Proceedings of the Fourteenth International AAAI Conference on Web and Social Media (ICWSM 2020)

Empirical Evaluation of Three Common Assumptions in Building Political Media Bias Datasets

Soumen Ganguly,[1] **Juhi Kulshrestha,**[2] **Jisun An,**[3] **Haewoon Kwak**[3]
[1]Saarland Informatics Campus, Germany
[2]GESIS - Leibniz Institute for the Social Sciences, Germany
[3]Qatar Computing Research Institute, Hamad Bin Khalifa University, Qatar
soumen.ganguly@dfki.de, juhi.kulshrestha@gesis.org, {jisun.an, haewoon}@acm.org

Abstract

In this work, we empirically validate three common assumptions in building political media bias datasets, which are (i) labelers' political leanings do not affect labeling tasks, (ii) news articles follow their source outlet's political leaning, and (iii) political leaning of a news outlet is stable across different topics. We build a ground-truth dataset of manually annotated article-level political leaning and validate the three assumptions. Our findings warn that the three assumptions could be invalid even for a small dataset. We hope that our work calls attention to the (in)validity of common assumptions in building political media bias datasets.

Introduction

In today's world, bias and polarization are some of the biggest problems plaguing our society. In such volatile environments, news media play a crucial role as the gatekeepers of the information. Given the huge impact they can have on societal evolution, they have long been studied by researchers. Researchers and practitioners often build political bias datasets for a variety of tasks ranging from examining bias of media outlets and news articles to studying and designing algorithmic news retrieval systems for a healthy news diet.

The fundamental task at the core of building political bias datasets is '*inferring the political leaning of individual news outlets or articles.*' Inferring the political leaning by human experts, however, is not scalable, and thus researchers often label fewer items and then expand these labels to the whole dataset. One such widely-used practice is collecting media-level bias labels and then using the same to annotate article-level bias. For example, in this case, all the news articles from Fox News would be considered right-leaning as Fox News is the right-leaning news media. While such 'propagation' of media-level leaning to article-level may *not* always be correct, it is widely used without proper verification in the wild. Through a comprehensive review of previous work on building political media bias datasets, in this paper, we have compiled three common assumptions that are widely-used in practice and which might pose potential risks to the validity of the constructed political media bias datasets. The three assumptions are:

A1: **Raters' bias:** *Political leanings of raters do not affect their ratings of political content* (Gentzkow and Shapiro 2010; Iyyer et al. 2014; Bamman and Smith 2015).

A2: **Media-level bias and article-level bias:** *News articles follow the political leaning of their source outlet* (Potthast et al. 2018; Kulshrestha et al. 2018).

A3: **Topic-level bias:** *Political leanings of news outlets do not change while reporting on different topics* (Groseclose and Milyo 2005; Kulkarni et al. 2018; Bakshy, Messing, and Adamic 2015; An et al. 2011; An, Quercia, and Crowcroft 2013; Kulshrestha et al. 2017).

In this work, we empirically validate these three assumptions for building political media bias datasets. For this purpose, we collected news articles published by 18 U.S. news outlets of diverse political leaning on the topics of 'Gun policy' and 'Immigration' over a 3-month period in 2018. We then built a ground truth dataset of manually annotated article-level political leanings using Amazon Mechanical Turk (MTurk) and used it to validate the aforementioned assumptions.

Surprisingly, or not surprisingly, we discovered that (i) in certain cases, liberal and conservative raters label leanings of news articles differently, (ii) in many cases, the news articles do not follow the political leaning of the source outlet, and (iii) the political bias of news outlet while reporting on different topics does not remain unchanged. We hope that our work calls attention to the need for validating common assumptions used for building political media bias datasets.

Related Work

As manually detecting the political leaning of outlets or articles is challenging at large scale, significant research effort has focused on computationally inferring the political leanings of texts documents such as articles, blogs, political statements, social media posts and congressional speeches (Gentzkow and Shapiro 2010; Sim et al. 2013; Iyyer et al. 2014; Preoţiuc-Pietro et al. 2017; Bamman and Smith 2015; Kulshrestha et al. 2018; Kulkarni et al. 2018;

Potthast et al. 2018). However, since inferring the political leanings in an automated, scalable manner is a hard task, researchers often assume certain conditions that make the problem simpler.

