Mining sentiment and Twitter response to policy discussions

Jacqueline Gutman (jg3862@nyu.edu) and Alex Pine (akp258@nyu.edu)

3932 words

Introduction

The first public debate for the 2016 presidential campaign was held on August 6, 2016 for the Republican party's primary election. It featured ten candidates running for the Republication nomination--an unprecedentedly large number. Its viewership was the largest-ever for a political program ("24 million watched US presidential debate: Nielsen"). CrowdFlower, an web-based data-collection company, collected Twitter messages published during the debate and shortly afterwards, and hired individuals to label each one relevant to the debate with the subject matter, tone, and president candidate referenced. One would expect these labels to facilitate an understanding of how Twitter users' feel about the policy positions put forth by the candidates in the debate. Unfortunately, 62% of the Twitter messages labeled did not fall in one of the 11 subject-matter categories provided by Crowdflower, indicating the categories did not properly capture the breadth of matter discussed during the debate.

We attempt to discover a different set of policy categories that better encompass the scope of the debate so that we can better understand how Twitter users' reacted to the policy ideas given by the candidates. We did this by training a Latent Dirichlet Allocation (LDA) model on the Twitter text directly, as well as on the transcript of the debate itself. We then trained two logistic regression models on the Twitter text as represented by these two LDA models, and examined

their effectiveness in predicting the sentiment labels of the Twitter messages. We then interpreted the relationships uncovered by these logistic regressions as proxy measure of how Twitter users felt about particular policy-oriented topics, controlling for other differences between tweets on these topics.

Overall, our models were able to predict sentiment of a tweet nearly as well from the debate text LDA as from the Twitter text LDA or manually supplied labels. While the human-annotated labeled provided by crowdsourcing may be the gold standard, the LDA-based topics were nearly as effective in understanding the relationships between policy issues and public sentiment. Moreover, these topics do not require the expense of manual labeling. The topics learned from Twitter do not inform our understanding of key policy issues as well as the topics drawn from the debate. By extracting these topics and modeling the relationship between topic prevalence and Twitter reaction, we can better understand the public reaction to key issues raised in the debate and how the reactions differ by candidate.

Data

We collected two sources of data related to the August 6th debate: one of Twitter messages written during and shortly after the debate, and the transcript of the debate itself. The Twitter data comes from crowdflower.com, a website where one can provide an unlabeled dataset, and have it manually annotated through crowdsourcing. Raters were asked to review debate-related tweets written during the debate and the morning after. Workers rated whethre the tweet was relevant to the debate, and if so, what the subject matter was, which presidential candidate was

referenced, and if the sentiment of the tweet was positive, neutral, or negative about the candidate ("CrowdFlower: Who won the GOP debate?", 2015).

The available data includes only tweets labeled as “relevant” to the debate. No details are provided on how the unfiltered were extracted from the Twitter API, but they were likely filtered by related hashtags, such as “#GOPDebate”, and timestamp of the message. The debate transcript was provided by the Washington Post (Washington Post, "Annotated transcript: The Aug. 6 GOP debate", 2015). We processed the transcript by separating the text into “documents” consisting of each utterance by the moderators or candidates.

Theory and Hypotheses

Our primary goal is to understand how Twitter users felt about the policies discussed by the candidates in the Republican presidential debate. The labels in the Crowdflower Twitter dataset ostensibly provide this information, but they were insufficiently comprehensive. Only 62% of the tweets in the data set were assigned a subject-matter label related to a policy issue. Additionally, only 10 policy issues were presented as options to the people choosing the labels, which do not reflect the range of the topics that were discussed in the debate.

We investigate whether topics generated by an LDA model applied to the text of the tweets can more accurately reflect the policies mentioned in the debate and provide a better understanding of the topics discussed in the 38% of unlabeled tweets. If this topic-discovery technique is successful, each tweet could be placed in the vector-space of the new topics, allowing to model the relationships between these topics and the sentiment and candidate labels. This method

provides a rough way to gauge public sentiment about different candidates' positions concerning several policy topics of interest.

For example, imagine the LDA model indicates that a particular tweet is about the nuclear arms and US-Iran relations, its sentiment label is "negative", and its "candidate" label is "Cruz". If tweets concerning this topic are generally more negative for Cruz than expected given the aggregate sentiment towards Cruz and towards the nuclear Iran topic, then this information could reasonably be interpreted to mean that Twitter users did not agree with Ted Cruz's proposals on that issue.

