Stylometry as a measure of media bias

Emily Jia*

May 4, 2018

1. Introduction

News reporting is frequently accused of political slant, but it is difficult to find metrics that quantify media bias. Current methods of measuring bias evaluate the content of an article to determine its framing. Jacobi et al. demonstrate that topic modelling through Latent Dirichlet Allocation can be used to classify journalistic documents [1]. Budak et al. use supervised learning and then crowdsource human judges to organize political articles by topic and political slant [2]. Niculae et al. find systematic bias in the way news sources quote presidential speeches and relate this bias to slant [3]. Groseclose and Milyo tabulate news articles citations of think tanks to score their political alignment with members of Congress [4]. Although existing work demonstrates news outlets in the US do differ in their position in the liberal-to-conservative spectrum, the magnitude of the differences are under debate, likely due to variation in bias across different news events [2].

To categorize news sources across multiple news events, we turn to stylometry, which is the quantitative study of literary style. One of the most conventional uses of stylometry is in determining or disputing authorship of older texts. Common metrics include the frequency of stylistic choices made at clause, phrase, and word level [5]. Using the Stanford Natural Language processor, we compute these frequencies across thousands of news articles from five news sources to see if stylistic choices can be used as a metric for political slant. We also compute the distribution of lexical diversity across articles published by the same news source.

2. Method

For data, we collect the text of thousands of political news articles from Fox News, Breitbart News, Associated Press, The New York Times, and The Washington Post. The NYT API or NewsAPI was used to fetch the URLs of recently published articles. Articles were required to contain the word Trump to ensure political content. Due to varying limits on API queries by source and API, the number of URLs collected varies across news sources (Table 1).

*Department of Mathematics, Harvard College, Email: ejia@college.harvard.edu

	URLs	Articles Recorded	Proportion of URLs Recorded
Fox	24100	3250	0.13
Breitbart	12800	4703	0.37
AP	10000	1616	0.16
NYT	453	174	0.38
WaPo	4980	1639	0.33

Table 1: URL and article counts by news source

After collecting the URLs, the BeautifulSoup library was used to extract the text from each article and the RAKE (Rapid Automatic Keyword Extraction) library was used to determine keywords of the article. If Trump appeared in the top 10 keywords, then the text of the article was recorded. The Trump family generally played a passive role in articles that did not have Trump in the top keywords. (Example.) Across news sources, 13-37% of articles mentioning Trump had Trump as a keyword.

Once the text of the articles was collected, the Stanford NLP Parser tagged each sentence in the text with the Penn Treebank II tags. The LexicalizedParser englishPCFG with Penn-style output was used with maximum sentence length of 120. Each tree contains information about the word, clause, and phrase structure of a sentence. From the tree files, we collected all information about clause and phrase level structures, as well as word tags related to adjectives and adverbs.

```
(ROOT
  (S
    (NP (NNP Kirby))
    (VP (VBD said)
      (SBAR
        (S
          (NP
            (NP (DT a) (JJ big) (JJ military) (NN parade))
            (PP (IN in)
              (NP (NNP Washington))))
          (VP (VBZ is)
            (NP (DT a) (JJ bad) (NN idea))))))
    (. .)))
```

Figure 1: Tags for the sentence “Kirby said a big military parade in Washington is a bad idea”

Because we had the article texts, we were also able to compute vocabulary richness using Yules K characteristic. Yules K has the special property that its value converges rather than decreases as article length varies. Other measures are very sensitive to text length because new words are introduced at a slower rate as the length of an article increases [6].

3. Results

To see if these metrics were good predictors of bias, we use Budak etal's crowdsourced results on political slant to quantify bias [2]. Budak et al. measure bias as positive for conservative outlets, zero for center, and negative for liberal outlets (see Table 3).

	bias
Breitbart	0.180
Fox	0.100
AP	0
WaPo	-0.020
NYT	-0.050

Table 2: Bias, as measured by Budak et al. [2]

Out of the 34 tags we counted, three of them appeared to correlate, all negatively, with bias such that the value for AP was between the values for Breitbart/Fox and NYT/WaPo. The tags are:

- Subordinate clause (tag: sbar)
  - "After the project," in "After the project, Emily was really glad she did it."
- Wh-noun phrase (tag: whnp)
  - "what a wh-noun phrase is" in "I learned what a wh-noun phrase is."
- Wh-preposition phrase (tag: whpp)
  - "with whom I worked" in "I met Ski, with whom I worked."

A linear regression of these three tags against bias gives a multiple R of 0.998 and high t-statistics.

	coefficient	t-Stat	p-value
intercept	2.54	8.77	0.072
sbar	-3.68	-7.28	0.09
whnp	0.78	1.88	0.31
whpp	59.06	6.97	0.09

Table 3: Bias, as measured by Budak et al. [2]

We also compute the distribution of Yule's K characteristic per article, which measures vocabulary richness, for all five of the news sources. Although the means are roughly the same, variance is positively correlated with bias.