The first category of prior studies assumes that the political leaning of raters does not affect their ratings of political content. Gentzkow and Shapiro (Gentzkow and Shapiro 2010) validate their computed slant indices of outlets against reader-submitted ratings of outlet slant from media directory website Mondo Times. Mondo Times did not take into account the political leaning of readers while aggregating their ratings. Similarly, Iyyer et al. (Iyyer et al. 2014) do not fully consider the leanings of Crowdflower workers while collecting ideological labels for the text from IBC dataset (Gross et al. 2013). As do Bamman and Smith (Bamman and Smith 2015) while collecting labels for political beliefs for their dataset of assertions by showing them to MTurk workers.

The second set of studies assume that the bias of articles follows the bias of the outlet publishing it. For instance, Kulshrestha et al. (Kulshrestha et al. 2018) label tweets with the political leaning of the authors of the tweet, Kulkarni et al. (Kulkarni et al. 2018) label each article in their dataset with the label assigned to its source outlet by *Allsides.com*, and Potthast et al. (Potthast et al. 2018) assume that all articles published by outlets with different orientations (mainstream, left-wing, or right-wing) reflect the orientation of the publisher.

Finally, many, if not most, studies (Gentzkow and Shapiro 2010; Groseclose and Milyo 2005; Kulkarni et al. 2018; Bakshy, Messing, and Adamic 2015; An et al. 2011; An, Quercia, and Crowcroft 2013; Kulshrestha et al. 2017) assume that the political leanings of news outlets are stable while reporting on different topics.

In this work, we empirically validate these assumptions for the task of building political bias datasets, on our manually annotated ground-truth dataset of article-level political leanings.

Building Ground-Truth Dataset

Collecting news articles

We collected 114,218 news articles published by 18 U.S. news media outlets, between 26 June and 25 September 2018. 13 of these outlets account for the most percentage of weekly usage during 2017 (Newman et al. 2017). We also gathered their political leanings reported by *Allsides.com* along five bias categories – strongly liberal (SL), moderately liberal (ML), neutral (N), moderately conservative (MC), and strongly conservative (SC).

Furthermore, to curate a more balanced set of outlets, we additionally included two strongly conservative, two moderately conservative, and one strongly liberal news outlet, giving us a total of 18 media outlets. Figure 1 depicts these 18 outlets across the political spectrum on the x-axis (as reported by *allsides.com*) ranked from SL to SC.

Obtaining ground truth bias labels for articles

We randomly selected an equal number of articles from the outlets in each bias categories, giving us a total of 460 articles (230 articles each on the topics of 'Gun policy' and 'Immigration' as determined by keywords search). As MTurk has been widely used to label the dataset in this field, we also used MTurk to collect the political leaning of these 460 articles from 5 workers each. To ensure high quality of labels, we recruited workers who: (i) reside in the U.S., (ii) had successfully completed 1000 MTurk HIT's with at least 95% approval rating, and (iii) have successfully completed our political bias qualification test. We paid the workers $0.10 for each article they reviewed because each news article is long enough to take a few minutes to read.

During the survey, the MTurk workers were shown the headline and body text of an article and were asked, "What is the political leaning of the article?". They were shown answer choices on a five-point scale, between strongly liberal (-2.0) and strongly conservative (+2.0). We computed the political leaning score of an article as the average score of the five workers, following the previous literature (Budak, Goel, and Rao 2016). Splitting the range of political leaning score, we obtained the political leaning label for each article within the five bias categories (SL, ML, N, MC, SC). The political leaning score ranges from -2.0 to 2.0, with [-2.0,-1.2] for Strongly Liberal (SL), (-1.2,-0.4] for Moderately Liberal (ML), (-0.4, 0.4) for Neutral (N), [0.4, 1.2) for Moderately Conservative (MC) and [1.2,2.0] for Strongly Conservative (SC). Table 1 shows the distribution of the articles across the five bias categories and the two topics.

Leaning	SL	ML	N	MC	SC	Total
Gun policy	24	90	78	33	5	230
Immigration	23	81	65	47	14	230

Table 1: Distribution of the articles on two topics along ground truth bias labels annotated by MTurk workers

Inter-rater agreement

When we investigate the 460 articles labeled by 5 workers each, we observe a high agreement among the workers. On average, the political leaning reported by the workers varied by less than one point (0.9) on a five-point scale, with only 5% cases with workers rating the article on opposite sides of the political spectrum (i.e., one rater labeled it as liberal-leaning while the other labeled it as conservative-leaning).

