Paper 4053-2019

Identifying Partisanship in Media Articles

Alex Lobman, Lohit Bhandari, Ravi Josyula, Nhan Nguyen

Oklahoma State University

ABSTRACT

In the United States, political polarization has become a significant issue. News media bias and the manner in which social networks deliver news to individual users have both been implicated as playing an essential role in increasing this polarization. Many have suggested that identification and disclosure of media bias would help address this divisive issue. In this paper, a method is demonstrated for data scientists to detect media bias without the bias of the data scientist influencing the results. The paper examines only partisan bias and focuses only on the two major political parties in the United States. The methods used were to build a predictive model on congressional house speeches, labeling each speech based on the speaker's party and then using the model to score news media articles. By using a text topic node combined with logistic regression in SAS Enterprise Miner, just over a 92% accuracy rate at distinguishing whether a Republican or Democrat made a congressional speech was achieved. This method could be used to create “bias checkers” to create flags on articles which would assist readers to evaluate their content and allow social media sites to ensure opposing viewpoints are displayed.

INTRODUCTION

In the United States, political polarization has become a significant issue. A Pew Research Poll found that over 27% of each political party view the other political party as “a threat to the nation’s well-being,” up from 17% in 1994 (Suh, 2016). One reason polarization has increased is due to the emergence of ‘echo chambers,’ wherein individuals are increasingly consuming news which conforms to their political affiliation (Flaxman, 2016). An ‘echo chamber’ can occur when an individual uses a search engine, news aggregator or social network which displays personalized content using machine learning (Das, 2007; Flaxman, 2016). In other words, as people increasingly access the news online, the news which is displayed is skewed toward that they have previously selected, thereby giving them results more likely containing opinions conforming to what they had previously read. Researchers have found that exposure to opposing viewpoints can increase political tolerance by improving awareness and rationales of opposing views (Muts, 2002). Thus, this ‘echo chamber’ is an essential factor in increasing polarization.

The mainstream media has been criticized for being biased both by President Donald Trump, who even calls it “Fake,” (Borchers, 2018) and by U.S. adults, 62% of whom believe the news is biased (Jones, 2018). There is value in Data Science being able to evaluate such claims. However, there are two problems from a data perspective in verifying a biased news outlet claim. The first is determining what kind of bias may be involved. Some of the types of bias include ideological — partisan, religious. Second, modeling cannot be done using rules chosen by the modeler who is also likely to be at least subconsciously biased, and training models must also be developed without pre-classifying news articles as biased.

This paper aims to solve the problem in two phases: 1) building a predictive model to identify the political party of a congressional speech and 2) scoring news media articles based on the model. In Phase 1, the scope of bias was narrowed to partisan bias and only considered Democrats and Republicans. To prevent bias on the part of the data scientist, a textual analysis was done of congressional floor speeches, which are, by definition, biased by the party speaker. A predictive model was then built to classify the party of the speaker of these floor speeches. Phase 2 then used the predictive model to score news media articles. Partisan bias is identified based on whether an article is more similar to a Democrat's or Republican's floor speech.

METHODS

DATA COLLECTION

The data collected consisted of a corpus of 1000 political floor speeches from both Democrat and Republican politicians in both houses of Congress. To minimize the bias of the corpus, the speeches gathered were equally balanced. Two hundred fifty speeches were copied to text files from each party in each house of Congress. Partisan bias likely changes over time, and therefore only 2017-2018 floor speeches for the 115th United States Congress (January 3rd, 2017 – October 22nd, 2018) were retrieved. The speeches came from www.votesmart.org, a non-profit that has transcriptions of many, but not all, congressional floor speeches. Speeches were selected starting with politicians in a leadership position to ensure they were in the 250 speeches selected. No more than ten speeches were chosen from an individual politician and most had fewer based on availability. Speeches were only excluded if they were about a natural disaster or a specific event such as acknowledging national event such as recognizing National Police Week. That was done to focus on political divides most useful for checking for news articles' political bias.

