Dataset Annotation and Model Building for Identifying Biases in News Narratives ⋆

Shaina Raza[1,*], Mizanur Rahman[2] and Shardul Ghughe[1]

[1]Vector Institute for Artificial Intelligence, Toronto, ON, Canada
[2]Royal Bank of Canada

Abstract

In the digital information age, detecting and mitigating linguistic biases, particularly in political discourse, presents a critical challenge. This study addresses this issue by developing a comprehensive pipeline for annotating and analyzing a dataset of news articles. The process involved expert-driven rule creation for identifying bias indicators, followed by annotation using a large language model (LLM). The integration of a Named Entity Recognition (NER) model is tailored to identify and categorize biases within the annotated dataset. The study's novelty lies in its blend of human expertise with LLM, enhancing the accuracy and relevance of bias detection. We conducted a comparative analysis of several language models. This research contributes significantly to both the theoretical and practical domains of bias detection in texts. The annotated dataset and the best model are made available for use and application.

Keywords

Bias Detection, Misinformation, Annotations, Large Language Model

1. Introduction

In recent decades, the domain of news reporting and content generation has undergone a profound transformation [1, 2]. Traditionally, news articles were developed by journalists and reporters, who relied on their skills and ethical guidelines to convey information. With technological advancements, there has been a significant shift towards AI-driven approaches [3]. The emergence of advanced artificial intelligence (AI) and Large Language Models (LLMs) [4] has increasingly played a dominant role in various aspects of news production, such as editorial assistance, multi-language translation, content summarization, and social media updates [5].

The integration of AI in journalism has enhanced news production's efficiency and scope but also introduced complex challenges, particularly in bias detection. The term bias' has diverse definitions [6], but in linguistics [7], it generally refers to the presence of non-neutral or prejudiced viewpoints in content. Historically, biases in news were attributed to the subjective perspectives of human authors, writers, editors and publishers [8]. However, AI's involvement adds a layer of complexity, as AI algorithms often mirror the biases in their training datasets

In: Text2Story 2024: Seventh International Workshop on Narrative Extraction from Texts held in conjunction with the 46th European Conference on Information Retrieval.Glasgow, Scotland

*Corresponding author.
shaina.raza@utoronto.ca (S. Raza); mizanur.york@gmail.com (M. Rahman); shardul.ghuge@mail.utoronto.ca (S. Ghughe)
0000-0003-1061-5845 (S. Raza); 0009-0001-0728-073 (S. Ghughe)

[9, 10]. These biases, whether explicit or implicit, can profoundly influence public opinion and societal discourse.

This paper addresses the challenge of detecting biases within political news available on web. We compiled an extensive dataset from diverse sources (CNN, BBC, Global etc.) via Google news feeds, consisting of articles selected over five months (May-October 2023). Post-collection, we conducted comprehensive preprocessing to structure and annotate the data for bias detection in both binary label and multi-class token classification tasks. This work focuses particularly on the multi-class token classification task and the methods used for annotating this data. The main contributions of this work are as follows:

Annotation Framework: We introduce an annotation framework wherein a team of expert annotators first developed rules to identify indicators of bias in news content. These indicators were then translated into structured prompts for analysis using OpenAI's advanced LLM [11]. This annotation scheme, combining the efforts of human experts and LLMs, aligns with recent trends in NLP research [12] and has shown promising results [13]. The aim of this blended approach is to efficiently and accurately label news content as biased or unbiased and to pinpoint biased terms within the text.

Named Entity Recognition (NER) Model Development: Following the identification of biased terms, we introduce a multi-label token classification task and develop a Named Entity Recognition (NER) model specifically tailored to this annotated dataset. This model is designed to focus on the multi-label classification of biased terms found in news articles. To establish benchmarks, we evaluated various state-of-the-art models for this purpose.

This work offers a novel methodology for understanding the biases identified in political news content. This research is very timely for comprehending data construction and annotation, as well as establishing a baseline for bias detection. It also presents potential avenues for future work directed towards mitigating biases.

2. Related Work

In its broadest definition, bias can be defined as a tendency or preference towards or against certain groups or individuals [14]. Bias in linguistics is often identified through specific words, phrases, or patterns such as slurs, hate speech, and toxicity [15], or through more subtle means that reflect societal biases or subjective perspectives [16].