Manual Validation

We manually examined whether it is reasonable to use the average score of the five workers as political bias score of the article. For this analysis, we randomly selected 10 articles for each topic and two authors independently labelled those 20 articles on the five-point bias scale. We then compared the author's labels (average score of the two authors' labels) with Mturk worker's labels. On average, the political leaning reported by the Mturk workers differed by 0.525 from the author's labels, with zero cases where the two labels were on opposite sides of the political spectrum. Our analysis validates that there is high agreement in MTurk workers and authors' assessment of the political leaning of articles.

Publicly sharing dataset

To encourage further research on inferring political leanings of articles and news outlets, we make our dataset along with publisher bias and ground truth labels judged by MTurk workers publicly available.[1]

Experimental Evaluation

In this section, we evaluate the three assumptions by statistical tests and prediction models.

A1: Raters' political leanings do not affect their ratings.

To evaluate the first assumption, we selected 20 random articles (four random articles by outlets in each of the five bias categories) for each topic, giving a total of 40 articles. We recruited MTurk workers by their political affiliation[2], while also ensuring that they satisfy the three conditions from Section . For each article, we obtained political leaning annotations (between +2.0 and -2.0) from 5 conservative and 5 liberal MTurk workers.

To evaluate whether the conservative and liberal workers label the political leaning of articles similarly or not, we performed paired t-test over the ratings of liberal and conservative workers, for these 40 articles. In this case, our null hypothesis is – *the difference between the political leaning labels of the same news article by the liberals and conservatives is zero*.

Our paired t-tests show inconsistent trends; while we can reject our null hypothesis for the 'Gun policy' articles (t = 2.19, p-value = 0.03), we fail to reject the null hypothesis for 'Immigration' (t = 0.45, p-value = 0.65) and 'All' articles (t = 1.85, p-value = 0.06) at the significance level of 0.05. In other words, we do observe a statistical difference between the labels of news articles made by liberals and by conservatives on the topic of 'Gun policy', but this difference is not significant for the topic of 'Immigration.' Therefore, our results suggest that this widely held assumption is not always valid.

This finding, however, does not mean that all the labels obtained by crowdsourcing without consideration of labelers' political leaning are incorrect. Although the liberals and conservatives label the articles differently, no one side is always correct. It is thus reasonable to expect that averaging the ratings from both sides would reduce this bias, which is what most studies in the past have done. In other words, we do not expect that our findings affect the correctness of previous work. Rather, based on our findings, we would like to caution the researchers that there can be scenarios where these labeling differences may play a significant role. For example, this might be an issue when the rater population is skewed (e.g., labeling volunteers are students from a university in a strongly liberal region). In this case, researchers may need to account for it by either explicitly selecting a balanced set of raters, or by formulating survey questions that mitigate the differences in the responses of raters based on their political belief (Budak, Goel, and Rao 2016). Similarly, the extent to which a topic is polarizing may determine how much care needs to be paid to methods for mitigating the impact of labeling differences between liberals and conservatives.

[1] https://github.com/soumenganguly/icwsm-20-media-bias

[2] This requires additional charges in MTurk.

A2: News articles follow the political leaning of their source outlet.

To give an idea of how the political leaning of the news articles can be different from those of their source outlets, we first simply count the number of articles for which the media-level and article-level political leanings differ. Using three bias categories (Liberal, Neutral, and Conservative), we find that for 58.3% (134 out of 230) and 50.9% (117 / 230) of the articles on Gun policy and Immigration, respectively, the article-level leanings do not match their media-level political leanings. As expected, the differences become larger when using five bias categories–73.9% and 74.8% of the Gun policy and Immigration articles have different political leanings from their media-level political leanings.

To evaluate the second assumption, we consider a typical use-case scenario of building a model for predicting the political leaning of a news article using the media bias dataset. We want to investigate if we use a dataset following the second assumption, how does the performance of the prediction model change?.

We build and compare the performance of two models, which are (i) M_g: model trained using articles labeled with ground-truth political leaning, and (ii) M_a: model trained using articles labeled with source outlet's political leaning.