Although analyzing the text of the tweets to build an LDA model is the most obvious way to discover topics in those tweets, we believe that topics drawn from the debate itself might actually generate a more useful model. A cursory glance at the Twitter text suggests that much of the content is not immediately about policy. More often, it is about the candidates themselves, or other Twitter users. An LDA generated on the Twitter data directly might therefore have more predictive value, but not help the researcher extract topics which speak to policy questions of interest.

The debate text, in contrast, is much more focused on policy directly, and therefore might do a better job of discovering the words that comprise a particular policy topic. For this reason, we trained an LDA model on the debate text in addition to the one trained on the tweets directly. Once trained, we used the topic model to understand their prevalence in the Twitter text by applying its posterior term-topic distribution on the tweets. Tweets that don't share terms from the debate-trained LDA would not be assigned a policy topic distribution.

Literature

In “Estimating Policy Preferences From Written and Spoken Words” (Benoit and Herzog, 2015), the authors discuss several statistical techniques used to infer policy preferences of individuals given documents they have written on the subject. Two of the three techniques discussed, Wordfish and word scores, place all documents on a binary spectrum, and require the researcher to choose anchor documents that represent its extremes. These approaches do not work well for this data. In the debate, all candidates are from the same political party, so placing their text on a binary political spectrum would not reasonably represent the nature of their differences. Only the third technique, LDA, is applicable to our data. It enables us to place any document into a space of multiple, not just binary, topics. LDA seems a natural choice to discover more policy topics within Twitter data.

The paper “Empirical Study of Topic Modeling in Twitter” (Hong and Davison, 2010) recommends techniques for maximizing the effectiveness of LDA of Twitter-based data. It’s primary recommendation is to aggregate tweets from the same user together into a single document. Unfortunately, this technique would diminish the primary advantage of our dataset: the sentiment and candidate labels for each tweet. Aggregating the text would require aggregating these labels as well, potentially losing a rich source of data. Applying this technique to the debate text topic model seemed more likely to be fruitful. Each candidate’s responses to moderator questions are typically short--not unlike a tweet--suggesting that aggregating their speech into single “documents” might improve the topic model.

Methods

Debate Text LDA for Sentiment Classification

The debate text provided a rich source of data for terms that clearly align to policy topics, so an LDA was trained on all exchanges from the debate, including moderator questions. We then applied its posterior term-topic distribution to infer a topic distribution for every tweet that shared at least one term in the debate LDA's vocabulary. These topic distributions were later provided as input features to a logistic regression classifying tweet sentiment.

Debate Text LDA Construction

We used Quanteda's list of SMART stop-words as a starting point for filtering uninformative terms, and additionally removed terms that were used in the transcript to describe speech unrelated to the content of the debate, e.g., APPLAUSE and BOOING. We filtered out terms not used in at least 3 debate exchanges (including moderator questions). The resulting model has a vocabulary of only 212 unique words.

We explored two approaches for building a latent topic model on the debate data: the Structural Topic Model (STM) (Roberts et al. 2016), and the standard LDA model (Blei et al., 2003). In both cases, we pre-processed the data similarly. We used the candidate associated with each debate excerpt as the content covariate in the STM model, and used the SearchK function (Roberts et al., 2016) to determine the number of topics (K) to set in our model. However, we found the results inconclusive: some values of K had better likelihood or residual values, but low semantic coherence, while other settings of K showed the reverse trend. As extracting

policy-oriented topics from the debate was critical to our approach, we ultimately chose the number of topics subjectively: by generating a set of candidate models for a range of K, and choosing the model that resulted in the most sensible and substantive topics.

We chose the number of topics to use for the standard LDA model in similar fashion. In comparing the STM and standard LDA models, we saw no advantage to the STM. The topics generated by STM often contained only a few words, and did not have a clear interpretation, while the topics chosen by standard LDA spanned the whole vocabulary, and were generally more coherent.

We experimented with aggregating all of the parts in the debate from the same speaker into a single document, resulting in one document per candidate, and an additional document for the moderators. We expected that longer documents might improve the relevance of the resulting topics, but the topics were not subjectively more coherent than the topics drawn from the unaggregated text.

