Intertwined Biases Across Social Media Spheres: Unpacking Correlations in Media Bias Dimensions

Yifan Liu$^{[0009-0003-6658-8089]}$, Yike Li$^{[0009-0001-9805-0692]}$, and Dong Wang$^{[0000-0002-9599-8023]}$

School of Information Sciences, University of Illinois Urbana-Champaign, Champaign IL 61820, USA

Abstract. Biased information on social media significantly influences public perception by reinforcing stereotypes and deepening societal divisions. Previous research has often isolated specific bias dimensions, such as *political* or *racial bias*, without considering their interrelationships across different domains. The dynamic nature of social media, with its shifting user behaviors and trends, further challenges the efficacy of existing benchmarks. Addressing these gaps, our research introduces a novel dataset derived from five years of YouTube comments, annotated for a wide range of biases including gender, race, politics, and hate speech. This dataset covers diverse areas such as politics, sports, healthcare, education, and entertainment, revealing complex bias interplays. Through detailed statistical analysis, we identify distinct bias expression patterns and intra-domain correlations, setting the stage for developing systems that detect multiple biases concurrently. Our work enhances media bias identification and contributes to the creation of tools for fairer social media consumption.

Keywords: Social Media · Bias Identification · Benchmark · Datasets

1 Introduction

In our digital era, the widespread distribution of information across social media often includes user-generated content that can perpetuate stereotypes, discrimination, and hatred. We categorize such content as a form of media bias [22], emphasizing its significant influence on these platforms. Beyond reinforcing existing social biases, online media bias interacts with cognitive biases [14], fostering information bubbles [16] that distort public perception and exacerbate social divisions. This highlights the urgency of developing robust systems for identifying online media bias to mitigate these effects. In our discussion, we specifically address online media bias arising from user-generated content, retaining the term 'media bias' for consistency. The concept of media bias, historically lacking a consensus definition, has recently been addressed through comprehensive literature reviews. Recent research proposes a unified definition that categorizes skewed portrayals into distinct dimensions of media bias [22], reflecting its complex nature. To counteract media bias, various frameworks have been developed to support automated identification efforts.

The identification of media bias has primarily been addressed by the machine learning (ML) and natural language processing (NLP) communities [19, 11]. In recent years, methods for identifying media bias have progressed from relying on hand-crafted features [18, 10] to the employment of advanced transformer-based models [17, 23, 27]. However, many existing research efforts still focus predominantly on detecting a *single* type of media bias, typically evaluating models against benchmarks that assess only one dimension of bias [25, 1]. Such a narrow focus presents significant challenges for developing comprehensive bias identification systems: 1) there is a disproportionate focus within the media bias detection community on different bias dimensions, resulting in a lack of high-quality benchmark datasets along some bias dimensions [26], and 2) without a thorough understanding of the various media bias dimensions, it is challenging to develop a bias identification system that is capable of jointly detecting and analyzing multiple bias dimensions.

To this end, we propose a new media bias identification benchmark that annotates and analyzes multiple dimensions of media bias across various topic domains. Drawing from prior work [26], we select specific bias dimensions tailored to our social media dataset. Our dataset comprises YouTube user comments collected over the last five years. Through rigorous statistical analysis, we find that the politics domain exhibits significantly higher proportions of biased content compared to other topics. Additionally, our analysis identifies domain-specific patterns in the expression of bias. For example, biased posts in politics often manipulate narratives to support specific agendas, while in sports, bias may manifest through specific word choices or jargon that convey prejudices. Additionally, temporal analysis shows that the correlations between different bias dimensions are dynamic, fluctuating in response to spikes in discussion volume.

2 Related works

2.1 Media Bias Dimensions

Compared to previous studies on media bias, our work emphasizes the multi-dimensional aspects of media bias and the domain-specific occurrences of bias dimensions on social media platforms. Recent research has utilized the interrelationships between different types of biases as a foundation for developing more robust bias identification systems [26]. Specifically, this approach is applied within the multi-task learning (MTL) framework [4, 9], which provides a joint optimization framework for various media bias dimensions. It is important to note that both task selection and data selection play critical roles in the success of MTL [3, 20]. To effectively address such challenges, automated task-selection algorithms are considered to be a promising enhancement [12]. However, gradient-based automated task selection schemes, which rely on monitoring the training dynamics of different sub-tasks, are susceptible to discrepancies between data sources and could require a large number auxiliary tasks to perform effectively [9].

