Assessing Political Bias using Crowdsourced Pairwise Comparisons

Tzu-Sheng Kuo, McArdle Hankin, Miranda Li, Andrew Ying, Cathy Wang
Stanford University
{tskuo, mcardle, mirandal, androo, cathywnx}@stanford.edu

Abstract

We propose a crowdsourced approach for assessing political bias among news publishers. We aggregate pairwise comparisons between publishers into a relative ranking of leanings, in contrast to prior crowdsourcing approaches that request independent ratings of publishers. We verify our approach by a preliminary study with 45 participants, who were asked to rank the political bias of 8 news publishers and to provide qualitative feedback on the study. We report our experimental results and describe our envisioned application scenario.

Introduction

While differing political bias among news publishers is not a new phenomena (Groseclose and Milyo 2005), methods of news consumption have changed with the rise of news aggregators like Facebook (Farrell 2012). A recent study found that readers desire less biased news (Mitchell et al. 2018), but often choose articles that confirm their biases (Bakshy, Messing, and Adamic 2015; Jakesch et al. 2019), in part due to news aggregation algorithms (Conover et al. 2011). Since these algorithms are often trained by a user's browsing history, the resulting news cycle often reinforces a reader's political leaning (Munson, Lee, and Resnick 2013).

Much effort has been made to identify biased news on news aggregators. For example, Facebook manually examines news articles by experts (Bhuiyan et al. 2020). However, since expert feedback is hard to scale, initiatives such as WikiTribune resort to scalable crowdsourcing methods, which suffer from inconsistent evaluation standards that vary among crowds (Bhuiyan et al. 2020). While automated algorithms help with scaling (Baly et al. 2018), they still require crowds labeling in the first place (Adair et al. 2019).

We propose a method that evaluates news articles by crowds while avoiding their inconsistent evaluation standards. We are motivated by our daily observation that regardless of personal political leaning, people generally agree about the relative leanings of different sources. For example, someone who self-identifies as staunchly conservative might consider The New York Times to be extremely left-leaning, while someone who self-identifies as more liberal might see it as more centrist, but both would agree that Breitbart is more right-leaning than The New York Times.

Based on this observation, we ask participants to read political articles from different publishers and compare their relative leanings. We then aggregate the results to a spectrum from left- to right-leaning. While similar ranking approach has been widely used in other domains (Asudeh et al. 2015; Xu et al. 2016), it has not been applied to ranking political bias among news publishers to the best of our knowledge. In a preliminary study, we examine the effectiveness of our approach and compare our result with other studies. We report our findings and envision an application that help news aggregators increase the politically balanced coverage.

Ranking Algorithm

In our study, we had participants view two articles at a time and assign one as more left-leaning and the other as more right-leaning or say that they were unsure. The results from multiple samples were aggregated into a ranking of article sources from relative left- to right-leaning. Analogously, in sports, games happen between two teams and there is either a winner and a loser or a tie. These results are aggregated into a seasonal ranking of teams from best to worst.

One common sports ranking algorithms is the Plackett-Luce algorithm (Turner et al. 2018), which deals with precisely our problem of turning pairwise comparisons into a ranking. It also handles an incomplete ranking so we can generate a scale without all-pairs comparison. This property is crucial for scaling because it will be difficult to compare an article with all the others when there are many articles.

Preliminary Experiment

We selected two articles from each of the eight publishers across the political spectrum as previously ranked by other studies. The articles were released within days of each other and were about the same story. Because our goal was to test variation in articles that are sold as evidence-based truth, we selected articles from the news section rather than opinions.

We only showed the headline and first paragraph of each article for three reasons: first, 59% of people only read headlines (Gabielkov et al. 2016); second, political news are usually formatted in an inverted pyramid style (Pöttker 2003); and third, to account for respondent fatigue (Johnson, Lehmann, and Horne 1990). Finally, we standardized fonts

Table 1: The number of times that each source was ranked by participants as relatively left or right leaning.