PROJECT APPROACH

Refer to figures A.(i) and A.(ii) in Appendix A to visualize the project approach.

DATA PREPARATION AND VALIDATION

The congressional floor speech files were gathered into two folders, one for Republican speeches and one for Democrat speeches, each containing the 500 speeches gathered from VoteSmart. Two Text Import nodes were used to import the speeches from their respective folders. The imported textual data was validated, in that, it showed no omissions or truncations. Using two SAS Code nodes, each observation in the two datasets created from the Text Import Nodes were labeled by creating a new variable, 'party,' which took the value 'Democrat' for Democrat floor speeches and 'Republican' for Republican floor speeches. Next, the dataset was partitioned into 70% training and 30% validation.

DATA CLEANING, MANIPULATION AND RATIONALE

After the data partitioning, text parsing was used on the speeches. The parsing process cleaned and modified the floor speeches' textual data. The speeches were tokenized, meaning each word became a separate variable, and a number would be assigned to each speech representing the number of times each word was used in that speech. Spaces, stop words, numbers and punctuation marks were removed. There were no changes to the default settings to minimize added human bias.

After the text was parsed, it was filtered to reduce the number of words. If a word did not appear in at least four congressional floor speeches, it was removed. Next, text topics were created. The idea behind the text topics was to find similar patterns of words occurring in multiple documents, and then grouping those words together to form a topic. That process was unsupervised, meaning that the topics created were not influenced by the data scientists or their bias. Text topics then have a many to many relationship to documents, meaning that one document can fall into many topics. That methodology was superior to text clustering, which is similar to text topics, except, each speech could only belong to one cluster. The ability of a speech to fall into many text topics allows for the detection of multiple different indicators of bias to be considered. For example, if a speech had mentioned both the 'Dreamers' and the 'Paris Agreement,' it would have fallen into two different text topics both highly correlated with the Democrats' Congressional speeches, whereas, if text clustering were used it would have been assigned only one cluster. Raw text topic probabilities were used in the predictive models. What that means is that each document had a probability of belonging to each of the 25 generated text topics. f a news article about immigration uses words associated with Democrats such as 'Dreamers' as well as words associated with Republican such as 'chain migration' when predicting the bias of the article, the model factors the use of the terms on each side of the political spectrum.

RESULTS

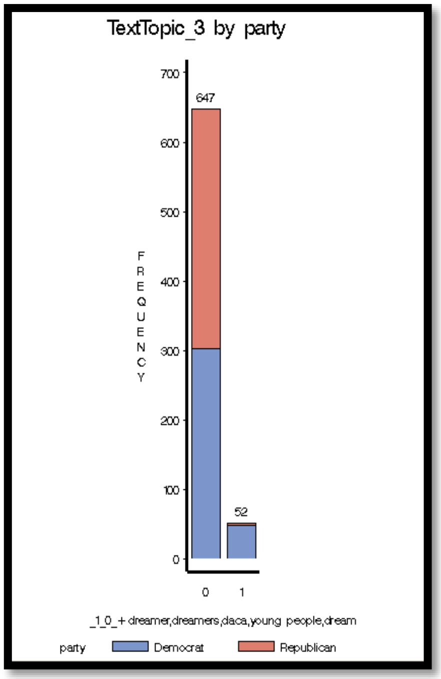
TEXT TOPICS

25 text topics were able to group together partisan terms, which when used together form common partisan issues from the text corpus. Additionally, the text topics were able to separate topics relating to similar issues based on the words commonly used when members of different parties talk about the related policy legislation. That is because each party had distinct words it used when talking about the certain political issues.