2.2 Comparison with Other Media Bias Datasets

Within the field of media bias research, various datasets are specifically designed to support the analysis of a single media bias type, employing multi-level labeling to capture its nuances. For example, RTGender dataset adopt a multi-categorical labeling to characterize different aspects of gender bias [24]. Similarly, prior research on political bias categorizes texts into five distinct political tenancies, ranging from *left* to *right* [1]. For more nuanced labeling, CMSB employs a continuous sexism scale to measure subtleties in Twitter data[21]. While such fine-grained categorization in each bias dimension provides a detailed view of media bias, our work primarily focuses on exploring inter-correlations among different bias dimensions.

Similar to our work, multidimensional bias dataset collects dataset with a bias dimension specification based on hidden assumptions, subjectivity and representation tendencies [7]. However, multidimensional bias dataset has a focus on news articles with a special emphasis on political tendencies, while our work focuses on the social media space with a wider range of topics. Recent media bias identification benchmark(MBIB) summarizes a rich list of publicly available datasets following a set of media bias dimension specification [26, 22]. Despite the well rounded bias dimensions discussed, MBIB provides a single label for each post, limiting the analysis of across-dimension analysis of media bias. Compared to other existing bias identification datasets, our dataset is the first to account for the joint occurrence of multiple bias dimensions with a joint labeling scheme. Additionally, we have intentionally segregated data collections across different domains using general keyword choices.

3 Dataset Creation

In this section, we elaborate on our dataset creation process, which includes the following steps: 1) Retrieval of domain-specific data; 2) Examination of bias dimensions leading to the development of our bias specification framework; and 3) Generation of multi-dimensional bias labels. Furthermore, we compare our dataset with existing datasets to underscore the research gap in the study of intertwined media biases. Our data collection process is designed to investigate dimensions of media bias within and across various domains, namely politics, healthcare, sports, entertainment and job & education.

To analyze media bias across social media, we collected comments from specific YouTube domains, using Google Trends to select representative keywords for each domain. For example, "COVID" was chosen for the healthcare sector due to its high search frequency over the past five years. The keywords we used are listed in Table 1. We employed YouTube's API to gather relevant posts from the last five years, collecting approximately 2,000 comments per domain. Our dataset includes only comments directly related to video topics, excluding replies to other comments to maintain focus on domain-specific content and avoid off-topic discussions. For preparing our data for media bias annotation, we filtered out non-English and overly long posts (over 200 words). Unlike previous studies, we converted emojis into text tokens to preserve the semantic content in our bias annotations, detailed further in Section 3.2.

Table 1. Data Collection by Domain

Domain	**Keywords**	**Comment Count**
Politics	election contest, election result, voting	1993
Healthcare	COVID, pain injury, symptom	2580
Sports	NBA, NFL, MLB	2398
Entertainment	film, lyrics, episodes	998
Job & Education	career, college, job	1455

3.1 Bias Dimension Specifications

In our analysis, we ground our investigation under the umbrella of media bias introduced in recent works [26, 22]. Our focus is on dimensions of media bias that are defined solely by post-level social media contents. In our exploration, we investigate a subset of media bias types summarized in prior work [26] including: linguistic bias [2], political bias [8], gender bias [26], hate speech [5, 13], racial bias [6] and text-level context bias [26]. We defer to the related works for the specific definitions of these bias dimensions.