A diverse collection of datasets has been assembled for various bias detection NLP tasks. For example, Toxigen [17] leverages GPT-3 with 274K prompts for implicit hate speech classification, while Stereotype [18] gathers 17K Wikidata entries for quantifying biases based on gender, profession, and race. HolisticBias [19] , MBIB [20], RedditBIAS [21], BIGNEWS [22] are among valuable resources for advancing research in the areas of bias detection, and linguistic analysis.

Research studies have underscored the inadvertent introduction of societal and cultural biases into training data [23, 24, 25, 26, 9, 27], as well as the amplification of disparities in NLP outcomes [28, 25]. Recent work has examined various dimensions of bias, including gender, race, religion, and emotions, with a focus on bias mitigation strategies [29, 30, 31, 32]. Additionally, efforts have been made to evaluate LM's understanding of bias-related tasks [33], introduce frameworks for bias reduction [34], and detect biases at both the sentence level within news

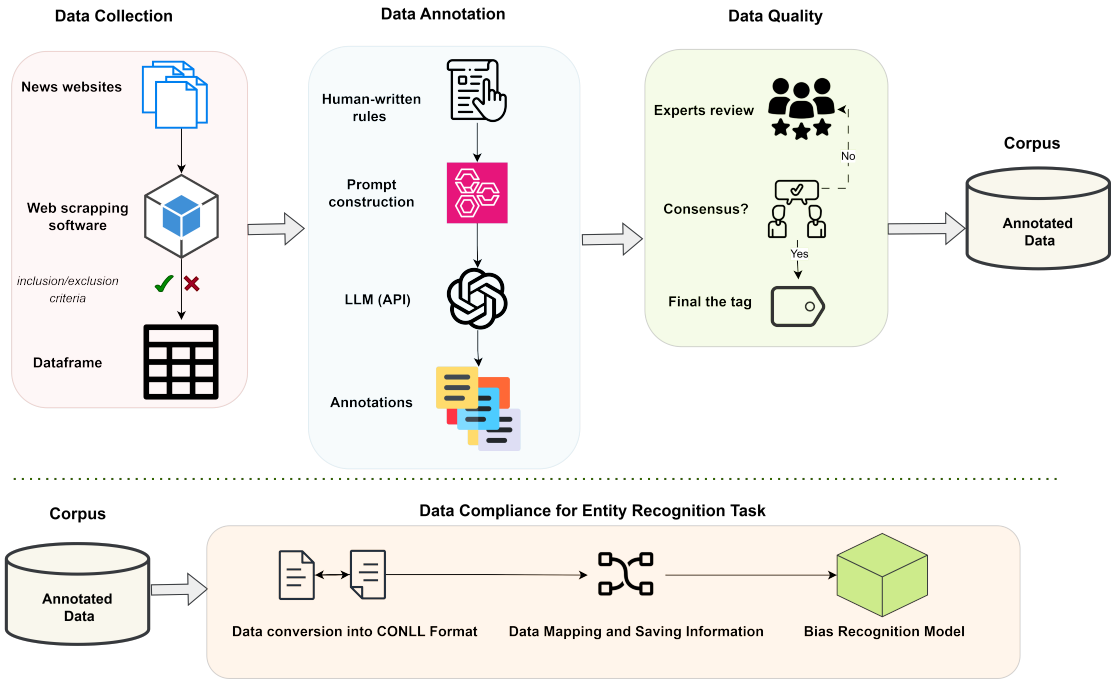

Figure 1: Data Annotation and Model Training Pipeline

articles [35] and in letters of recommendation [36]. These interdisciplinary studies emphasize the need for ongoing vigilance and proactive measures to promote fairness in NLP models.

Our study introduces a novel approach, inspired by prior research, where we employ human-designed rules to guide a LLM in annotating bias within news articles. We then use this data as a baseline for multi-label bias detection methods.

3. Data Annotation and Model Training Pipeline

We propose a data annotation and model training pipeline in this work, shown in Figure 1. The initial stage consists of a *data collection* phase, where data is scraped from news websites and processed into a dataframe based on inclusion/exclusion criteria. The second stage consists of *annotation* process, which involves human-written rules, prompt construction, and the use of a LLM API for generating annotations. The third stage is the *data quality* stage, where experts assess the annotations, reach a consensus, and finalize the tags. The final stage is the data compliance and model building phase, the where annotated data is converted into a specific format (e.g., CoNLL), mapped, saved, and used to train a bias recognition . Next, we explain each phase of the pipeline.