The optimal LDA model for subjective semantic coherence was a standard (i.e. non-STM) Gibbs LDA with 15 topics, trained on debate exchanges where each candidate's answer to a moderator's question was treated as a single document. In examining the words belonging to each topic, three of them were related to debate moderation, and the remaining twelve pertained to these policy topics: "Common Core" education reform, foreign policy, Social Security reform, immigration reform, economics, Donald Trump's proposed border wall, Budgetary matters, Paul Ryan, the military, the general election against Hillary Clinton, the Iran nuclear deal, and gay marriage.

Training models with more than 15 topics resulted in essentially the same twelve topics, plus several incoherent ones. There are only two additional topics in this set compared to the ten policy labels that came with the CrowdFlower dataset. These twelve topics, however, are more specific than the ones that came with the Crowdflower dataset: abortion, foreign policy, gun control, healthcare, immigration, LGBT, race, religion, "women's issues (not abortion)".

Logistic Regression using Debate LDA Topics

We first performed a logistic regression to predict sentiment of a tweet given the posterior topic distribution over tweets of the 15-topic LDA fitted on the text of the debate itself. We were interested in whether topics learned from the candidate exchanges during the course of the debate could inform our understanding of the Twitter reaction to the debate. The original Twitter data contained 13,869 tweets, but after removing stopwords, Twitter related terms (such as hashtags, hyperlinks, RTs, and Twitter handles) and terms that appeared in fewer than 5 tweets in the entire corpus, there were 10,929 tweets remaining. Finally, all neutral sentiment tweets removed to transform the multiclass problem into a simple binary sentiment classification, leaving 8,722 tweets in the data, the overwhelming majority of which were classified as having negative sentiment.

The 15-topic LDA model was used to predict a distribution of topics over the Twitter data, with In addition to the predicted topic distribution, we used the annotated candidate labels of each tweet from the crowdsourcing data as additional features for the logistic regression. In order to fit a more robust and parsimonious model, we performed a stepwise regression procedure using

Akaike Information Criterion (AIC) as a model selection criterion to prune uninformative and redundant features from the scope of the model.

The resulting model retained the posterior topic distribution parameters for topics 3, 4, and 9, topics which roughly correspond to foreign policy, social security, and budget issues, respectively. Topics 3 and 4 were associated with reduced likelihood of positive sentiment under the selected model, meaning that tweets which included more terms related to the foreign policy and social security topics were more likely to be negative, controlling for the name of the candidate referenced in the tweet. Topic 9, related to budget and tax issues, was associated with increased likelihood of positive sentiment under the selected model, meaning that tweets which included more terms related to budget issues were more likely to be positive, controlling for the name of the candidate referenced in the tweet. A plot of the regression coefficients of the final debate topic model and sentiment classifier is shown below. In general, sentiment of a tweet was more strongly influenced by the name of the candidate it referenced than by the policy-oriented topic distribution learned from the text of the debate.

Final Model: **sentiment ~ topic.3 + topic.4 + topic.9 + candidate**

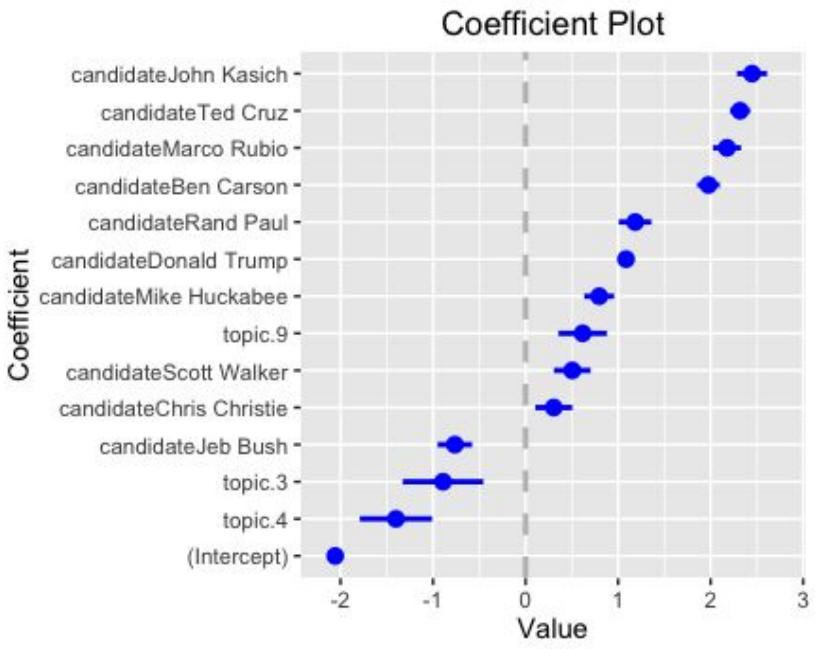

Figure 1: Logistic regression coefficients (log odds ratios) and confidence intervals for the logistic regression model chosen by stepwise and trained on the debate LDA topics