3.2 Bias Annotation

To annotate the social media posts we collected, we employed a mixed approach combining manual and automated annotations. For each domain dataset, 100 samples were annotated by two annotators. Each annotator was responsible for annotating 60 samples, with 20 samples overlapping between the two annotators for consistency checks. In table 2, we report the performance of automated annotation, together with the inter-rater agreement scores (Cohen's κ) for each domain. Across all bias dimensions, we observe substantial inter-rater agreement (Cohen's $\kappa > 0.8$), with the conflicting annotations being reviewed and inspected. Building on the manual annotations, we evaluated two groups of automated annotations: 1) a shallow pretrained models trained on existing benchmark bias evaluation datasets; and 2) zero/few-shot annotations from pretrained large language model [15]. Our results indicate that while smaller-scale transformer baselines achieve good robustness and generalizability with cross-validated train-test splits, they exhibit poor generalization when applied to the noisy social media posts we collected along some bias dimensions. Overall, we utilize the predictions generated by the best performing considering both shallow models and LLMs evaluated on our manually annotated test set in our further analysis of bias dimensions.

Table 2. Weighted F1-Scores of Automated Annotation & Inter-Rater Agreement for Bias Dimensions (HS: Hate Speech, PB: Political Bias, GB: Gender Bias, RB: Racial Bias, LB: Linguistic Bias, TLCB: Text-level Context Bias)

Bias Dimension	**Politics**	**Sports**	**Healthcare**	**Job& Education**	**Entertainment**	**Model**	**Cohen's κ Score**
GB	0.90	0.99	0.99	0.97	0.88	GPT-Turbo-3.5	1.00
RB	0.97	0.98	1.00	0.93	0.96	GPT-Turbo-3.5	1.00
HS	0.77	0.85	0.92	0.95	0.91	Roberta-Twitter	0.82
LB	0.63	0.75	0.81	0.84	0.63	GPT-Turbo-3.5	0.93
TLCB	0.56	0.77	0.86	0.84	0.87	ConvBert	0.89
PB	0.64	1.00	0.94	0.98	1.00	GPT-Turbo-3.5	0.81

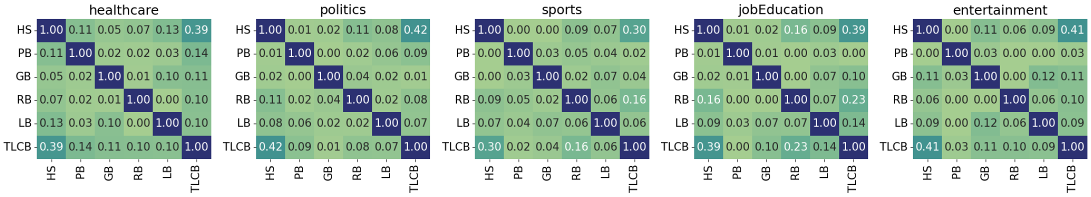
Table 3. Number of Biased Content Along Different Dimensions from Automated Annotation with Percentage of Total Posts Shown in Brackets (%)

Bias Dimension	**Politics**	**Sports**	**Healthcare**	**Job&Education**	**Entertainment**
Gender Bias	91 (4.61)	76 (3.21)	77 (3.09)	72 (5.15)	44 (4.53)
Racial Bias	99 (5.02)	39 (1.64)	30 (1.21)	35 (2.50)	14 (1.44)
Hate Speech	464 (23.54)	330 (13.91)	244 (9.82)	172 (12.29)	101 (10.40)
Linguistic Bias	499 (25.32)	501 (21.13)	382 (15.37)	239 (17.08)	192 (19.77)
Text-level Context Bias	480 (24.35)	204 (8.60)	247 (9.94)	123 (8.79)	56 (5.77)
Political Bias	578 (29.32)	35 (1.48)	60 (2.42)	18 (1.29)	3(0.31)
Total Posts	1971	2371	2484	1399	971

4 Experiments

4.1 Distribution Shift

For all annotated data, we report the number of samples and their percentage occurrences across various bias dimensions in Table 3. Notably, the politics domain exhibits a significantly higher proportion of biased content, especially in *text-level context bias* and *political bias*. Additionally, biased content in most domains primarily manifests as *linguistic biases*, characterized by discriminatory word usage. In the politics domain, biased content appears through both specific word choices (*linguistic bias*) and skewed descriptions (*text-level context bias*), occurring in similar proportions as detailed in Table 3.