3.1. Dataset Construction

Our dataset is focussed on racial slurs and biases in political speeches in North America. The data is curated using Google RSS by using ‘feedparser‘ and ‘newspaper.Article‘ API to extract article text, publication dates, news outlets, and URLs. These articles represent a wide range of political news sources, such as ’The Daily Chronicle’, ’Global Times News’, ’Liberty Voice’, ’New Age Politics’, and ’The Public Perspective’. The data collection strategy is as:

- **Time Frame:** Five months from May to October 2023.
- **Inclusion/Exclusion Criteria:** Relevant, accessible articles included; incomplete, non-English, redundant articles excluded.
- **Keywords**: Political, economic, social, cultural, environmental, scientific, global affairs, historical, and emerging social issues.
- **Total Articles Scraped:** 40,000.
- **Articles for Annotation:** 4,000, covering a broad range of topics.
- **Annotation Method:** Initial annotations with GPT-Turbo-3.5 API, followed by expert verification.
- **Basic Fields:** Text, biased-words (n-gram words can be tokens, phrases or a pattern in a sentence), bias-label.

3.2. Data Annotation

Rule Development : The initial phase of our dataset construction involves a collaborative effort from five experts across computer science, social science, political science, and media analysis. These experts developed a comprehensive set of rules to identify various forms of political bias in news narratives. The experts focussed on linguistic patterns and narrative structures indicative of biases like ideological leanings, emotive language, and selective reporting. An example of a rule is as:

- *Rule*: Identification of Emotive Language
- *News Text*: “The government’s disastrous policies have resulted in economic chaos.”
- *Annotation*: Words such as “disastrous” and “chaos” are tagged as emotionally charged, which signifies a negative bias against the government’s policies.

Prompt Construction : Following the rule development, these guidelines are synthesized into a single, comprehensive prompt designed for the OpenAI [11] interface. This annotation scheme is inspired by recent works related research [12, 13]. Specifically, we used the *gpt-3.5-turbo*, which is from the ChatGPT model family. We provided the few-shot (5) examples to the API along with the prompt. Our prompt is:

- *Prompt*: “Carefully analyze the given news text. First, determine whether each sentence exhibits political bias. If a sentence is biased, identify and categorize instances of political bias within it, such as ideological leanings, emotive language, and selective reporting. Highlight specific words or phrases that demonstrate these biases. Pay attention to the context, subtleties of phrasing, and underlying implications in your analysis. For biased sentences, organize the identified words and phrases into a separate dataframe, categorizing them based on the type of bias they represent.”

- *Few-Shot Examples*: This prompt is accompanied by a set of annotated examples, which include both biased and unbiased sentences. These examples are intended to guide the LLM in distinguishing biased sentences from unbiased ones and then in identifying specific biased words and phrases within the biased sentences.

We obtained the dataframe with the text column, bias label and 'bias-words' identified in this step.

3.3. Data Quality Assurance

We implemented a human verification system to validate the annotations produced by the OpenAI model. This verification process was carried out over a two-month period by 5 experts and 2 students to ensure the uniformity and reliability of all annotations. Where necessary, adjustments were made based on a consensus among the team. To measure the success of this collaboration between our human experts and the AI model, we adopted an Inter-Rater Reliability Score. Currently, this metric stands at approximately 78%, which indicates a high level of concordance between human and AI evaluations.

3.4. Data Compliance

We transformed our annotated dataset with biased words into the CONLL (Conference on Natural Language Learning) [37] format to align with standard NLP practices with FAIR (Findability, Accessibility, Interoperability, and Reusability) principles [38]. This data format involves structuring each word from our 'text' data as a separate entity, with corresponding 'biased_words' annotations, in a tabular layout for a sequence labeling task. Below is an example of a politically biased sentence formatted in the CONLL style, identifying specific biased words:

- **Word:** The, government's, reckless, policies, harm, the, economy;
- **Bias Tag:** O, O, B-BIAS, I-BIAS, B-BIAS, I-BIAS, I-BIAS.

In this example, "B-BIAS" marks the beginning of a biased phrase, "I-BIAS" continues the phrase, and "O" indicates words that are not part of a biased phrase.

