Research Progress of Causal Inference in Bias Elimination

Limanxi Chen

College of Computer and Big Data, Fuzhou University, Fuzhou, 350000, Fujian, China

Abstract. The natural language processing (NLP) models have recently gained widespread attention and are increasingly being applied to real-world tasks. However, due to the presence of bias, the application of NLP models in specialized fields has led to various issues. This paper introduces various types of biases and, through a comparative analysis of multiple existing methods, explains why previous approaches cannot fundamentally solve the bias problem. Additionally, this paper proposes that using causal inference to eliminate bias is an advanced method, and research teams that reduce bias by building causal relationship models have already studied this approach. This paper conducts a detailed analysis of several cutting-edge studies, exploring the practical application of causal inference in bias elimination and the challenges involved. Experimental results indicate that although causal inference methods have eliminated bias to some extent, further research and optimization are still needed. Finally, this paper summarizes the previous sections and provides an outlook on future research directions.

1 Introduction

With the increasing size of modern pre-trained models, the biases they inherit are also growing, which makes the computational cost of eliminating these biases quite high [1]. Deep learning-based models for natural language processing (NLP) (e.g., large language models like LLMs) rely on the training of massive datasets. These large-scale datasets primarily come from social media, knowledge bases, and search engines. However, data from these sources often carry societal biases related to gender, age, and race due to stereotypes and cultural attributes of the user groups. NLP models can learn these biased patterns from the data, which can further influence the performance of NLP-based applications in specific fields. Moreover, since the field of natural language processing encompasses many sub-tasks, the differences between these tasks make it difficult for models designed to handle specific tasks to generalize across other sub-tasks. This is referred to in this paper as a form of genre bias. Moreover, when addressing real-world social issues, NLP models may still exhibit algorithmic biases, such as the phenomenon of emotional polarization caused by recommendation algorithms on social media platforms.

In previous studies, researchers have employed various methods to eliminate bias. An

Corresponding author: ronghong@ldy.edu.rs

example would be manually labeled samples, but this method presents the issue of being excessively costly [2]. Some researchers have also adopted data augmentation algorithms to achieve bias elimination without altering the biased training data. Moreover, Bias fine-tuning involves adjusting large models with smaller datasets by applying a two-stage fine-tuning process to a pre-trained neural model. First, an iterative algorithm is chosen, followed by the application of a selected debiasing model to achieve effective bias elimination [3]. However, this method is prone to overfitting when sample sizes are limited, reducing the model's generalization ability. Additionally, in downstream natural language processing tasks (e.g., sentiment classification models), Even after removing biases, fine-tuning pre-trained language models can still show or make stereotypes worse [4]. Currently, there is a method called prompt-tuning, which effectively resolves the issues in fine-tuning, such as the significant gap between the objectives of the pre-training and fine-tuning stages, and also the overfitting problems. Experimental teams have proposed using prefix tuning, prompt tuning, and adapter tuning to debias various language processing models, and to some extent, have achieved improvements in time and memory efficiency for debiasing compared to fine-tuning [1]. However, these methods can only partially mitigate the impact of bias, as they suppress correlations rather than improve causality. Recently, the use of causal inference to address biases in models has attracted the attention of many researchers. Casual inference aims to identify and eliminate the occurrence of bias from the perspective of causal relationships, which, intuitively, makes it a more promising approach. Structural Causal Models (SCM) can be used in causal inference to analyze associated factors and identify confounders [5]. This paper provides an overview of causal concepts in Section 2. Section 3 reviews related work on using causal inference to address bias issues, comparing and analyzing their respective advantages and disadvantages. It also presents improvements in effectiveness and bias elimination based on practical experiments. The paper discusses the challenges of applying causal inference to bias issues, analyzes reasons for suboptimal results, and introduces other advanced algorithms that hold promise for solving bias problems, such as prior awareness and capsule networks. This paper summarizes the entire study in Section 4 and provides outline prospects for future related work.

2 Overview of Causal Inference

2.1 Ladder of Causal Relationships

In causal inference, there is an important concept called the causal ladder. The causal ladder is divided into three levels: association, intervention, and counterfactual [6]. First, patterns can be identified through observation in the association layer. This paper uses a classic economic problem as an example, “In a supermarket, what is the likelihood that a customer who buys toothpaste also buys dental floss?” Solving this problem requires observation and prediction based on that data to determine the degree of association between the behaviors of buying toothpaste and buying dental floss. Secondly, the second layer is intervention, which is higher than the association level because it involves not only passive observation but also active action and intervention to change the situation. An example is used in this paper to illustrate: Can my headache be cured if I take aspirin? In this question, the aim is to intervene in the number of aspirin in the human body and observe the changes in headache status, which is another variable, through this intervention. The final level, or the third level, is counterfactual thinking, which is characterized primarily by imagination and reflection. This paper illustrates it with a simple question, “What if I had done... at that time? Why?” Similarly, using the example of headaches and aspirin, this paper raises the following question: Did aspirin cure my headache? Or would my headache have improved if I had not taken aspirin?

This issue illustrates the fundamental approach of counterfactual reasoning, which involves constructing a world that is contrary to reality— a counterfactual world. In this counterfactual world, I have not taken aspirin. By comparing this counterfactual world with the real world, people can then determine the causal relationship between aspirin and the alleviation of headaches.