Logistic Regression using Twitter LDA Topics

Next, we trained an LDA topic model on the tweets directly. We examined LDA models between 10 and 50 topics, but found these topics to be incoherent regardless of the number of topics. So instead of selecting the number of topics to optimize subjective coherence, we chose the model which maximized the log likelihood of the observed tweets under the LDA, which was the model with 25 topics. We confirmed this choice of topic number by comparing the logistic regressions with all potential predictors included in the model, and found that under the classification task, the 25-topic model also outperformed the other models by minimizing the Bayesian Information Criterion (BIC).

As 25 topics was optimal for both supervised and unsupervised models, we retained a posterior topic-document distribution over these 25 topics for all tweets that remained after pre-processing. As with the topics learned from the debate data, we performed a stepwise regression procedure using AIC as a model selection criterion, and also included the name of the candidate referenced in the tweet as a potential predictor for the model.

In the final model selected by the stepwise procedure, we retained 15 of the original 25 topics, and of these remaining 15 topics, 12 were considered statistically significant predictors of sentiment (with an additional topic considered marginally significant). A plot of the regression coefficients of the final Twitter topic model and sentiment classifier is shown below. Compared to the regression using the debate topics, in this model, we see that sentiment of a tweet was much more strongly influenced by prevalence of certain topics than by the candidate referenced in the tweet. However, these topics were generally more sentiment and candidate-oriented, with very limited weight placed on policy-relevant terms in the term-topic distribution for this model.

Final Model: **sentiment ~ topic.1 + topic.2 + topic.3 + topic.4 + topic.6 + topic.10 + topic.13 + topic.14 + topic.15 + topic.17 + topic.19 + topic.20 + topic.22 + topic.23 + topic.24 + candidate**

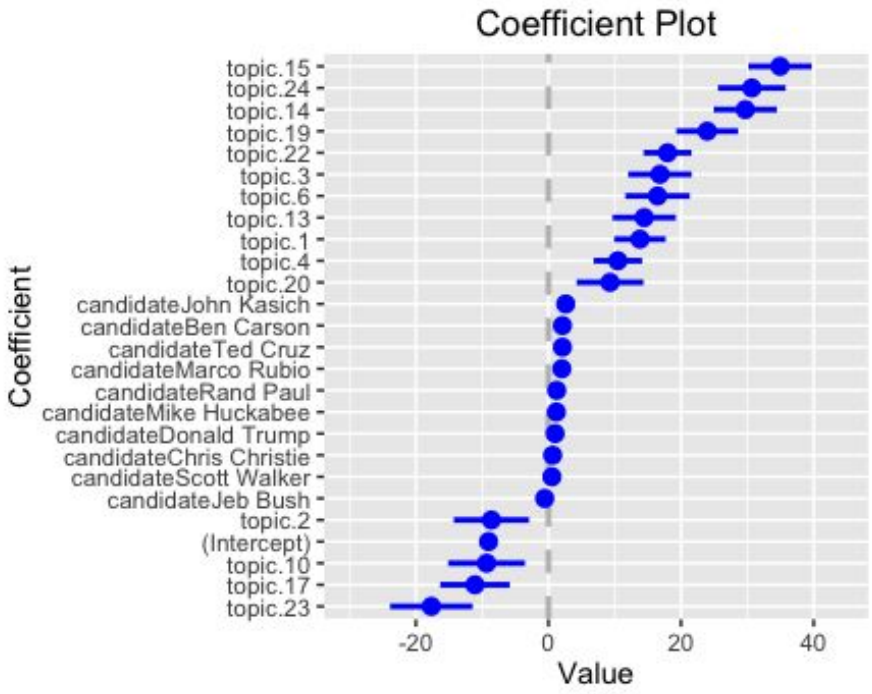

Figure 2: Logistic regression coefficients (log odds ratios) and confidence intervals for the logistic regression model chosen by stepwise and trained on the Twitter LDA topics

