Proposal

Meaghan Burke, Justin Herman, Vikas Sinha, Joseph Garcia

9/20/2019

Introduction

Our capstone project aims to measure the bias in media coverage of candidates in US Presidential Elections. Assuming that such bias is found, we aim to quantify it for a number of selected outlets in print media and cable television.

Various studies such as [5] have measured the effect of media coverage on presidential elections and concluded that it affects election outcomes in small but measurable ways. Measuring media bias in political coverage and presidential elections in particular, is important as it can potentially shape public opinion and policy regarding media regulation.

Previous Approaches

There have been a number of approaches directed at the general problem of measuring and quantifying media bias.

Sheth[1] describes an approach for measuring media bias in which news items from selected new sources are assigned bias scores for several types of biases. Then the bias score for each news source is computed as a distance between its extreme values in a vector space of different types of bias represented numerically. The bias score is also scaled by the number of new items from that news source.

In Budak[2] bias (or "slant") is quantified based on a scoring of news articles on a range of topics spanning the ideological spectrum, such as Environment, Gun Control and Healthcare. The scoring is performed by people through use of the crowdsourcing platform Amazon Turk.

One way in which our proposal differs from previous approaches is that it focuses on bias in coverage of individual candidates, not political bias in general.

Hypothesis

Our hypothesis is that publications negatively portray certain candidates more than other candidates. We will be using sentiment score to track a custom aggregated score for each candidate in each individual publication. Our custom KPI (Negative bias) will attempt to track negative mentions of a candidate in several publications. Depending on data collection, we may need to deal with differences in volume of mentions in the publications. We could do so by clustering candidates into categories based upon either volume or polling numbers

Cluster 1: Biden,Bernie, Warren

Cluster 2: Buttigeg, Harris, Booker,

Cluster 3- Yang, Beto

Our hypothesis test will look like this:

Washington Post(given Clusters):

Null Hypothesis: there is no difference in Negative_bias within Cluster 1 mean score

Alternative Hypothesis: there is a statistical difference in the Negative_bias mean score

Null Hypothesis: there is no difference in Negative_bias within Cluster 2 mean score

Alternative Hypothesis: there is a statistical difference in the Negative_bias mean score

Null Hypothesis: there is no difference in Negative_bias within Cluster 3 mean score

Alternative Hypothesis: there is a statistical difference in the Negative_bias mean score

Washington Post(Without Clusters):

Null Hypothesis: there is no difference in Negative_bias mean score

Alternative Hypothesis: there is a statistical difference in the Negative_bias mean score

Solution

As with any issue of bias, knowledge and awareness is important. Our data will hopefully allow publications to realize that unintendedly they are expressing bias. Most news organizations value truth and will hopefully take measures to figure out how to address that bias. Otherwise, public pressure is typically the best method to force corporations to act. Even if we can't get corporations to act on minimizing their bias, making the public aware of potential bias allows them to understand that disproportionately certain organizations may favor certain ideological positions.

Dataset & Related Technologies

The group plans to primarily use Python modules to access, scrape, consolidate, explore, classify, and model the newspaper data.

- **requests** module to get the HTML code from the newspaper's webpage
- **BeautifulSoup** module to navigate through and extract elements from the HTML
- **Pandas**module to perform exploratory data analysis
- **Seaborn** module to visualize and further explore the data
- **nltk** and **scikit-learn** modules to perform all-natural language processing (NLP) and modeling components
  - Word Frequency "Bag-of-Words"
    - CountVectorizer (scikit-learn)
    - TfidfVectorizer (scikit-learn)
    - FreqDist (nltk)
  - Split dataset into train/test components (scikit-learn)
  - Train models (scikit-learn)
  - Random Forest
  - Support Vector Machine
  - Prediction and evaluation of model (scikit-learn)

References

1. Sheth, Dev. Measuring Ideological Bias in News Coverage of Political Events by Print Media using Data Analytics. https://pdfs.semanticscholar.org/791d/c8b2f3feb1d731bce0a8b9fa46f2ec0516f1.pdf
2. Budak et al. Fair and Balanced? Quantifying Media Bias through Crowdsourced Content Analysis. https://www8.gsb.columbia.edu/media/sites/media/files/JustinRaoMediaBias.pdf
3. Is the media biased toward Clinton or Trump? Here is some actual hard data. https://www.washingtonpost.com/news/monkey-cage/wp/2016/09/20/is-the-media-biased-toward-clinton-or-trump-heres-some-actua

4. Peng, Yilang. Same Candidates, Different Faces: Uncovering Media Bias in Visual Portrayals of Presidential Candidates with Computer Vision. https://academic.oup.com/joc/article-abstract/68/5/920/5113150?redirectedFrom=fulltext

5. Media Bias and Voting. https://www.nber.org/digest/oct06/w12169.html