Source	Left	Right	Total
HuffPost	115	17	132
The New York Times	75	57	132
The Washington Post	64	51	115
Wall Street Journal	64	56	120
CNN	63	58	121
BuzzFeed	59	73	132
Fox	44	77	121
Breitbart	21	116	137

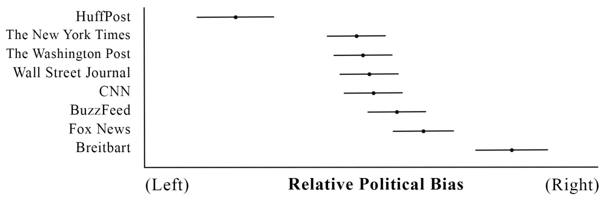

Figure 1: Political bias scale for selected sources.

and removed publishers' identifying information so that participants only made decisions based on the content.

45 US university students participated the study. Each of them was compensated $10 for evaluating 14 random article pairs. 27 self-identified as left-leaning and the rest as neutral. Although our participants may be relatively left-leaning comparing to the general population, we believe that people could determine relative leanings regardless of personal political standpoints. Our approach is also designed to mitigate subjectiveness by asking only for relative comparisons.

Result and Discussion

Of the 630 comparisons made, 125 pairs were marked as unsure. The remaining results are shown in Table 1. The Left and Right columns denote how many times an article from a source was ranked as relative left- or right-leaning. We utilize these data to construct Figure 1 by running the Plackett-Luce algorithm. The y-axis denotes the source and the x-axis denotes relative political leaning. For clarity, values (log worth) on the x-axis have been omitted because they are uninformative for our purposes; relative position of sources is the key. The horizontal bar for each datapoint denotes quasi-standard error. Overlap between two bars reflects low confidence in the relative ordering between the two sources.

We then compared our results with two representative methods that evaluate political bias. In Figure 2, the left diagram reflects the comparison with a content-based method that asks participants to rate articles using a five-point scale (Budak, Goel, and Rao 2016); the right diagram reflects the comparison with an audience-based method that measures political leaning of sources only based on the audience's political demographics (Mitchell et al. 2014). In each diagram, the x- and y-axis juxtapose ours and their rankings. The numbers on both axes (from 1-8) denote the sources that were ranked from relative left to right leaning. For example, the left-most point in the left diagram denotes that the

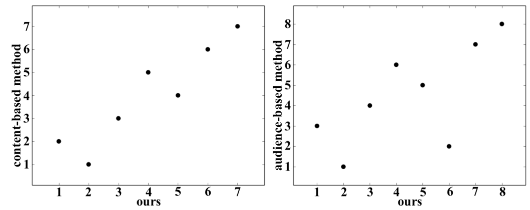

Figure 2: Comparison of our results with existing methods. Since the content-based method did not rank BuzzFeed, the left diagram only has seven sources.

most left leaning source in our result (HuffPost) is ranked the second-most left-leaning in the content-based method.

Spearman correlation assesses the consistency of our ranking with others'. For the left diagram, the analysis finds a rho value of 0.93 and a p-value of 0.002, demonstrating strong correlation. In contrast, the right diagram yields comparatively weaker rho and p-value of 0.69 and 0.06. The lower correlation is caused by discrepancies in the rankings. For example, BuzzFeed appears much further right in our ranking. Since we only chose two articles per source in this preliminary study, further experiments with more articles are needed for verification of this positioning.

According to qualitative feedback, participants generally agree that relative political leanings may be detected using tangible measures such as the use of language and the arrangement of sentences. For example, a participant replied: "*It (political bias) is evident by the use of language.*" Others reported the structure of articles also affected their decision. Examples include: "*Bias can't be found in THIS sentence or THAT sentence, but rather how the sentences fit together.*"

In summary, while this study has demonstrated the potential of our approach, future experiments with more sources, full articles, and participants with varying political leanings are required for further verification.

Application Scenario

Most existing measurements on political bias ask for and provide users with ratings. However, on the data collection end, ratings are susceptible to participants' inconsistent use of scales (Bolt and Johnson 2009), lower response rates, and respondent fatigue (Dolnicar, Grün, and Leisch 2011); on the application end, low ratings for biased articles actually backfire and increase readers' belief in what they have read (Berinsky 2017). To avoid backfiring, today's designs on Facebook and Twitter suggest related fact-checked articles to readers when they attempt to share misinformation.

Our approach can be integrated into news aggregators and help suggest related articles with different political perspectives, while preventing readers from facing direct refutations to their beliefs. Once readers have seen an article suggested based on the previous one they read, the system may ask them for a single-click binary comparison of political leaning between two articles. The collected feedback is aggregated into the system for future suggestions. It is our hope that our approach helps move towards a future where the wisdom of crowds makes the whole crowds wiser.