Immigration		
Model	**Classifier**	**H,T and P**
M_g	MNB_g	0.36 ± 0.03
	SVM_g	$\mathbf{0.40 \pm 0.12}$
	LR_g	0.38 ± 0.07
	RF_g	0.37 ± 0.08
M_a	MNB_a	0.21 ± 0.01
	SVM_a	0.22 ± 0.01
	LR_a	$\mathbf{0.23 \pm 0.01}$
	RF_a	$\mathbf{0.23 \pm 0.01}$
Gun policy		
Model	**Classifier**	**H,T and P**
M_g	MNB_g	$\mathbf{0.39 \pm 0.04}$
	SVM_g	0.35 ± 0.11
	LR_g	0.37 ± 0.10
	RF_g	$\mathbf{0.39 \pm 0.08}$
M_a	MNB_a	0.19 ± 0.01
	SVM_a	0.23 ± 0.01
	LR_a	$\mathbf{0.25 \pm 0.01}$
	RF_a	0.24 ± 0.01

Table 2: Average F1-scores and 90% confidence intervals across the 50 runs

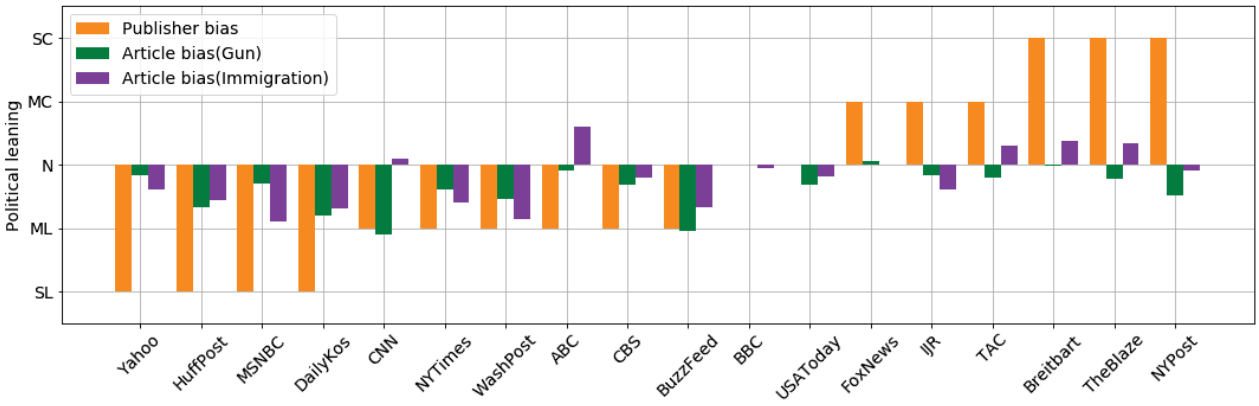

Figure 1: Political leaning of media outlets: Publisher bias (determined by Allsides.com) & aggregate bias of outlet's articles as judged by MTurkers. Outlets are ranked on x-axis by their publisher bias from SL to SC.

We apply four types of supervised learning classifiers for our models M_g and M_a – Multinomial Naive Bayes (MNB), Support Vector Machines (SVM), Logistic Regression (LR) and Random Forest (RF). We train each classifier using three categories of features: (i) text of the headline (H), (ii) text (T) of the article, and (iii) political leaning of the source outlet (P) of the article. Additionally, to account for the lack of balance in our dataset, we utilize class weights for adjusting the balance of the training set.

For training our models, we use 5-fold cross-validation, where the dataset is partitioned into 5 samples, out of which 4 are used for training and the remaining one is used for testing. Since each sample is used exactly once for testing, we produce 5 results. The entire 5-fold cross-validation process is repeated 10 times with different seeds for shuffling and partitioning the whole dataset. We report the average F1-scores and 90% confidence intervals across these 50 runs for the 230 articles on the topics of 'Immigration' and 'Gun policy' in Table 2.

Comparing the performance of the two models in Table 2, we observe that M_g, which is trained with ground truth data, performs 56-74% better than M_a, which is trained with the dataset following **A2**. Even though it is expected that the gap of the performance between M_g and M_a with larger data would decrease, the potentially huge impact of the dataset built based on **A2** should be carefully understood.

A3: Political leanings of news outlets do not change across topics.

Finally, to evaluate the last assumption, we split the data into two sets of 230 articles each on the topics 'Gun policy' and 'Immigration.' For each topic, we compute the average political bias score for each outlet based on the bias scores of the articles (as determined by MTurk workers) published by that outlet.