Annotated Dataset: The annotated dataset consists of 4,000 records, comprising 50.8k tokens in total, with 20% of the dataset specifically reserved for testing. The annotated dataset is available here [1].

3.5. Bias Recognition Model

We develop a multi-label token classification model, an Named Entity Recognition (NER) system, which is designed to identify linguistic biases in textual content. The proposed model innovates beyond traditional NER approaches by incorporating features sensitive to the manifestation of bias. It aims to show not only the presence of bias but also its scope and boundary within text.

The foundation of the model is a transformer-based neural network, employing the BERT [39] architecture for contextualized token embeddings. The classification layer of the model

[1]https://huggingface.co/datasets/newsmediabias/FAKE-NEWS-BIASES-LABELLED

is a softmax layer, fine-tuned to distinguish between three classes: B-BIAS for the beginning of a biased entity, I-BIAS for tokens inside a biased entity, and O for tokens outside of biased entities. The classification of each token is computed as:

$$E_{\text{bias}} = \text{softmax}(W \cdot T_{\text{BERT}} + b) \tag{1}$$

where E_{bias} signifies the probability distribution over the three classes for each token, W is the weight matrix, b is the bias vector, and T_{BERT} denotes the embeddings from BERT. The corpus used for training has been annotated with these bias-specific entity tags, as elaborated in Section 3.2. To optimize the parameters of our model, we employ the cross-entropy loss function:

$$L = -\sum_{c \in \{B,I,O\}} y_c \cdot \log(\hat{y}_c) \tag{2}$$

where y_c is the ground truth label, and $\hat{y}_c$ is the predicted probability for each class. The summation runs over the classes B (B-BIAS), I (I-BIAS), and O (Outside).

4. Experiments

We present a comparative analysis of various LMs on the task of detecting linguistic biases within our annotated dataset. Each model is evaluated based on its ability to accurately classify the pre-defined bias entity categories: B-BIAS, I-BIAS, and O (non-biased).

4.1. Baseline Models

In our study, we compare several baseline models, each fine-tuned on our annotated dataset for bias detection. We utilize BERT ‘bert-base-uncased‘ as a foundational model for deep bidirectional training, serving as our primary baseline. Additionally, we include DistilBERT ‘distilbert-base-uncased‘, a lighter version of BERT that offers a balance between performance and efficiency. ALBERT ‘albert-base-v2‘ is employed for its parameter-reduction techniques, which enhance efficiency and highlight the trade-offs in bias detection. Furthermore, we useRoBERTa ‘roberta-base-uncased‘, a BERT-alike model but trained on a larger corpus, to examine its effectiveness in bias detection [40]. We also use Llama-7b-chat [41] by Meta, which is a generative AI LLM and we tested its bias recognition ability with few-shot learning, where a limited number of examples, or *shots*, are provided to perform a new instance of the task.

4.2. Settings and Hyperparameter Configuration

The dataset, detailed in Section 3, was partitioned into three distinct subsets: training (80), validation (10), and testing (10). We use precision, recall, and the F1-score to evaluate the model’s performance on the bias entity recognition task.

Our experiments were conducted using Google Colab Pro, with GPUs such as the NVIDIA Tesla P100 and A100 (at times), along with a cloud-based environment. The software stack consisted of Python 3.9, PyTorch 1.7.1 as the primary deep learning framework, and the Hugging Face Transformers library version 4.36 for the implementation of transformer-based models.

Hyperparameter tuning played a vital role in optimizing the performance of each baseline model. For BERT, DistilBERT , and RoBERTa, we chose a learning rate of 2e-5, a batch size of 16 (smaller to save memory cycles), and a training duration of 4 epochs. For ALBERT, we opted for a learning rate of 1e-5, a smaller batch size of 16, and a similar epoch count of 4. Hyperparameters were specifically adjusted to suit the few-shot learning requirements of Llama 2, aiming to maximize its efficacy in learning from a limited number of examples. All models were fine-tuned on the same annotated dataset (except Llama 2), to ensure a basis for comparison across the different approaches.

5. Results and Analysis

5.1. Performance Comparison of Baselines

Table 1
Performance Comparison of Baseline Models

Model	Precision	Recall	F1-Score
BERT (Base-Uncased)	0.82	0.80	0.81
DistilBERT (Base-Uncased)	0.79	0.77	0.78
ALBERT (Base-v2)	0.81	0.79	0.80
RoBERTa (Base-Uncased)	**0.90**	**0.89**	**0.89**
Llama 2 with Few-Shot Prompts	0.78	0.76	0.77

The comparative analysis of baseline models, as detailed in Table 1. The results provides insightful findings regarding the effectiveness in the task of bias detection.