Debate topic classifier vs. Twitter topic classifier

After fitting the two sentiment classification models on the learned topics, we compared the two models to determine how much information about the sentiment of a tweet could be recovered by fitting the topic models on the more policy-oriented text of the debate instead of the more directly relevant text of the tweet. A likelihood ratio test of the deviance statistic between the two non-nested models revealed, as expected, that the sentiment classifier trained on the 25 fitted Twitter topics performed significantly better than the sentiment classifier trained on the 15 fitted debate topics. This significant difference in model performance does not necessarily translate to

a meaningful difference in performance, however. As the Twitter-trained topics do not relate to key policy issues discussed in the debate, using this model requires that we sacrifice the coherence and political relevance of the topics, and therefore our ability to interpret the regression coefficients as proxy measures for public sentiment towards the policy topics of interest discussed in the debate.

Results

We compare the model performance and find that the classifier trained on the topics drawn from the debate is a more than passable approximation of the classifier trained on the Twitter text directly. BIC is marginally lower for the Twitter-trained model--7804 for the Twitter topic model compared to 7866 for the debate topic model. The area under the receiver operating characteristic curve (AUC), is also close for both models--0.754 for the Twitter topic model compared to 0.725 for the debate topic model. The ROC curves for both models are shown together below. Thus, the debate topic model provides a reasonable approximation of the Twitter topic model while allowing us to more substantively gauge Twitter reaction to the key topics discussed in the course of the debate.

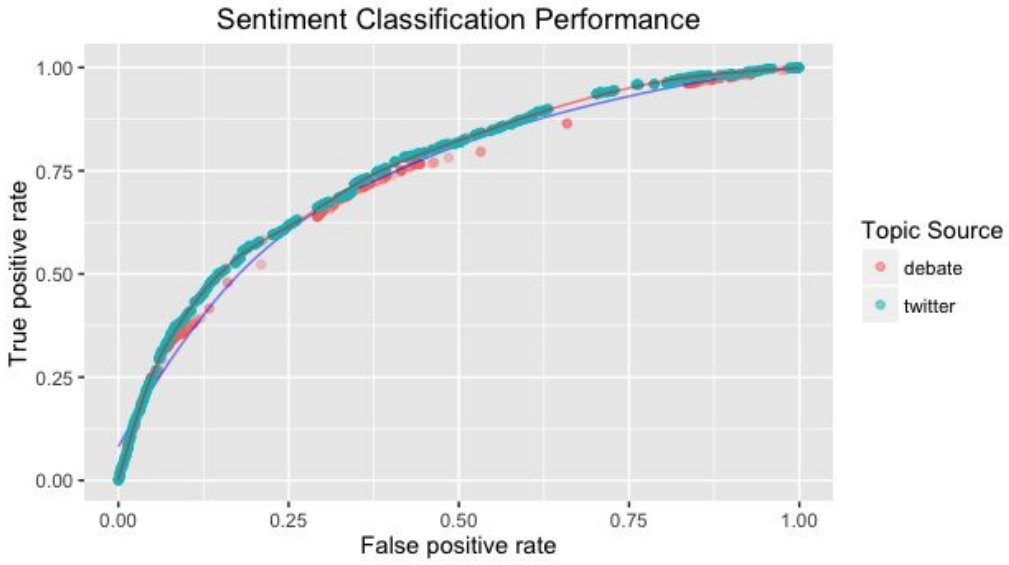

Figure 3: Comparison of sentiment classifier performance between models trained on the topics drawn from the debate text versus the Twitter text

Finally, to compare our latent topic models to the performance given by the manually annotated subject matter labels, we considered a third logistic regression using only these subject labels and the referenced candidate labels as predictors for the model. While most of these subject categories were statistically significant predictors of sentiment, the model did not substantially outperform the topic-model trained logistic regressions. The AUC for the fully supervised classifier was 0.749, in between the performance of the debate topic model (0.725) and the Twitter topic model (0.754). BIC was only slightly better than the less parsimonious Twitter topic model (7740 compared to 7804), and this advantage was mainly due to the smaller number of parameters in the fully supervised model.

Many of the specific policy-oriented topics learned from the debate topic model may have been more associated with positive or negative sentiment simply because these topics were strongly associated with a particular candidate, and the Twitter response to that candidate was largely

negative regardless of the particular policy topic in question. We are interested in how Twitter users responded to these policy issues, regardless of differences in their relative frequency within messages referring to each candidate.