In Figure 1, we present the bias scores for each outlet, averaged across the articles on each topic. Here, the outlets are ranked on the x-axis from strongly liberal to strongly conservative according to the outlet-level bias determined by *Allsides.com*, while the y-axis depicts the political leaning scores. We observe that several outlets have differing political bias scores for the articles on the two topics, as shown in the figure. Moreover, for some media outlets, their political leaning scores of the two topics are on opposite sides of the political spectrum. As an example, while The Blaze is considered strongly conservative, it shows opposing biases for articles on Gun policy and Immigration. Similarly, while ABC is considered a neutral media, the aggregate bias of its Immigration-related articles is conservative and that of Gun policy-related articles is liberal. Overall, our results demonstrate that assuming that the political leaning of outlet remains unchanged across the topics is not always correct; particularly, when the dataset covers multiple topics.

Concluding Discussion

In this paper, we empirically evaluated three common assumptions for building political bias datasets. For doing so, we constructed a manually annotated dataset of news articles and their political leaning labels. Our experimental evaluation reveals that these widely held assumptions do not always hold true. Particularly, we found that: (i) in certain cases, the political leaning of the rater can affect their rating of political leaning of the news article, (ii) the political leaning of news articles does not always follow the leaning of the publisher, and (iii) the political leaning of the publisher sometimes changes while reporting on different topics.

While we tried to sample articles from a larger dataset from multiple outlets with different political leanings and handled class imbalance by weights, one limitation of this work is the small size of the ground truth dataset. It is mainly because of the cost of recruiting MTurk workers with additional constraints (e.g., the political leaning of workers). We hope that our work will inspire further larger-scale investigation of these common assumptions based on bigger datasets. Also, self-reported political leanings of MTurk workers could potentially introduce bias. It, however, is not ad hoc and is repeatedly confirmed by their participation in other tasks.

We believe that our findings will guide researchers and practitioners to (i) make more informed assumptions while building manually-annotated political bias datasets, and (ii) be more aware of the potential biases of existing datasets built without carefully considering the validity of assumptions made.

References

An, J.; Cha, M.; Gummadi, K.; and Crowcroft, J. 2011. Media landscape in twitter: A world of new conventions and political diversity. In *Fifth International AAAI Conference on Weblogs and Social Media*.

An, J.; Quercia, D.; and Crowcroft, J. 2013. Fragmented social media: a look into selective exposure to political news. In *Proceedings of the 22nd International Conference on World Wide Web*, 51–52. ACM.

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

Bamman, D., and Smith, N. A. 2015. Open extraction of fine-grained political statements. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 76–85.

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.

Gentzkow, M., and Shapiro, J. M. 2010. What drives media slant? evidence from us daily newspapers. *Econometrica* 78(1):35–71.

Groseclose, T., and Milyo, J. 2005. A measure of media bias. *The Quarterly Journal of Economics* 120(4):1191–1237.

Gross, J. H.; Acree, B.; Sim, Y.; and Smith, N. A. 2013. Testing the etch-a-sketch hypothesis: a computational analysis of mitt romney's ideological makeover during the 2012 primary vs. general elections. In *APSA 2013 Annual Meeting Paper*.

Iyyer, M.; Enns, P.; Boyd-Graber, J.; and Resnik, P. 2014. Political ideology detection using recursive neural networks. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, 1113–1122.

Kulkarni, V.; Ye, J.; Skiena, S.; and Wang, W. Y. 2018. Multi-view models for political ideology detection of news articles. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 3518–3527. Brussels, Belgium: Association for Computational Linguistics.

Kulshrestha, J.; Eslami, M.; Messias, J.; Zafar, M. B.; Ghosh, S.; Gummadi, K. P.; and Karahalios, K. 2017. Quantifying search bias: Investigating sources of bias for political searches in social media. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, 417–432. ACM.

Kulshrestha, J.; Eslami, M.; Messias, J.; Zafar, M. B.; Ghosh, S.; Gummadi, K. P.; and Karahalios, K. 2018. Search bias quantification: investigating political bias in social media and web search. *Information Retrieval Journal* 1–40.

Newman, N.; Fletcher, R.; Kalogeropoulos, A.; Levy, D.; and Nielsen, R. K. 2017. Reuters institute digital news report 2017.

Potthast, M.; Kiesel, J.; Reinartz, K.; Bevendorff, J.; and Stein, B. 2018. A stylometric inquiry into hyperpartisan and fake news. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 231–240. Melbourne, Australia: Association for Computational Linguistics.

Preoţiuc-Pietro, D.; Liu, Y.; Hopkins, D.; and Ungar, L. 2017. Beyond binary labels: political ideology prediction of twitter users. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 729–740.

Sim, Y.; Acree, B. D.; Gross, J. H.; and Smith, N. A. 2013. Measuring ideological proportions in political speeches. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 91–101.