RoBERTa (Base-Uncased) showed best performance in our analysis, achieving the highest scores across precision (0.90), recall (0.89), and F1-score (0.89). This high performance can be attributed to RoBERTa's enhanced training and larger model architecture. BERT (Base-Uncased) and ALBERT (Base-v2) display competitive results, with BERT marginally outperforming in precision and F1-score. This suggests that while ALBERT's design optimizes memory efficiency, but it may slightly reduce effectiveness in complex tasks like bias detection.

DistilBERT (Base-Uncased) and Llama 2 with Few-Shot Prompts show decent, although at relatively lower performance. The reduced complexity of DistilBERT, designed for efficiency, likely impacts its ability to fully understand the intricacies of bias. Meanwhile, Llama 2's performance suggests that while few-shot learning is good rapid adaptation, but it might effect the bias detection tasks.

Overall, the results highlight the effectiveness of comprehensive training and model complexity in bias detection. These results also pointing to the nuanced trade-offs between model size, efficiency, and task-specific performance.

5.2. Qualitative Comparison of Model Performances

To provide a deeper understanding of each model's ability in bias detection, we present comparing instances, in Table 2, where each model successfully detected bias against instances where it

Table 2
Examples of Bias Detection Performance Across Models

Model	Successful Bias Detection	Missed Bias Detection
BERT	*"Unprecedented policy blatantly disregards minority rights."* - Correctly identified as biased.	*"The policy has received mixed reactions from various groups."* - Missed implicit bias.
Distil-BERT	*"The decision, clearly partisan, has sparked debate."* - Accurately detected bias.	*"Critics argue the decision could have unforeseen consequences."* - Overlooked slight bias.
ALBERT	*"This overtly discriminatory act has been condemned."* - Successfully identified bias.	*"There are concerns about the effectiveness of the new regulation."* - Failed.
RoBERTa	*"The policy blatantly favors the wealthy, ignoring others."* - Correctly identified bias.	*"Some experts are questioning the necessity of this policy."* - Missed nuanced bias.
Llama 2	*"A clear case of nepotism has been observed in the selection process."* - Accurately identified bias.	*"The project's relevance is being debated in academic circles."* - Failed to detect implicit bias.

failed. For example, we observe that models like BERT and RoBERTa effectively identify explicit biases but sometimes miss more nuanced ones. DistilBERT and ALBERT tend to overlook subtle biases, which hints a need for refinement in detecting less apparent bias instances. The performance varies greatly with the context of the bias, as seen in Llama 2's accurate detection in cases of 'nepotism' but could not perform good in abstract scenarios.

6. Conclusion

This research presents the development and implementation of a comprehensive data annotation and model building pipeline. A key outcome of this research has been the successful annotation of a dataset comprising 4,000 records, alongside the identification of the most effective model for bias detection. We make both the annotated data and best model weights (RoBERTa) publicly available for further research and application. This study also acknowledges the dynamic nature of language, highlighted by the phenomenon of data drifts [27, 42]. This underscores the need for ongoing monitoring and updating of our data annotation protocols. The limited size of the dataset, at 4k records, poses potential limitations on the depth and diversity of biases that can be captured. Therefore, expanding the dataset is crucial for enhancing the robustness and generalizability of our models. The Llama 2 model [41], with its focus on few-shot learning, demonstrated promising results. However, to optimize its performance, we recommend fine-tuning it with more explicit instructions. This research lays a foundation for data annotation and model building in the evolving landscape of language and bias.

Acknowledgments

References

[1] T. A. Van Dijk, et al., Power and the news media, Political communication and action 6 (1995) 9–36.

[2] S. Raza, C. Ding, News recommender system: a review of recent progress, challenges, and opportunities, Artificial Intelligence Review (2022) 1–52.

[3] J. Pavlik, The impact of technology on journalism, Journalism studies 1 (2000) 229–237.

[4] Y. Chang, X. Wang, J. Wang, Y. Wu, K. Zhu, H. Chen, L. Yang, X. Yi, C. Wang, Y. Wang, et al., A survey on evaluation of large language models, arXiv preprint arXiv:2307.03109 (2023).