We use the stepwise regression results as a measure of the importance of each topic to understanding Twitter sentiment above and beyond the overall sentiment of Twitter users towards particular candidates. Of the 15 topics learned from the debate, 3 were still important in predicting sentiment after controlling for these candidate-level effects. We find that Twitter users were generally more negative towards the topics concerning foreign policy and social security issues compared to their baseline opinion of the candidate. In particular, Twitter users were more critical of Jeb Bush, Ted Cruz, and Ben Carson on questions of social security, and more critical of Marco Rubio and Scott Walker on foreign policy issues.

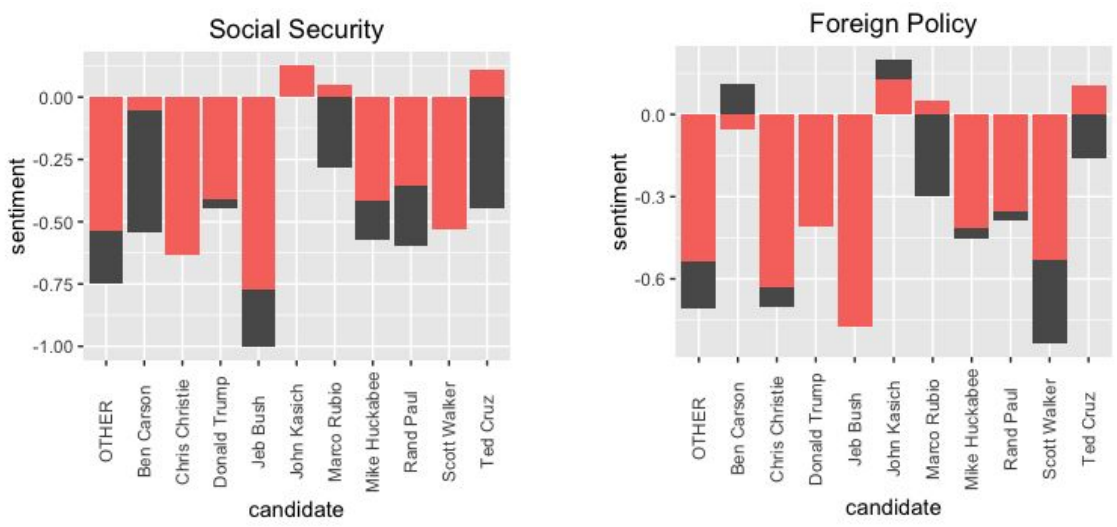

Figure 4: Overall (red) and topic-specific (gray) Twitter sentiment towards each candidate for the social security, and foreign policy topics drawn from the debate text

Some examples of tweets in this category help illustrate this increase in negative sentiment towards candidates for this topic. In the graph on the left, we see that while overall sentiment towards Ben Carson is fairly close to neutral, when it comes to tweets mentioning the social security topic, opinion is strongly negative. As an example, one tweet predicted by the LDA to be related to social security states: “Ben Carson: I will end polarization by removing health insurance from 17 million Americans. #Obamacare #GOPDebate”. In the graph on the right, we see similarly steep drops in sentiment towards Marco Rubio and Scott Walker on issues of foreign policy. A tweet predicted to be related to foreign policy states: “Yes Marco Rubio Actually Said The Iraq Invasion 'Was Not A Mistake’ ”. However, for Scott Walker, many of the tweets contributing to this perceived dip in public sentiment concerning foreign policy are in fact more closely related to questions about abortion or about unions, such as “How does plan on giving priority to working American families and wages if he is so anti-union?”. The low coherence in the topic models extracted here contribute to a few poorly aligned topics that compromise the quality of inferences that can be drawn from this model, particularly for candidates who are tweeted about less often.

Twitter users were generally more positive towards the budget topic compared to compared to their baseline opinion of the candidate. In particular, Twitter users gave unusually high praise to Ted Cruz and Rand Paul in tweets concerning budget issues. However, most tweets in this topic seem only very loosely related to the topic--because the documents are so short, the word “job” becomes very important for the budget topic, but in the Twitter data, this term is often used in contexts like “Governor did a bang up job in the #GOPDebate” that do not speak to issues of

budget, taxes, and job--a critical difference between the way language is used within the debate itself and by Twitter users.

