Proceedings of the Sixteenth International AAAI Conference on Web and Social Media (ICWSM 2022)

Measuring Media Bias via Masked Language Modeling

Xiaobo Guo, Weicheng Ma, Soroush Vosoughi

Department of Computer Science, Dartmouth College Hanover, New Hampshire
{xiaobo.guo.gr, weicheng.ma.gr, soroush.vosoughi}@dartmouth.edu

Abstract

Bias in news reporting can lead to tribalism and division on important issues. Scalable and reliable measurement of such biases is an important first step in addressing them. In this work, based on the intuition that media bias is captured by the tone and word choices in articles, we propose a framework for modeling the relative bias of media outlets through masked token prediction via large-scale pretrained masked language models fine-tuned on articles form news outlets. Through experiments on five diverse and politically polarized topics we show that our framework can capture media bias towards these topics with high reliability. Additionally, our experiments show that our framework is general, in that language models fine-tuned on one topic can be applied to other topics with little drop in performance.

Introduction

Partisanship in news outlets has been shown to heavily sway public opinion (Eveland Jr and Shah 2003; Heimlich 2011; Morales 2011). Such bias operates via two mechanisms: selective coverage of issues, known as *issue filtering*, and presentation of issues, known as *issue framing* (Budak, Goel, and Rao 2016). Compared to issue filtering, bias via issue framing is more nuanced and difficult to identify. Furthermore, framing bias falls under different categories: bias relative to "neutral" opinion, median preference of citizens, and other media outlets (Puglisi and Snyder Jr 2015). Among these categories, bias relative to other media outlets is one that can be objectively quantified. Thus, in this paper, we focus on measuring the relative bias of media outlets with respect to issue framing.

Most prior work examining media bias use qualitative methods, however these methods are subjective, expensive and cannot be easily reproduced. Others have used quantitative methods, specifically audience-based and content-based approaches. The audience-based method is premised on the assumption that readers prefer news outlets closer to their ideology (Mullainathan and Shleifer 2005). Although this method has produced an ideological ordering of outlets (Zhou, Resnick, and Mei 2011; Ribeiro et al. 2018; Gentzkow and Shapiro 2011; Mitchell 2016; Spinde et al. 2021a), it relies on readership surveys which introduces its own subjective bias, is costly and not easily scalable.

Content-based methods on the other hand measure media bias directly from the published content. These methods usually utilize hand-crafted features such as the tone or the choice of phrases when referring to the same events or entities (e.g., undocumented immigrant vs. illegal alien). These methods either rely fully on human annotation (Lim, Jatowt, and Yoshikawa 2018; Golez and Karapandza 2020; Färber et al. 2020; Gentzkow and Shapiro 2010; Budak, Goel, and Rao 2016; Fan et al. 2019; Spinde et al. 2021c)– which makes them not scalable and prone to human bias– or utilize (semi-) automatic methods by representing the problem as a standard supervised learning task (Spinde, Hamborg, and Gipp 2020a,b; Spinde et al. 2021b; Chen et al. 2020), relying on topic-specific hand-crafted features.

To address these challenges, we utilize language models (LMs) to capture the opinions "hidden" in news articles. Specifically, we capture the tone and word preference of media outlets by fine-tuning BERT models (Devlin et al. 2019) on corpora from different outlets and topics collected from Media Cloud, a publicly available news API (see Section for more detail).

Through the masked language modeling (MLM) objective, we then use these fine-tuned models to predict the choice of word by individual media outlets for different contexts, such as different issues or events. In other words, prompting these models with sentences such as "the greatest threat to the immigration system is ___", they will output the word most likely to complete the sentence given the associations learned from their fine-tuning texts. Comparing outputs from models fine-tuned on different outlets can thus illuminate key attitudinal differences between. The prompts used in this paper come directly from the development set of our corpora.

To validate the relative bias between outlets as reported by our method, we compare the results to three independent news bias datasets: two based on surveys of the US populace conducted by Pew Research (Jurkowitz et al. 2020), and another based on expert curation of bias in media outlets from the website Allsides.com. Our proposed framework for estimating relative bias between media outlets through prompting fine-tuned masked language models is scalable in that it eliminates the need for data annotation or hand-crafted fea-

tures.

Datasets

We collect articles from Media Cloud [1] to construct our dataset. Specifically, we sample news under five diverse topics: "Climate Change", "Corporate Tax", "Drug Policy", "Gay Marriage" and "The Affordable Care Act", from 10 news outlets in the US: Breitbart News Network, CBS News, CNN, Fox News, HuffPost, New York Times, NPR, USA Today, Wall Street Journal, and Washington Post.