[5] C. Chin, Navigating the risks of artificial intelligence on the digital news landscape (2023).

[6] E. Ntoutsi, P. Fafalios, U. Gadiraju, V. Iosifidis, W. Nejdl, M.-E. Vidal, S. Ruggieri, F. Turini, S. Papadopoulos, E. Krasanakis, et al., Bias in data-driven artificial intelligence systems—an introductory survey, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 10 (2020) e1356.

[7] C. J. Beukeboom, Mechanisms of linguistic bias: How words reflect and maintain stereotypic expectancies, in: Social cognition and communication, Psychology Press, 2014, pp. 313–330.

[8] S. H. Stocking, P. H. Gross, Understanding errors, biases that can affect journalists, The Journalism Educator 44 (1989) 4–11.

[9] S. Dev, E. Sheng, J. Zhao, A. Amstutz, J. Sun, Y. Hou, M. Sanseverino, J. Kim, A. Nishi, N. Peng, et al., On measures of biases and harms in nlp, arXiv preprint arXiv:2108.03362 (2021).

[10] B. Hutchinson, V. Prabhakaran, E. Denton, K. Webster, Y. Zhong, S. Denuyl, Social biases in NLP models as barriers for persons with disabilities, in: Proceedings of the Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 5491–5501. doi:10.18653/v1/2020.acl-main.487. arXiv:2005.00813.

[11] OpenAI, Gpt-3.5, https://openai.com/, 2023. Accessed: 2023-12-24.

[12] J. White, Q. Fu, S. Hays, M. Sandborn, C. Olea, H. Gilbert, A. Elnashar, J. Spencer-Smith, D. C. Schmidt, A prompt pattern catalog to enhance prompt engineering with chatgpt, arXiv preprint arXiv:2302.11382 (2023).

[13] F. Gilardi, M. Alizadeh, M. Kubli, Chatgpt outperforms crowd-workers for text-annotation tasks, arXiv preprint arXiv:2303.15056 (2023).

[14] I. O. Gallegos, R. A. Rossi, J. Barrow, M. M. Tanjim, S. Kim, F. Dernoncourt, T. Yu, R. Zhang, N. K. Ahmed, Bias and fairness in large language models: A survey, arXiv preprint arXiv:2309.00770 (2023).

[15] S. Raza, M. Garg, D. J. Reji, S. R. Bashir, C. Ding, Nbias: A natural language processing framework for bias identification in text, Expert Systems with Applications (2023) 121542.

[16] Z. Yanbo, Implicit bias or explicit bias: an analysis based on natural language processing, in: 2020 International conference on computing and data science (CDS), IEEE, 2020, pp. 52–55.

[17] T. Hartvigsen, S. Gabriel, H. Palangi, M. Sap, D. Ray, E. Kamar, ToxiGen: A Large-Scale Machine-Generated Dataset for Adversarial and Implicit Hate Speech Detection, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 3309–3326. URL: https://aclanthology.org/2022.acl-long.234. doi:10.18653/v1/2022.acl-long.234.

[18] M. Nadeem, A. Bethke, S. Reddy, StereoSet: Measuring stereotypical bias in pretrained language models, 2020. URL: http://arxiv.org/abs/2004.09456, arXiv:2004.09456 [cs].

[19] E. M. Smith, M. Hall, M. Kambadur, E. Presani, A. Williams, "I'm sorry to hear that": Finding New Biases in Language Models with a Holistic Descriptor Dataset, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022, pp. 9180–9211. URL: https://aclanthology.org/2022.emnlp-main.625. doi:10.18653/v1/2022.emnlp-main.625.

[20] M. Wessel, T. Horych, T. Ruas, A. Aizawa, B. Gipp, T. Spinde, 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, 2023, pp. 2765–2774.

[21] S. Barikeri, A. Lauscher, I. Vulić, G. Glavaš, RedditBias: A Real-World Resource for Bias Evaluation and Debiasing of Conversational Language Models, in: C. Zong, F. Xia, W. Li, R. Navigli (Eds.), Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Association for Computational Linguistics, Online, 2021, pp. 1941–1955. URL: https://aclanthology.org/2021.acl-long.151. doi:10.18653/v1/2021.acl-long.151.