healthcare

	HS	PB	GB	RB	LB	TLCB
HS	1.00	0.11	0.05	0.07	0.13	0.39
PB	0.11	1.00	0.02	0.02	0.03	0.14
GB	0.05	0.02	1.00	0.01	0.10	0.11
RB	0.07	0.02	0.01	1.00	0.00	0.10
LB	0.13	0.03	0.10	0.00	1.00	0.10
TLCB	0.39	0.14	0.11	0.10	0.10	1.00

politics

	HS	PB	GB	RB	LB	TLCB
HS	1.00	0.01	0.02	0.11	0.08	0.42
PB	0.01	1.00	0.00	0.02	0.06	0.09
GB	0.02	0.00	1.00	0.04	0.02	0.01
RB	0.11	0.02	0.04	1.00	0.02	0.08
LB	0.08	0.06	0.02	0.02	1.00	0.07
TLCB	0.42	0.09	0.01	0.08	0.07	1.00

sports

	HS	PB	GB	RB	LB	TLCB
HS	1.00	0.00	0.00	0.09	0.07	0.30
PB	0.00	1.00	0.03	0.05	0.04	0.02
GB	0.00	0.03	1.00	0.02	0.07	0.04
RB	0.09	0.05	0.02	1.00	0.06	0.16
LB	0.07	0.04	0.07	0.06	1.00	0.06
TLCB	0.30	0.02	0.04	0.16	0.06	1.00

jobEducation

	HS	PB	GB	RB	LB	TLCB
HS	1.00	0.01	0.02	0.16	0.09	0.39
PB	0.01	1.00	0.01	0.00	0.03	0.00
GB	0.02	0.01	1.00	0.00	0.07	0.10
RB	0.16	0.00	0.00	1.00	0.07	0.23
LB	0.09	0.03	0.07	0.07	1.00	0.14
TLCB	0.39	0.00	0.10	0.23	0.14	1.00

entertainment

	HS	PB	GB	RB	LB	TLCB
HS	1.00	0.00	0.11	0.06	0.09	0.41
PB	0.00	1.00	0.03	0.00	0.00	0.03
GB	0.11	0.03	1.00	0.00	0.12	0.11
RB	0.06	0.00	0.00	1.00	0.06	0.10
LB	0.09	0.00	0.12	0.06	1.00	0.09
TLCB	0.41	0.03	0.11	0.10	0.09	1.00

Fig. 1. Correlation heatmap for each bias dimension of different domains, calculated using Cramér's $\mathcal{V}$. Higher values indicate stronger correlations. (HS: Hate Speech, PB: Political Bias, GB: Gender Bias, RB: Racial Bias, LB: Linguistic Bias, TLCB: Text-level Context Bias)

4.2 Correlated Bias Dimensions

Based on the definitions of our media bias dimensions, we divide media bias dimensions into two categories: i) style-based bias (*linguistic bias, text-level context bias*), which focuses on the phrasing of biased texts, and ii) content-based bias (*hate speech, gender bias, political bias, racial bias*), concerning the topics of bias within the content. Noting that these groups contribute differently to the media bias spectrum and may require distinct identification frameworks, we analyze correlations within and between these categories. For each bias dimension, we examine their co-occurrence within topic domains through pairwise chi-square tests using binary labels. We report the results as Cramer's $\mathcal{V}$ values in Figure 1. Our analysis confirms strong statistical significance across all pairwise correlations, with the highest p-value at 1.87×10^{-3}.

In order to understand how media bias are expressed in different domains, we first investigate the correlations we observe associated with style-based bias dimensions. Specifically, we observe that for all topic domains, *text-level context bias* is mostly associated with *hate speech*. In sports and job & education domain, we observe *hate speech* also has a moderate positive correlation with racial bias with Cramer's $\mathcal{V} > 0.15$. Unlike *text-level context bias*, *linguistic bias*, which primarily focuses on specific word choices, does not exhibit a clear positive correlation with some specific types of biases, but more evenly correlated with all content-based bias dimensions. This observation suggests that content-based bias dimensions, particularly hate speech, are more often expressed through biased descriptions rather than specific biased terms.