Example:

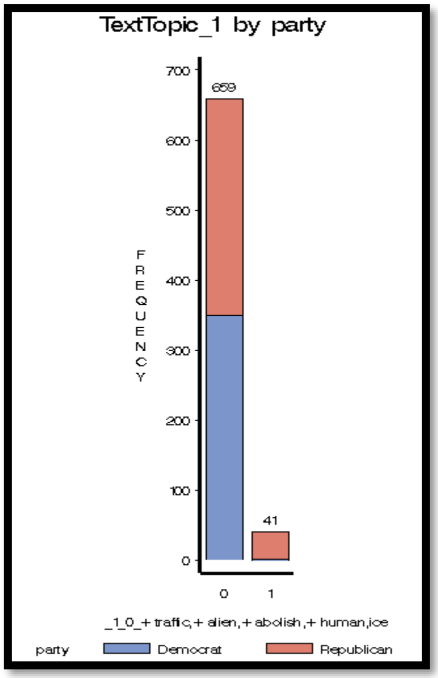
Primarily Democratic Text Topic – Undocumented Immigration:

Key Terms: dreamer, DACA, young people



Primarily Republican Text Topic – Illegal Immigration:

Key Terms: traffic, alien, abolish, ICE



MODELS

The champion model was a logistic regression with a validation accuracy rate over 92%. The classification table can be found in appendix B, figure B. (i).

SCORING RESULTS

The champion model was used to score articles from https://www.huffpost.com, who's viewers according to a Pew Research study are more often on the political left, and from https://www.breitbart.com, who's viewers are more often on the political right (Mitchell, 2018). Moreover, both websites were founded by Andrew Breitbart, who self described Breitbart as "the Huffington Post of the right" (Cox, 2016). 75 articles from each site were gathered by searching for the term "immigration." In order to minimize scoring bias all 75 articles returned when searching on Breitbart were used as well as the first 75 articles returned from the Huffpost Post. Unfortunately, a specific date range cannot be chosen when searching. The term immigration was used to ensure that the articles selected would be related to text topics created. The data was imported to SAS Enterprise Miner using the same techniques as the congressional speeches. The resulting confusion matrix is below.

Actual	Predict: Democrat	Predict: Republican
Breitbart (Republican)	16	59
The HuffPost (Democrat)	64	11

Table 1. Confusion Matrix for the Regression on the Scoring articles

GENERALIZATIONS

The confusion matrix shows that the model predicts the political party of the two websites as aligning with the expected political party approximately 80% of the time. The scoring results are an expected result but not a true measure of accuracy because not all articles on these websites are likely to be biased. However, the scoring results demonstrate that the model created on political speeches was not only able to predict the political affiliation of congressional floor speeches, but it was also able categorize news articles as expected based on the expected bias of the article's website. When trying to generalize this model to a specific news author, it is useful to see if there is a consistent pattern to their articles. Some individual authors were considered, including Peggy Noonan, a conservative writer for the Wall Street Journal, and her articles scored as being most similar to the Democrats nearly 50% of the time. That is a good indication that even though Noonan might have her political affiliation, her articles appear to be more balanced. Adding a no bias category to the target might be difficult because humans would have to agree on a news article that is not biased. Thus, when looking for bias by specific authors or websites, it is possible to use the ratio of the number of articles which the model predicts have Democrat bias to the number of articles the model Predicts are Republican biased. Comparing that ratio to 1 provides insight into of whether that individual or news organization is biased.

FUTURE STUDIES

One limitation to the model presented in this paper is that it can only score based on the topics created from the Congressional speech text corpus. As an example, because there were no Congressional speeches in text corpus that referenced abortion, bias in news articles about abortion would not be appropriately classified. A future study could expand the number of congressional speeches in the corpus to generate additional text topics. Additionally, it is possible that a future study using a Recurrent Neural Net, available in SAS Viya, could produce better results. Finally, these same methods could be applied to find other kinds of bias. For example, the Corpus could be include writings of other political parties, religious groups, or other ideologies to expand the types of bias detected.

