clerk, US House of Representatives, "FINAL VOTE RESULTS FOR ROLL CALL 256," [Online]. Available: http://clerk.house.gov/evs/2017/roll256.xml. [Accessed 7 Dec 2018].

[7] United States Library of Congress, "Members of the U.S. Congress," [Online]. Available: https://www.congress.gov/members?q=%7B%22chamber%22%3A%22House%22%2C%22congress%22%3A%22115%22%7D. [Accessed 7 Dec 2018].

[8] "NLTK Corpora," [Online]. Available: http://www.nltk.org/nltk_data/. [Accessed 7 Dec 2018].

[9] C. a. E. G. Hutto, "VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text," AAAI, 2014.

[10] M. a. B. L. Hu, "Mining and Summarizing Customer Reviews," KDD, 2004.

[11] S. A. E. a. F. S. Baccianella, "SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining".

[12] J. M. J. a. T. G. Snyder, "Estimating Party Influence in Congressional Roll-Call Voting," *American Journal of Political Science,* vol. 44, no. 2, pp. 193-211, 2000.

[13] Cornell, "Movie Review Data," [Online]. Available: http://www.cs.cornell.edu/people/pabo/movie-review-data/. [Accessed 8 Dec 2018].

[14] A. Biggs, "Health," [Online]. Available: https://biggs.house.gov/issues/health. [Accessed 7 Dec 2018].