For content-based biases, we focus exclusively on correlations that are relatively strong, specifically where Cramer's $\mathcal{V} > 0.1$. Among the content-based bias dimensions, unlike other content-based biases, we observe *hate speech* often coexists with other types of content-based biases. Across all topic domains, we observe that *hate speech* is most significantly correlated with *political bias* in healthcare, *racial bias* in sports, *racial bias* in politics, *racial bias* in job & education, and *gender bias* in entertainment. The aforementioned correlations may reflect the nuanced ways in which content creators and social media users engage with topics sensitive to identity and political context. For instance, in the healthcare domain, political discussions often intersect with deeply polarized issues such as healthcare policy and reproductive rights, which often incite hate speech. In politics and job & education, racial discussions usually evoke strong biases, potentially escalating into hate speech.

4.3 Time Series Analysis

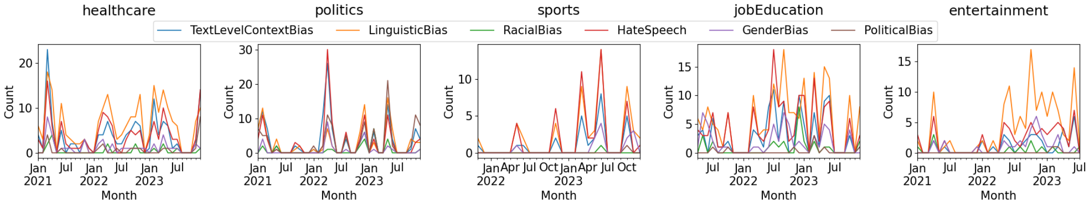


Fig. 2. Line plot visualizations of monthly aggregated counts of bias dimensions. Key observations include: 1) Hate speech manifests in varying proportions of the two types of style-based bias dimensions across different domains. 2) Notable surges in aggressive biases are observed in specific months within the politics domain, supporting our hypothesis that biases in these domains are more event-driven compared to others.

Social media is a rapidly evolving field with frequent shifts in content and interaction patterns. Recognizing the importance of temporal analysis, we monitor the dynamics of biased content by aggregating data within each bias dimension monthly. This data is analyzed across different topic domains and presented in Figure 2. We standardize the time series for each domain to facilitate pairwise comparisons and perform t-tests to investigate differences across domains. Additionally, we conduct intra-domain analysis, reporting on the top-2 correlated bias dimensions using Pearson's correlation.

We first investigate the strong (style, content)-based correlation pairs. Most notably, diverging from the typically dominant correlation of (*hate speech*, *text-level content bias*), *linguistic bias* emerges as the most closely correlated style-based bias with *hate speech* with a Pearson's coefficient of 0.86 in the aggregated time series for the entertainment domain. Furthermore, the month-interval aggregation significantly strengthens the (*hate speech*, *linguistic bias*) correlation across all domains, with the lowest Pearson's coefficient observed being 0.77 in the politics domain. The observation that temporal aggregation enhances the visibility of (*hate speech*, *linguistic bias*) correlation is likely due to the smoothing of outliers and noise in the data. The observed correlation in the entertainment domain, in particular, might reflect a trend where hate speech is more frequently expressed through nuanced language choices rather than content cues.

For content-based bias dimensions, we observe that aggregation introduces distinct correlations between bias dimensions. For clarity, we refer to correlations observed in monthly aggregated bias counts as 'short-term correlations,' and those across the entire dataset as 'long-term correlations.' We summarize the differences in the highest correlated pairs of biases as follows: 1) In the politics domain, the most closely correlated pair is (*hate speech, gender bias*) with a Pearson's coefficient of 0.92, which is significantly higher than the correlation between (*hate speech, political bias*) with a coefficient of 0.69. 2) In the sports domain, *gender bias* and *racial bias* show the closest correlation. 3) In job & education domain, *hate speech* is closely correlated with both *gender bias* and *racial bias* with Pearson's coefficients of 0.54 and 0.53 respectively.