CONCLUSION

The paper has demonstrated one approach for data science to answer some of the questions regarding whether news articles bias detection is possible in a non-biased way. The model accuracy was above 90% both on the congressional speech corpus as well as on the articles about immigration from the two politically biased websites. One application of this model is a browser extension that keeps track of the political bias of the news articles an individual reads, so that individuals could know if they are trapped in an echo. If they were trapped, they could escape by using the extension to identify articles on the other side of the issue. Another application is that a search engine such as Google could indicate which kind of bias a news article has, so that readers could be more conscious of how biased the information they are consuming is. Finally, search engines could manipulate results to ensure articles from multiple types of bias are displayed.

REFERENCES

VoteSmart. 2018. Accessed November 2018. https://votesmart.org.

Borchers, Callum. “The Many Ironies of Trump's Tweets about 'Unfair' Coverage of Melania's Public Absence.” *The Washington Post*, WP Company, 6 June 2018, www.washingtonpost.com/news/the-fix/wp/2018/06/06/the-many-ironies-of-trumps-tweets-about-unfair-coverage-of-melanias-public-absence/?utm_term=.2b9191893e23.

Cox Media Group National Content Desk. “Breitbart News: A Look at the Site Making Headlines for Its pro-Trump Stance.” *Ajc*, Cox Media Group National Content Desk, 1 Sept. 2016, www.ajc.com/news/national-govt--politics/breitbart-news-look-the-site-making-headlines-for-its-pro-trump-stance/4pll66YqrJtaLI7UxdboeM/.

Das, Abhinandan S., et al. “Google News Personalization.” *Proceedings of the 16th International Conference on World Wide Web - WWW '07*, 2007, doi:10.1145/1242572.1242610.

Flaxman, Seth, et al. “Filter Bubbles, Echo Chambers, and Online News Consumption.” *Public Opinion Quarterly*, vol. 80, no. S1, 2016, pp. 298–320., doi:10.1093/poq/nfw006.

Jones, Jeffery M. “Americans: Much Misinformation, Bias, Inaccuracy in News.” *Gallup.com*, Gallup, Inc, 20 June 2018, news.gallup.com/opinion/gallup/235796/americans-misinformation-bias-inaccuracy-news.aspx.

Mitchell, Amy, et al. “Political Polarization & Media Habits.” *Pew Research Center's Journalism Project*, Pew Research Center's Journalism Project, 26 Apr. 2018, www.journalism.org/2014/10/21/political-polarization-media-habits/.

Mutz, Diana C. “Cross-Cutting Social Networks: Testing Democratic Theory in Practice.” *American Political Science Review*, vol. 96, no. 01, 2002, pp. 111–126., doi:10.1017/s0003055402004264.

Suh, Michael. “Political Polarization in the American Public.” *Pew Research Center for the People and the Press*, Pew Research Center for the People and the Press, 11 Oct. 2016, www.people-press.org/2014/06/12/political-polarization-in-the-american-public/.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Alex Lobman
Oklahoma State University
alobman@okstate.edu