Acknowledgments

We would like to thank professors Maneesh Agrawala, Krishna Bharat, R.B. Brenner, and James Hamilton for their guidance on this study. We would also thank Kylie Jue and Danaë Metaxa for their feedback and support.

References

Adair, B.; Li, C.; Yang, J.; and Yu, C. 2019. Automated popup fact-checking: Challenges & progress. *Computation+Journalism Symposium*.

Asudeh, A.; Zhang, G.; Hassan, N.; Li, C.; and Zaruba, G. V. 2015. Crowdsourcing pareto-optimal object finding by pairwise comparisons. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, CIKM '15, 753–762. New York, NY, USA: Association for Computing Machinery.

Bakshy, E.; Messing, S.; and Adamic, L. A. 2015. Exposure to ideologically diverse news and opinion on facebook. *Science* 348(6239):1130–1132.

Baly, R.; Karadzhov, G.; Alexandrov, D.; Glass, J.; and Nakov, P. 2018. Predicting factuality of reporting and bias of news media sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 3528–3539. Brussels, Belgium: Association for Computational Linguistics.

Berinsky, A. J. 2017. Rumors and health care reform: Experiments in political misinformation. *British Journal of Political Science* 47(2):241–262.

Bhuiyan, M. M.; Zhang, A.; Sehat, C. M.; and Mitra, T. 2020. Investigating 'who' in the crowdsourcing of news credibility. *Computation+Journalism Symposium*.

Bolt, D. M., and Johnson, T. R. 2009. Addressing score bias and differential item functioning due to individual differences in response style. *Applied Psychological Measurement* 33(5):335–352.

Budak, C.; Goel, S.; and Rao, J. M. 2016. Fair and Balanced? Quantifying Media Bias through Crowdsourced Content Analysis. *Public Opinion Quarterly* 80(S1):250–271.

Conover, M. D.; Ratkiewicz, J.; Francisco, M.; Goncalves, B.; Flammini, A.; and Menczer, F. 2011. Political polarization on twitter. *Fifth International AAAI Conference on Weblogs and Social Media*.

Dolnicar, S.; Grün, B.; and Leisch, F. 2011. Quick, simple and reliable: Forced binary survey questions. *International Journal of Market Research* 53(2):231–252.

Farrell, H. 2012. The consequences of the internet for politics. *Annual Review of Political Science* 15(1):35–52.

Gabielkov, M.; Ramachandran, A.; Chaintreau, A.; and Legout, A. 2016. Social clicks: What and who gets read on twitter? *SIGMETRICS Perform. Eval. Rev.* 44(1):179–192.

Groseclose, T., and Milyo, J. 2005. A Measure of Media Bias*. *The Quarterly Journal of Economics* 120(4):1191–1237.

Jakesch, M.; Koren, M.; Evtushenko, A.; and Naaman, M. 2019. The role of source and expressive responding in political news evaluation. *Computation+Journalism Symposium*.

Johnson, M. D.; Lehmann, D. R.; and Horne, D. R. 1990. The effects of fatigue on judgments of interproduct similarity. *International Journal of Research in Marketing* 7(1):35 – 43.

Mitchell, A.; Gottfried, J.; Kiley, J.; and Matsa, K. E. 2014. Political polarization & media habits. *Pew Research*.

Mitchell, A.; Simmons, K.; Matsa, K. E.; and Silver, L. 2018. Publics globally want unbiased news coverage, but are divided on whether their news media deliver. *Pew Research*.

Munson, S.; Lee, S.; and Resnick, P. 2013. Encouraging reading of diverse political viewpoints with a browser widget. *International AAAI Conference on Web and Social Media*.

Pöttker, H. 2003. News and its communicative quality: the inverted pyramid—when and why did it appear? *Journalism Studies* 4(4):501–511.

Turner, H. L.; van Etten, J.; Firth, D.; and Kosmidis, I. 2018. Modelling rankings in r: the plackettluce package.

Xu, Q.; Xiong, J.; Cao, X.; and Yao, Y. 2016. Parsimonious mixed-effects hodgerank for crowdsourced preference aggregation. In *Proceedings of the 24th ACM International Conference on Multimedia*, MM '16, 841–850. New York, NY, USA: Association for Computing Machinery.