5 Conclusion

Our study aims to advance the understanding of media bias by introducing and investigating a social media dataset that spans multiple domains and bias dimensions, collected from YouTube over the past five years. Moreover, our findings reveal significant differences in how biases are expressed across various domains such as politics, sports, and healthcare. We also discover fluctuations in the correlations between bias dimensions in response to surges in social media posts. These findings underscore the complex and evolving nature of media bias and lay the foundation for the future development of multi-dimensional bias identification systems. By advancing investigations into media bias, we hope to equip both researchers and practitioners with the tools necessary to address and mitigate the impacts of media bias, ultimately fostering a fairer media environment.

Acknowledgement

This research is supported in part by the National Science Foundation under Grant No. IIS-2202481, CHE-2105032, IIS-2130263, CNS-2131622, CNS-2140999. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

References

1. Aksenov, D., Bourgonje, P., Zaczynska, K., Ostendorff, M., Moreno-Schneider, J., Rehm, G.: Fine-grained classification of political bias in german news: A data set and initial experiments. In: WOAH (2021), https://api.semanticscholar.org/CorpusID:236486143
2. Beukeboom, C.J., Burgers, C.: Linguistic bias (07 2017). https://doi.org/10.1093/acrefore/9780190228613.013.439, https://doi.org/10.1093/acrefore/9780190228613.013.439

3. Bingel, J., Søgaard, A.: Identifying beneficial task relations for multi-task learning in deep neural networks. In: Lapata, M., Blunsom, P., Koller, A. (eds.) Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers. pp. 164–169. Association for Computational Linguistics, Valencia, Spain (Apr 2017), https://aclanthology.org/E17-2026
4. Chen, S., Zhang, Y., Yang, Q.: Multi-task learning in natural language processing: An overview (2021)
5. Davidson, T., Warmsley, D., Macy, M.W., Weber, I.: Automated hate speech detection and the problem of offensive language. CoRR **abs/1703.04009** (2017), http://arxiv.org/abs/1703.04009
6. Dixon, T.L., Azocar, C.L.: Priming crime and activating Blackness: Understanding the psychological impact of the overrepresentation of Blacks as lawbreakers on television news. Journal of Communication **57**(2), 229–253 (2007). https://doi.org/10.1111/j.1460-2466.2007.00341.x, place: United Kingdom Publisher: Blackwell Publishing
7. Färber, M., Burkard, V., Jatowt, A., Lim, S.: A multidimensional dataset based on crowdsourcing for analyzing and detecting news bias. p. 3007–3014. CIKM '20, Association for Computing Machinery, New York, NY, USA (2020), https://doi.org/10.1145/3340531.3412876
8. Feldman, S.: Political ideology. In: The Oxford handbook of political psychology, 2nd ed., pp. 591–626. Oxford University Press, New York, NY, US (2013)
9. Horych, T., Wessel, M., Wahle, J.P., Ruas, T., Waßmuth, J., Greiner-Petter, A., Aizawa, A., Gipp, B., Spinde, T.: Magpie: Multi-task media-bias analysis generalization for pre-trained identification of expressions (2024)
10. Hube, C., Fetahu, B.: Detecting biased statements in wikipedia. In: Companion Proceedings of the The Web Conference 2018. p. 1779–1786. WWW '18, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE (2018). https://doi.org/10.1145/3184558.3191640, https://doi.org/10.1145/3184558.3191640
11. Kou, Z., Shang, L., Zeng, H., Zhang, Y., Wang, D.: Exgfair: A crowdsourcing data exchange approach to fair human face datasets augmentation. In: 2021 IEEE International Conference on Big Data (Big Data). pp. 1285–1290. IEEE (2021)
12. Ma, W., Lou, R., Zhang, K., Wang, L., Vosoughi, S.: GradTS: A gradient-based automatic auxiliary task selection method based on transformer networks. In: Moens, M.F., Huang, X., Specia, L., Yih, S.W.t. (eds.) Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. pp. 5621–5632. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic (Nov 2021), https://aclanthology.org/2021.emnlp-main.455
13. Mathew, B., Saha, P., Yimam, S.M., Biemann, C., Goyal, P., Mukherjee, A.: Hatexplain: A benchmark dataset for explainable hate speech detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 14867–14875 (2021)
14. Nickerson, R.S.: Confirmation bias: A ubiquitous phenomenon in many guises. Review of General Psychology **2**(2), 175 – 220 (1998). https://doi.org/10.1037/1089-2680.2.2.175, https://psycnet.apa.org/record/2018-70006-003, cited by: 4362
15. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C.L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., Lowe, R.: Training language models to follow instructions with human feedback (2022)
16. Pariser, E.: The filter bubble: What the Internet is hiding from you. penguin UK (2011)