Lohit Bhandari
Oklahoma State University
lohit.bhandari@okstate.edu

Ravi Josyula
Oklahoma State University
ravi.josyula@okstate.edu

Nhan Nguyen
Oklahoma State University
nhan.nguyen@okstate.edu

APPENDIX A: PROCESS FLOW

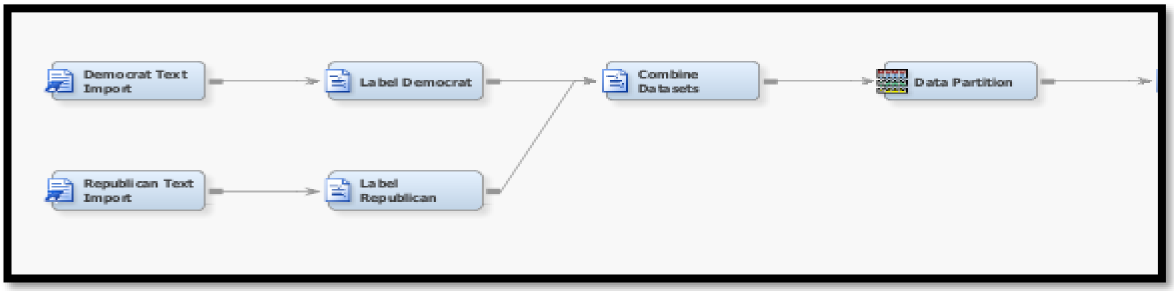


Figure A. (i) Process showing importing and partitioning raw text documents

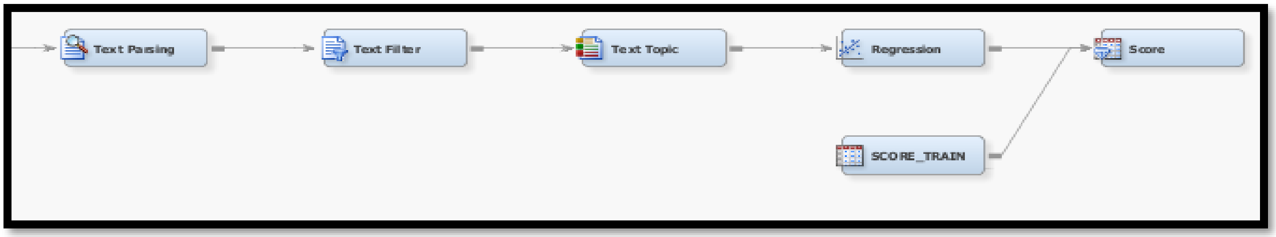
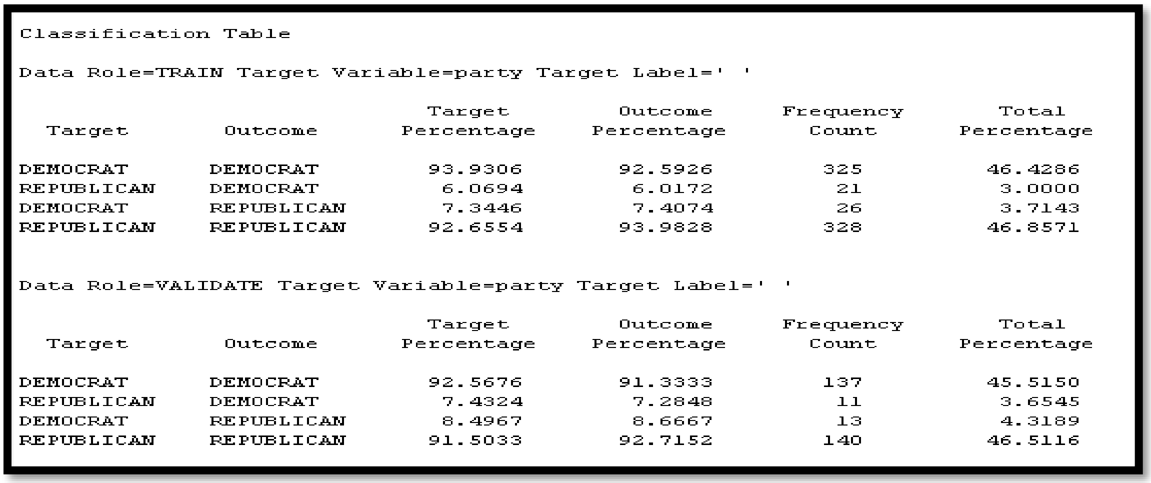