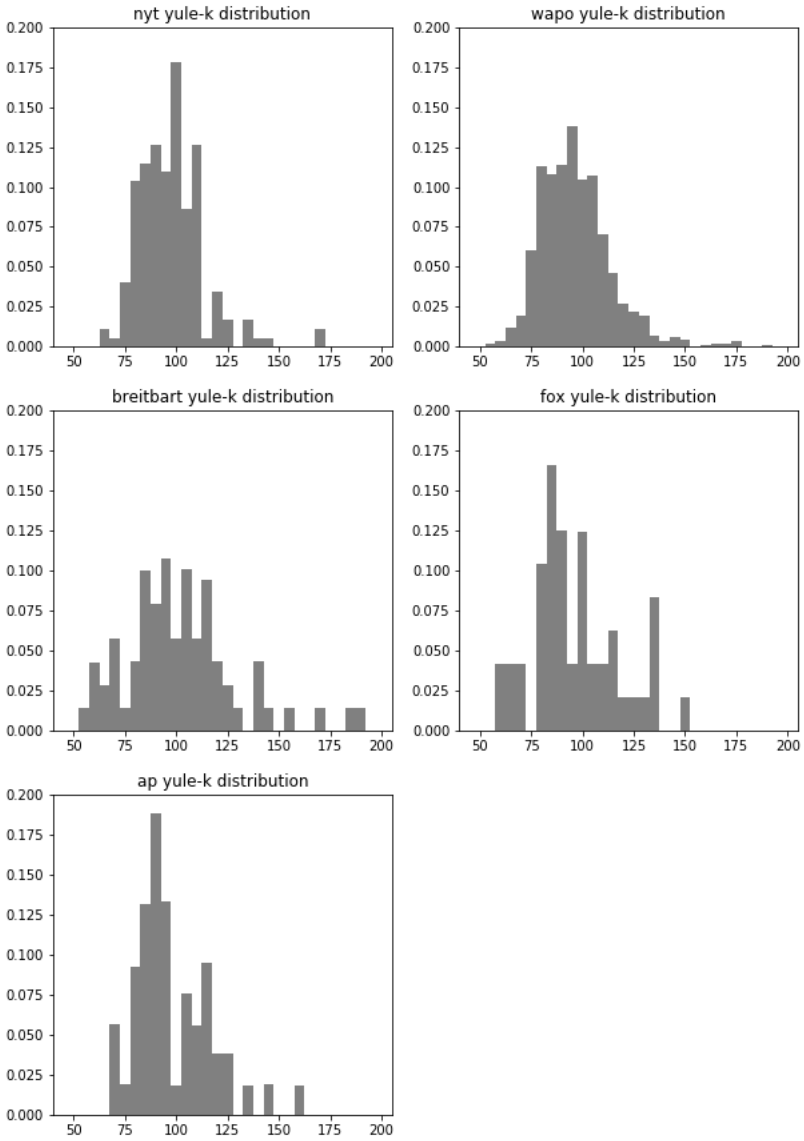

Figure 2: Yule's K across the news sources. Variance increases as sources become more conservative.

4. Discussion

Although our linear regression had a high R value, further work should test the linear regression model on more news sources. An intercept of 2.54 and a coefficient of 59.07 seems large for modelling bias that ranges from 0.2 to -0.2, but forcing a constant of 0 also does not seem reasonable. Also, the coefficients do not seem interpretable: conservative outlets use fewer WHPP, but the coefficient of WHPP is 59.06. It is necessary to collect more data for the model and then test the model on more data. Looking at Yule-K characteristic variance as a metric of bias is another direction that future work could consider. Following up on the paper by Niculae et al. on quoting patterns, it would also be interesting to look at the stylometry of quotes [3].

5. Project Reflections

The original project proposal, which proposed the tagging we did as well as more testing and possibly a machine learning model, estimated that data collection would only take a few weeks. However, the data collection process took much longer than expected. API requests had limits on 6-hour and 24-hour usage, collecting article texts took several hours per source, and the NLP took several more hours per source.

We also had to start the data collection process over completely. We initially used the python newspaper library but ran into an issue with the robot.txt of one of our news sources. However, we had been use the python newspaper NLP package to determine keywords, so we needed to re-collect all of the articles while using a different NLP package to determine keywords.

As mentioned in the discussion, the next direct steps are to collect more data and test our model. It would be interesting to compare dendrograms of all news sources, organized by bias vs. organized by similarity of styolmetric choices.

6. Acknowledgements

I would like to thank my mentor, Ski Krieger, for granting me both agency and guidance in this project. I appreciate his flexibility in letting me propose this topic and make decisions about how I wanted to explore it, and also his knowledge about cultural evolution and existing tools for NLP and data processing. I would also like to thank Sam Sinai and Professor Martin Nowak for running Math 243 and giving me the opportunity to learn about mathematical biology and complete my own project.

References

[1] C. Jacobi, W. van Atteveldt, and K. Welbers, "Quantitative analysis of large amounts of journalistic texts using topic modelling," *Digital Journalism*, vol. 4, no. 1, pp. 89–106, 2016.

[2] C. Budak, S. Goel, and J. M. Rao, "Fair and balanced? quantifying media bias through crowdsourced content analysis," *Public Opinion Quarterly*, vol. 80, no. S1, pp. 250–271, 2016.

[3] V. Niculae, C. Suen, J. Zhang, C. Danescu-Niculescu-Mizil, and J. Leskovec, "Quotus: The structure of political media coverage as revealed by quoting patterns," in *Proceedings of the 24th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 2015, pp. 798–808.

[4] T. Groseclose and J. Milyo, "A measure of media bias," *The Quarterly Journal of Economics*, vol. 120, no. 4, pp. 1191–1237, 2005.

[5] J. Grieve, "Quantitative authorship attribution: An evaluation of techniques," *Literary and linguistic computing*, vol. 22, no. 3, pp. 251–270, 2007.

[6] K. Tanaka-Ishii and S. Aihara, "Computational constancy measures of textsyule's k and rényi's entropy," *Computational Linguistics*, vol. 41, no. 3, pp. 481–502, 2015.