We divide the dataset into train and development (dev) sets with a 90/10 split. Due to the length limit of BERT, we break the news articles down to paragraphs with no more than 256 words and label the paragraphs with their respective media outlets. The paragraphs are split at sentence level to ensure the completeness of sentences.

The dataset contains 107,121 instances for "Climate Change", 104,316 instances for "Corporate Tax", 104,188 instances for "Drug Policy", 109,486 instances for "Gay Marriage", and 106,287 instances for "The Affordable Care Act". Each instance is one paragraph. In Table 1, we list the statistics of the train and dev sets for each topic and media outlet.

Additionally, we construct three ground truth datasets for evaluation purposes. Two of the datasets are based on survey data from Pew Research (Jurkowitz et al. 2020). These datasets are annotated on a 6-point scale based on the survey results regarding how much the respondents (1) trust a particular new outlet, and (2) use the outlet for political news. We call these datasets SoA-t and SoA-s, capturing the share of Americans who trust and receives news from each outlet respectively. The third dataset (called MBR) is based on expert curation of media bias by the wesbite Allsides.com [2]. This dataset labels each media outlet with five political leanings, from left to right, based on editorial review, third-party analysis, independent review, surveys and community feedback.

Methodology

As shown in Figure 1, our framework is comprised of three steps: (1) We first fine-tune an LM for each media outlet (2) We then leverage the fine-tuned LMs to create attitudinal representations of the media outlets via prompt-based mask token prediction (3) We utilize the generated representations to measure the relative bias of the outlets.

Language Model Fine-tuning

We fine-tune bert-base-cased (henceforth referred to as BERT) using the MLM task as done in its pre-training stage. While we leverage BERT in our experiments because of its popularity and high performance, any MLM-based language model can potentially be utilized.

[1] https://mediacloud.org/

[2] https://www.allsides.com/media-bias/media-bias-ratings

Media Attitude Representation

To generate the attitudinal representation of each media outlet, we use prompt-based mask token prediction in which we mask one word in a selected prompt and apply the fine-tuned LMs to predict it.

Inspired by work on authorship attribution (e.g., (Coyotl-Morales et al. 2006)), we pick the prompts and the word to be masked using the following approach: (1) We first create a list of bigrams that appear in the dev set of all ten media outlets (219 total) (2) For each instance in dev set containing these bigrams (11,500 total) we generate two masked prompts (per bigram), one where the mask is applied to the word preceding the bigram and one with the mask applied to the word following the bigram. We experiment with other methods to pick the prompts and the tokens to be masked, and the method described here was the best performing in our small-scale experiments.

We then use the fine-tuned LMs for each outlet to predict the masked token in our prompts. We retrieve the the top-10 candidate words with the highest probability. The attitude of each media outlet with respect to the masked prompt is then represented as a vector of the probability of these words. Note that the vectors for the outlets all have the same length and correspond to the same words, i.e., the union of all the top-10 candidate words from the outlets (if a word was not in the top-10 candidate of an outlet, its probability is set to 0). This allows for cross media comparisons. In Figure 2, we show an example of how the media attitudinal representations are generated for two media outlets using the top-3 candidate words (for simplicity).

At the end, for each topic t we have n_t vectors for each of the 10 media outlets, where n_t is the number of masked prompts for each topic. The set of these vectors represent the attitude of each media outlet with respect to different topics. In Table 2, we show the number of bigrams, instances in the dev set, and masked prompts (note that each instance produces one or more masked prompts) for each topic for the bigram outer method. We observe that other than the "The Affordable Care Act" all topics have a similar data size.

Measuring Relative Bias

We measure the relative bias of the media outlets for specific topics by first calculating the distance between each pair of outlets. This is done by calculating the mean of the euclidean distance across all aligned (by prompt) attitudinal representation vectors for the specified topics for the outlet pair.

Next, for each outlet we create a ranking of other media outlets based on their distance. Note that these ranking do not have to be symmetric, in that if outlet A has outlet B as the closest outlet, it doesn't necessarily follow that B will have A as its closest outlet. These rankings corresponding to relative attitudinal bias of each outlet with respect to other outlets. Outlets that are ranked closer to each other should have more similar attitudes towards the specified topics and vice versa.