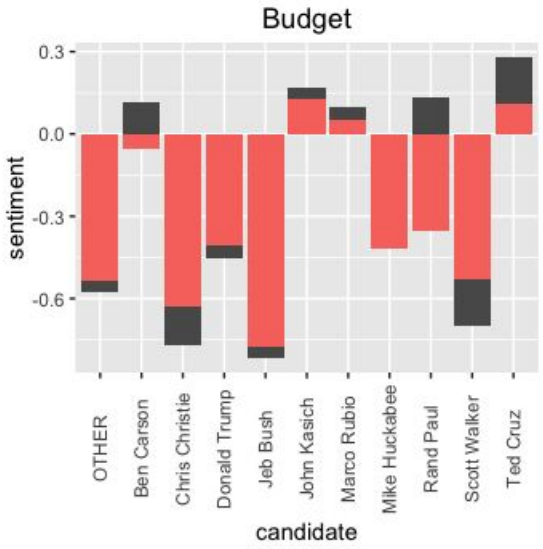

Figure 5: Overall and topic-specific Twitter sentiment towards each candidate for the budget topic drawn from the debate text

Discussion

There are great differences in how language is employed by candidates on national television for a political debate and Twitter users reacting to the debate from home. While these differences limit the utility of topic models trained directly on Twitter data itself for addressing questions of policy, topic models can be trained on a corpus of more politically relevant texts which discuss substantive issues for which sentiment mining provides a relatively inexpensive window onto public opinion without the use of polling data. Some information is certainly lost due to inherent differences in the format and style of these text domains. However, our ability to understand sentiment from these out-of-domain texts is nearly as strong as our ability to understand sentiment from the in-domain standard topic models or the manually annotated subject matter

labels collected by crowdsourcing. What we gain in exchange as a policy researcher is the ability to set, with a greater degree of refinement, the texts that comprise the set of policy issues of interest. We prioritize the semantic coherence of the topics we are interested in over their predictive power, and find that in fact, surprisingly, we have sacrificed very little predictive and inferential power after all.

Here, our ability to predict sentiment well is merely a confirmation that our procedure is working acceptably well--as researchers we care about the inferred relationships between topic and sentiment that our models describe. This process could be greatly improved by expanding the source of policy-oriented texts for training these topic models. In our example, we used the exchanges from a single Republican presidential debate, lasting about an hour and a half in duration. From this very limited text, with a vocabulary of only just over 200 words, we were able to predict sentiment nearly as well as some of the more straightforward models. However, by training these topic models on a more expansive body of text--for example, all televised debates, town halls, press releases, and issue statements--with the texts chosen by the researcher to encompass the breadth of topics of interest, we can greatly improve the model. Adding some degree of supervision to the topic models, for example, by using policy-relevant anchor terms in anchor-based topic models and spectral inference, would likely improve our ability to extract coherent topics that we are truly interested in gauging sentiment for.

Code

Our code can be found at https://github.com/pinesol/texas/tree/master/debate

References

"24 Million Watched US Presidential Debate": Nielsen. Web. 10 May 2016. <https://www.yahoo.com/news/massive-viewer-numbers-us-presidential-debate-fox-190446396.html?ref=gs#>.

Benoit, Kenneth, and Alexander Herzog. "Text Analysis: Estimating Policy Preferences From Written and Spoken Words." 17 Feb. 2015. Web. 08 May 2016. <http://www.kenbenoit.net/pdfs/HerzogBenoit_bookchapter.pdf>

Hong, Liangjie, and Brian D. Davison. "Empirical Study of Topic Modeling in Twitter." Proceedings of the First Workshop on Social Media Analytics - SOMA '10 (2010): n. pag. Web. <http://snap.stanford.edu/soma2010/papers/soma2010_12.pdf>.

Roberts, Margaret E., Brandon M. Stewart, and Edoardo M. Airoldi. "A Model of Text for Experimentation in the Social Sciences." Journal of the American Statistical Association (2016): 1-49. Web. <http://scholar.princeton.edu/files/bstewart/files/stm.pdf>.

D. M. Blei, A. Y. Ng, and M. I. Jordan. "Latent dirichlet allocation." *The Journal of Machine Learning Research*, 3:993–1022, 2003.

Roberts, Margaret E., Brandon M. Stewert, and Dustin Tingley. "Stm: R Package for Structural Topic Models." Journal of Statistical Software (2016) VV.II (n.d.): n. pag. Print.