Figure A. (ii) Process showing data preparation, modelling, and scoring

APPENDIX B: MODEL RESULTS



```
Classification Table

Data Role=TRAIN Target Variable=party Target Label=' '

                                    Target       Outcome     Frequency      Total
  Target         Outcome          Percentage   Percentage      Count      Percentage

DEMOCRAT       DEMOCRAT             93.9306      92.5926        325         46.4286
REPUBLICAN     DEMOCRAT              6.0694       6.0172         21          3.0000
DEMOCRAT       REPUBLICAN            7.3446       7.4074         26          3.7143
REPUBLICAN     REPUBLICAN           92.6554      93.9828        328         46.8571


Data Role=VALIDATE Target Variable=party Target Label=' '

                                    Target       Outcome     Frequency      Total
  Target         Outcome          Percentage   Percentage      Count      Percentage

DEMOCRAT       DEMOCRAT             92.5676      91.3333        137         45.5150
REPUBLICAN     DEMOCRAT              7.4324       7.2848         11          3.6545
DEMOCRAT       REPUBLICAN            8.4967       8.6667         13          4.3189
REPUBLICAN     REPUBLICAN           91.5033      92.7152        140         46.5116
```

Figure B.(i) Congressional speeches classification table

APPENDIX C

Primarily Democratic Text Topic – Climate Change:

Key Terms: climate change, Paris [agreement], Science

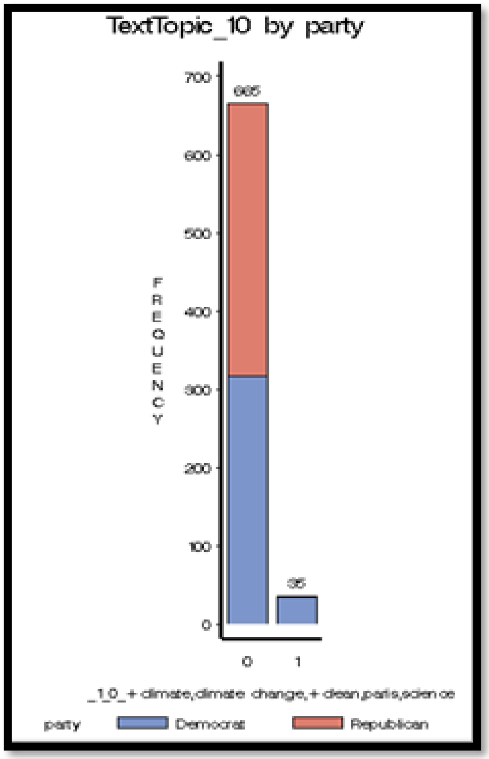


Figure C.(i) More Examples of Text Topics

Primarily Republican Text Topic – Overregulation of Financial Institutions:

Key Terms: Regulatory, regulation, financial institutions

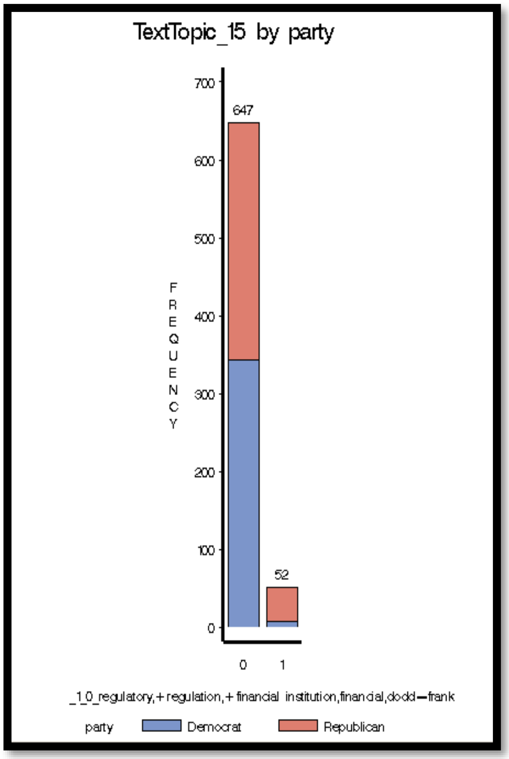


Figure C.(ii) More Examples of Text Topics