	Climate Change		Corporate Tax		Drug Policy		Gay Marriage		The Affordable Care Act	
	Train	Dev	Train	Dev	Train	Dev	Train	Dev	Train	Dev
Breitbart	8,267	807	8,357	756	7,765	934	7,663	1,016	6,665	732
CBS	11,515	1,680	9,002	1,026	9,313	1,273	12,320	2,268	7,968	799
CNN	12,587	1,414	12,982	1,529	13,158	1,565	13,949	1,841	16,030	1,805
Fox	8,487	450	8,526	1,284	9,445	602	6,473	695	7,272	735
HuffPost	10,272	1,102	11,044	1,087	9,780	1,117	9,385	1,054	10,138	1,132
NPR	15,260	1,509	14,730	1,997	14,934	1,503	17,285	2,036	13,057	1,281
NYTimes	3,113	322	2,678	297	3,077	271	3,803	289	4,936	452
USA Today	12,288	1,842	13,464	1,476	12,432	1,718	12,436	1,340	12,261	1,208
Wallstreet	4,914	447	4,040	433	4,403	469	6,469	635	3,003	407
Washington	9,747	1,098	8,788	816	9,545	911	7,666	861	14,459	1,947

Table 1: The statistics of the train and dev sets for each topic and media outlet."NYTimes" : "New York Times", "Wallstreet" : "The Wall Street Journal", "Washington" : "Washington Post"

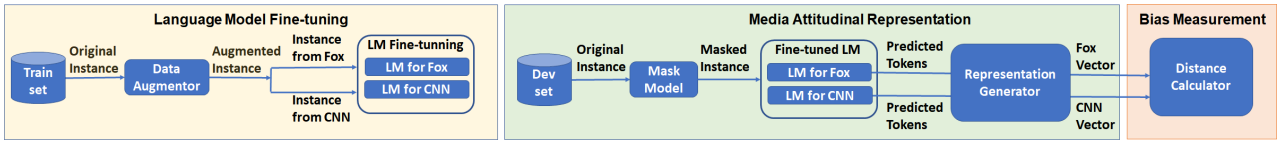

Figure 1: An overview of our framework. For simplicity, only two sample media outlets are shown here.

	Climate	Cor	Drug	SSM	Care	Total
# of bigrams	39	37	37	31	75	219
# of instances	1,856	2,067	1,992	1,514	4,071	11,500
# of masked prompts	4,299	5,152	5,161	3,550	13,160	31,322

Table 2: The statistics of the number of bigrams, instances in the dev set, and masked prompts for the BO prompt and mask selection method. "Climate" : "Climate Change", "Cor" : "Corporate Tax", "Drug" : "Drug policy", "SSM" : "Gay Marriage", "Care" : "The Affordable Care Act".

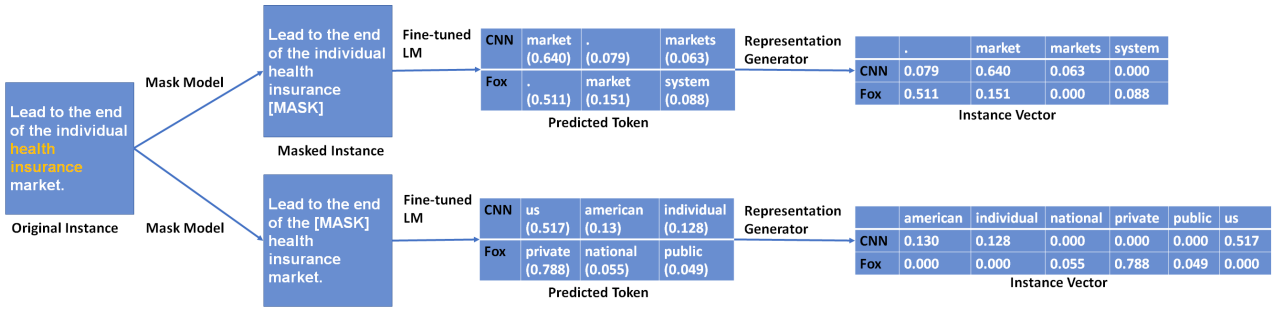

Figure 2: An example generation of media attitudinal representations for two outlets. The bigram is "health insurance" (shown in orange), and we mask the tokens "market" and "individual" separately to generate two masked instance. For simplicity, we generate the representations using the top-3 predicted words.