[22] Y. Liu, X. F. Zhang, D. Wegsman, N. Beauchamp, L. Wang, Politics: pretraining with same-story article comparison for ideology prediction and stance detection, arXiv preprint arXiv:2205.00619 (2022).

[23] A. Caliskan, J. J. Bryson, A. Narayanan, Semantics derived automatically from language corpora contain human-like biases, Science 356 (2017) 183–186.

[24] T. Spinde, M. Plank, J.-D. Krieger, T. Ruas, B. Gipp, A. Aizawa, Neural media bias detection using distant supervision with babe-bias annotations by experts, in: Findings of the Association for Computational Linguistics: EMNLP 2021, 2021, pp. 1166–1177.

[25] M. Färber, V. Burkard, A. Jatowt, S. Lim, A multidimensional dataset based on crowdsourcing for analyzing and detecting news bias, in: Proceedings of the 29th ACM International Conference on Information & Knowledge Management, 2020, pp. 3007–3014.

[26] T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, A. T. Kalai, Man is to computer programmer as woman is to homemaker? debiasing word embeddings, Advances in neural information processing systems 29 (2016).

[27] S. Raza, C. Ding, Fake news detection based on news content and social contexts: a transformer-based approach, International Journal of Data Science and Analytics 13 (2022) 335–362.

[28] S. Raza, D. J. Reji, C. Ding, Dbias: detecting biases and ensuring fairness in news articles, International Journal of Data Science and Analytics (2022). doi:10.1007/

s41060-022-00359-4.
[29] Y. Cai, A. Zimek, G. Wunder, E. Ntoutsi, Power of explanations: Towards automatic debiasing in hate speech detection, in: 2022 IEEE 9th International Conference on Data Science and Advanced Analytics (DSAA), IEEE, 2022, pp. 1–10.
[30] Y. Wang, J. Mansurov, P. Ivanov, J. Su, A. Shelmanov, A. Tsvigun, C. Whitehouse, O. M. Afzal, T. Mahmoud, A. F. Aji, et al., M4: Multi-generator, multi-domain, and multi-lingual black-box machine-generated text detection, arXiv preprint arXiv:2305.14902 (2023).
[31] L. Cheng, S. Ge, H. Liu, Toward understanding bias correlations for mitigation in nlp, arXiv preprint arXiv:2205.12391 (2022).
[32] V. S. Govindarajan, K. Atwell, B. Sinno, M. Alikhani, D. Beaver, J. J. Li, How people talk about each other: Modeling generalized intergroup bias and emotion, in: Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, 2023, pp. 2488–2498.
[33] L. Bauer, K. Gopalakrishnan, S. Gella, Y. Liu, M. Bansal, D. Hakkani-Tur, Analyzing the limits of self-supervision in handling bias in language, arXiv preprint arXiv:2112.08637 (2021).
[34] F. Zhou, Y. Mao, L. Yu, Y. Yang, T. Zhong, Causal-debias: Unifying debiasing in pretrained language models and fine-tuning via causal invariant learning, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2023, pp. 4227–4241.
[35] Y. Lei, R. Huang, L. Wang, N. Beauchamp, Sentence-level media bias analysis informed by discourse structures, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022, pp. 10040–10050. URL: https://aclanthology.org/2022.emnlp-main.682. doi:10.18653/v1/2022.emnlp-main.682.
[36] S. Fu, D. Q. Calley, V. A. Rasmussen, M. D. Hamilton, C. K. Lee, A. Kalla, H. Liu, Gender-based language differences in letters of recommendation, AMIA Summits on Translational Science Proceedings 2023 (2023) 196.
[37] E. F. Sang, F. De Meulder, Introduction to the conll-2003 shared task: Language-independent named entity recognition, arXiv preprint cs/0306050 (2003).
[38] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, et al., The fair guiding principles for scientific data management and stewardship, Scientific data 3 (2016) 1–9.
[39] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).
[40] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, arXiv preprint arXiv:1907.11692 (2019).
[41] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al., Llama 2: Open foundation and fine-tuned chat models, arXiv preprint arXiv:2307.09288 (2023).
[42] S. Raza, B. Schwartz, Entity and relation extraction from clinical case reports of COVID-19: a natural language processing approach, BMC Medical Informatics and Decision Making 23 (2023) 20. doi:10.1186/s12911-023-02117-3.