17. Raza, S., Reji, D.J., Ding, C.: Dbias: detecting biases and ensuring fairness in news articles. International Journal of Data Science and Analytics **17**(1), 39–59 (Jan 2024). https://doi.org/10.1007/s41060-022-00359-4, https://doi.org/10.1007/s41060-022-00359-4
18. Recasens, M., Danescu-Niculescu-Mizil, C., Jurafsky, D.: Linguistic models for analyzing and detecting biased language. In: Schuetze, H., Fung, P., Poesio, M. (eds.) Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1650–1659. Association for Computational Linguistics, Sofia, Bulgaria (Aug 2013), https://aclanthology.org/P13-1162
19. Rodrigo-Ginés, F.J., de Albornoz, J.C., Plaza, L.: A systematic review on media bias detection: What is media bias, how it is expressed, and how to detect it. Expert Systems with Applications **237**, 121641 (2024). https://doi.org/https://doi.org/10.1016/j.eswa.2023.121641, https://www.sciencedirect.com/science/article/pii/S0957417423021437
20. Ruder, S., Plank, B.: Learning to select data for transfer learning with Bayesian optimization. In: Palmer, M., Hwa, R., Riedel, S. (eds.) Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. pp. 372–382. Association for Computational Linguistics, Copenhagen, Denmark (Sep 2017). https://doi.org/10.18653/v1/D17-1038, https://aclanthology.org/D17-1038
21. Samory, M., Sen, I., Kohne, J., Flöck, F., Wagner, C.: "unsex me here": Revisiting sexism detection using psychological scales and adversarial samples. CoRR **abs/2004.12764** (2020), https://arxiv.org/abs/2004.12764
22. Spinde, T., Hinterreiter, S., Haak, F., Ruas, T., Giese, H., Meuschke, N., Gipp, B.: The Media Bias Taxonomy: A Systematic Literature Review on the Forms and Automated Detection of Media Bias. ACM Computing Surveys (2023)
23. Spinde, T., Krieger, J.D., Ruas, T., Mitrović, J., Götz-Hahn, F., Aizawa, A., Gipp, B.: Exploiting transformer-based multitask learning for the detection of media bias in news articles. In: Smits, M. (ed.) Information for a Better World: Shaping the Global Future. pp. 225–235. Springer International Publishing, Cham (2022)
24. Voigt, R., Jurgens, D., Prabhakaran, V., Jurafsky, D., Tsvetkov, Y.: RtGender: A corpus for studying differential responses to gender. In: Calzolari, N., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Hasida, K., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S., Tokunaga, T. (eds.) Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). European Language Resources Association (ELRA), Miyazaki, Japan (May 2018), https://aclanthology.org/L18-1445
25. Wang, W.Y.: "liar, liar pants on fire": A new benchmark dataset for fake news detection. In: Barzilay, R., Kan, M.Y. (eds.) Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). pp. 422–426. Association for Computational Linguistics, Vancouver, Canada (Jul 2017). https://doi.org/10.18653/v1/P17-2067, https://aclanthology.org/P17-2067
26. Wessel, M., Horych, T., Ruas, T., Aizawa, A., Gipp, B., Spinde, T.: Introducing mbib - the first media bias identification benchmark task and dataset collection. In: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 2765–2774. SIGIR '23, Association for Computing Machinery, New York, NY, USA (2023). https://doi.org/10.1145/3539618.3591882
27. Zhang, D.Y., Kou, Z., Wang, D.: Fairfl: A fair federated learning approach to reducing demographic bias in privacy-sensitive classification models. In: 2020 IEEE International Conference on Big Data (Big Data). pp. 1051–1060. IEEE (2020)