	In-domain Performance					Out-of-domain Performance				
	Climate	Cor	Drug	SSM	Care	Climate	Cor	Drug	SSM	Care
SoA-t	0.24	0.35	0.31	0.31	0.35	0.33 (0.08)	0.25 (0.07)	0.29 (0.07)	0.24 (0.05)	0.25 (0.01)
SoA-s	0.28	0.28	0.31	0.30	0.33	0.32 (0.08)	0.21 (0.05)	0.28 (0.07)	0.23 (0.06)	0.26 (0.02)
MBR	0.27	0.26	0.44	0.30	0.41	0.37 (0.07)	0.21 (0.03)	0.35 (0.09)	0.26 (0.05)	0.30 (0.05)

Table 3: Agreement between media similarity ranking using our framework and ground truth datasets. Agreement is calculated using Kendall's τ (higher magnitude is better). "Climate" : "Climate Change", "Cor" : "Corporate Tax", "Drug" : "Drug policy", "SSM" : "Gay Marriage", "Care" : "The Affordable Care Act". The out-of-domain results show the average performance of models trained on 4 different topics and tested on the remaining topic. The standard deviations are shown in the brackets.

Experiments and Analysis

To validate our method for estimating the relative attitudinal bias of news outlets, we run two sets of experiments: the first experiment trains and evaluates the relative bias of outlets on the same topic. while the other trains the model on one topic and evaluates it on other topics to assess the generalizability of the trained model. The two settings are respectively referred to as in-domain and out-of-domain media bias estimation tasks.

We compare the relative attitudinal bias of the outlets predicted by our framework with ground truth labels in the SoA-s, SoA-t, and MBR datasets. Specifically, we measure the relative bias of the outlets using these ground truth datasets using a similar procedure as described in Section : for each outlet we calculate the distance to the other outlets (based on their ground truth political ideology labels) and create a ranked list. For SoA-s and SoA-t we use cosine distance (since these datasets provide a distribution over ideologies for each outlet), while for MBR we look at the absolute distance between the outlets (since this dataset provides a single ideological score for each outlet).

We then calculate the similarity between the predicted and ground truth similarity rankings via Kendall rank correlation coefficients (Kendall's τ). As a Kendall's τ is calculated for each media outlet; we use the mean of all the τ's as the evaluation score. Note that τ ranges from -1 to 1 corresponding to perfect misalignment and alignment of ranked lists respectively. A τ of 0 corresponds to completely random alignment.

In Table 3, we show the in-domain and out-of-domain relative media bias estimation evaluations using the three ground truth datasets. We observe that the agreements as reported by τ are all significantly above random chance ($p < 0.05$ for all), with the 'The Affordable Care Act" dataset achieving the highest agreement in in-domain experiments (likely due to the extremely polarized conversation around this topic). As expected, the out-of-domain experiments on average have lower performance, but the drop is minimal and all results are still significantly above random chance. This results show that our framework can capture both the topic-specific and general attitudinal bias of outlets.

We also compare the performance of our models against two classic baseline methods used in prior media bias estimation experiments, namely models based on Latent Dirichlet allocation (LDA) and Term Frequency-Inverse Document Frequency (TF-IDF) features. Both baseline models are fine-tuned on the train set for deciding the hyper-parameters and encode the instances in the dev set[3]. The encodings are then used to calculate cosine distances between pairs of media outlets in a manner similar to what was done in Section . Same as before, for each outlet we use the distance to other outlets to create a ranked list and use Kendall's τ to measure agreement with the ground truth rankings.

We show the mean τ across all topics in Table 4. Our model outperforms the baselines on all three ground truth datasets, with the differences all being statistically significant ($p < 0.05$).

[3]We treat each instance as one document.

	SoA-t	SoA-s	MBR
LDA	0.15(0.08)	0.16(0.09)	0.22(0.11)
TF-IDF	0.23(0.07)	0.26(0.08)	0.29(0.07)
Ours	**0.31(0.05)**	**0.30(0.02)**	**0.34(0.08)**

Table 4: Comparison of agreement between ground truth datasets and the baselines and our method. Agreement is calculated using Kendall's τ. Results are averaged across 5 topics and the standard deviations are reported in the brackets. $p < 0.05$ for all values.

Conclusion and Future Work

We presented a framework for capturing attitudinal bias of media outlets using the MLM objective through masked prompting of fine-tuned large-scale pretrained language models (BERT specifically). Our experiments show that the predicted relative bias of outlets match 3 different ground truth datasets. Future work can explore other strategies for prompt and masked token selection.

Code & Data Availability: We will make the code and data publicly available here. [4]

Ethics Statement

This paper proposes a novel method for studying the relative bias of news outlets with respect to different issues using masked language modeling. Though the issue of bias in news is controversial, this paper only studies the *relative* bias between outlets. We do not make any judgements as to whether certain outlets are biased, our method only reports the relative difference between the outlets. All text used in this paper come from public news outlets and were collected using the publicly available API of Media Cloud. As such, the data does not contains any private information. Since we use mainstream news outlets for our data collection we believe there is less risk of overtly unethical information (though we cannot be sure given the current sociopolitical climate). Given the relatively large size of our dataset, we cannot manually examine all articles, however, the publicly released dataset will warn users of the possibility of the dataset containing unethical information and will allows users to flag unethical articles in our dataset.

Finally, as we use pre-trained language models in our paper, we must be aware of the inherent bias in such models (based on their pre-training data) and again err on the side of caution when utilizing such models for consequential applications.

References

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.

Chen, W.-F.; Al Khatib, K.; Wachsmuth, H.; and Stein, B. 2020. Analyzing Political Bias and Unfairness in News Articles at Different Levels of Granularity. In *Proceedings of*

[4]https://github.com/guoxiaobo96/media-bias

the Fourth Workshop on Natural Language Processing and Computational Social Science, 149–154.

Coyotl-Morales, R. M.; Villaseñor-Pineda, L.; Montes-y Gómez, M.; and Rosso, P. 2006. Authorship attribution using word sequences. In *Iberoamerican Congress on Pattern Recognition*, 844–853. Springer.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186.

Eveland Jr, W. P.; and Shah, D. V. 2003. The impact of individual and interpersonal factors on perceived news media bias. *Political psychology*, 24(1): 101–117.

Fan, L.; White, M.; Sharma, E.; Su, R.; Choubey, P. K.; Huang, R.; and Wang, L. 2019. In Plain Sight: Media Bias Through the Lens of Factual Reporting. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 6343–6349.

Färber, M.; Burkard, V.; Jatowt, A.; and Lim, S. 2020. A Multidimensional Dataset Based on Crowdsourcing for Analyzing and Detecting News Bias. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 3007–3014.

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

Gentzkow, M.; and Shapiro, J. M. 2011. Ideological segregation online and offline. *The Quarterly Journal of Economics*, 126(4): 1799–1839.

Golez, B.; and Karapandza, R. 2020. Home-country media slant and equity prices. *Available at SSRN 3396726*.

Heimlich, R. 2011. Press widely criticized, but trusted more than other information sources. *Pew Research Center*.

Jurkowitz, M.; Mitchel, A.; Shearer, E.; and Walker, M. 2020. U.S. Media Polarization and the 2020 Election: A Nation Divided. *Pew Research Center*.

Lim, S.; Jatowt, A.; and Yoshikawa, M. 2018. Understanding Characteristics of Biased Sentences in News Articles. In *CIKM workshops*.

Mitchell, A. 2016. Key findings on the traits and habits of the modern news consumer. *Pew Research Center*.

Morales, L. 2011. Majority in US continues to distrust the media, perceive bias. *Gallop Politics*.

Mullainathan, S.; and Shleifer, A. 2005. The market for news. *American economic review*, 95(4): 1031–1053.

Puglisi, R.; and Snyder Jr, J. M. 2015. Empirical studies of media bias. In *Handbook of media economics*, volume 1, 647–667. Elsevier.

Ribeiro, F. N.; Henrique, L.; Benevenuto, F.; Chakraborty, A.; Kulshrestha, J.; Babaei, M.; and Gummadi, K. P. 2018. Media bias monitor: Quantifying biases of social media news outlets at large-scale. In *Twelfth international AAAI conference on web and social media*.

Spinde, T.; Hamborg, F.; and Gipp, B. 2020a. An integrated approach to detect media bias in German news articles. In *Proceedings of the ACM/IEEE joint conference on digital libraries in 2020*, 505–506.

Spinde, T.; Hamborg, F.; and Gipp, B. 2020b. Media bias in german news articles: A combined approach. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 581–590. Springer.

Spinde, T.; Hamborg, F.; Kreuter, C.; Gipp, B.; Gaissmaier, W.; and Giese, H. 2021a. How Can the Perception of Media Bias in News Articles Be Objectively Measured? Best Practices and Recommendations Using User Studies. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries (JCDL)*.

Spinde, T.; Rudnitckaia, L.; Mitrović, J.; Hamborg, F.; Granitzer, M.; Gipp, B.; and Donnay, K. 2021b. Automated identification of bias inducing words in news articles using linguistic and context-oriented features. *Information Processing & Management*, 58(3): 102505.

Spinde, T.; Rudnitckaia, L.; Sinha, K.; Hamborg, F.; Gipp, B.; and Donnay, K. 2021c. MBIC–A Media Bias Annotation Dataset Including Annotator Characteristics. *arXiv preprint arXiv:2105.11910*.

Zhou, D. X.; Resnick, P.; and Mei, Q. 2011. Classifying the political leaning of news articles and users from user votes. In *Fifth International AAAI Conference on Weblogs and Social Media*.