2.2 Three methods of integration

2.2.1 A→B→C

Controlling B in the link A→B→C can prevent information about A from flowing to C or information about C from flowing to A.

2.2.2 A←B→C

In a cross-connection or mixed connection A ← B → C, there is a factor B, which simultaneously affects both A and C, controlling B can prevent information about A from flowing to C or information about C from flowing to A. In this type of connection, A and C share a common cause B, leading to a spurious correlation between A and C that is non-causal, therefore, by controlling B, the spurious correlation between A and C can be eliminated.

2.2.3 A→B←C

In the collision junction A→B←C, the information flow rules are completely opposite to those of the first two types. Variables A and C are originally independent, so information about A cannot provide any information about C. However, if B is controlled, information will start to flow through the 'pipeline' due to the mediation effect.

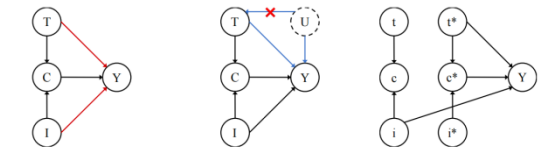
2.3 Do operator, Confounding factor and backdoor criterion

In the context of causal inference, the do operator removes all arrows pointing to X, which prevents any information about X from flowing in non-causal directions. (Randomized treatment has the same effect.) A confounding factor is any factor that causes P(Y|do(X)) to differ from P(Y|X). The backdoor criterion ensures that when estimating causal effects, only direct causal relationships are considered, without being influenced by other confounding factors. Therefore, if a set of variables Z blocks all backdoor paths from X to Y (non-causal paths from X to Y, and does not include any descendant nodes of X, then Z satisfies the backdoor criterion.

3 The application of causal inference in bias elimination

3.1 Application

3.1.1 First application



(a) The proposed causal graph (Factual world). (b) Causal graph with confounder. (c) Counterfactual world.

Fig. 1. Causal Diagram for Fake News Detection [7]

As shown in Fig. 1, the paper presents a causal graph for fake news detection [7] in the essay. In this causal graph, T represents text features, I represents image features, and C represents multimodal features, which can be understood as the fusion of image and text features. Y denotes news labels, and U represents confounding factors. The essay specifically notes that an asterisk (*) denotes a reference value. The figure illustrates that the model contains a backdoor path, this creates a false connection between the text's characteristics and the news categories, thereby leading to the generation of bias. Through the study of the essay, this paper finds that, to minimize psycholinguistic biases as much as possible, the essay employs the method of backdoor adjustment in causal inference and calculates the causal effects during the training phase using calculus P(Y|do(T)), which fundamentally differs from the traditional probability P(Y|T). Using counterfactual reasoning, the paper applies causal interventions by imagining a counterfactual world (c), which is also the third layer mentioned earlier. The essay does not provide the features T and the fused feature C in addition to the image feature I. Instead, a different approach is used: reference values *t and *c are represented, aiming to estimate bias by calculating the immediate impact of I on Y through a cause-and-effect relationship. Another way to deal with bias is by removing this influence from the overall impact on Y.

3.1.2 Second application

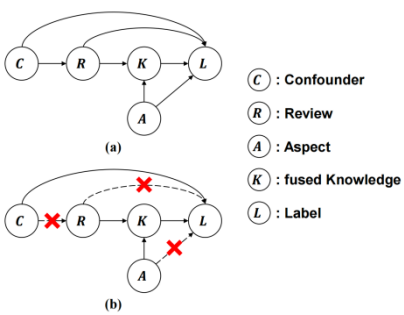


Fig. 2. SCM of ABSA [8]

This paper presents the structure causal model (SCM) of the sentiment analysis model (ABSA) shown in Fig. 2. In this study, the SCM of ABSA is represented as a directed acyclic graph. After analyzing the SCM, the paper derived the formula for causal effects [8]. In this study, the researchers chose backdoor adjustment to eliminate bias in the branches. Considering that SCM only includes R, C, and L, with C satisfying the backdoor criterion, the essay presents:

$$P(L|do(R)) = \sum_c P(L|R,C)P(C) = \sum_c \frac{P(L|P,C)P(C)}{P(R|C)} \quad (1)$$

Based on the previous introduction of the do operator, it can be understood that the do(R) operator signifies a causal intervention, cutting off the direct effect of R on L.

3.1.3 Third application

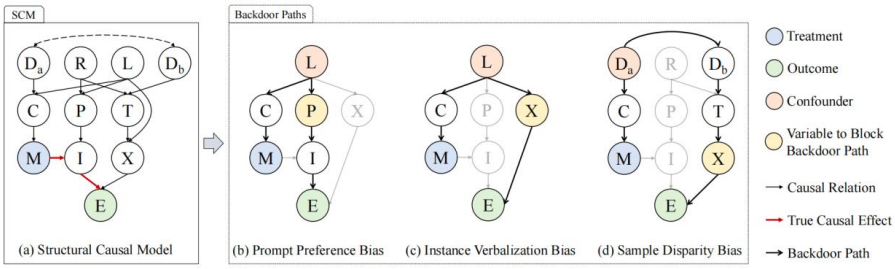


Fig. 3. Causal Framework Diagra [9]

Through backdoor adjustment for causal intervention to reduce bias, this paper aims to achieve more precise, stable, and dependable detection based on the given conditions [9]. The essay does not use a fixed standard method to determine biases but instead highlights subtle, often hard-to-detect implicit risks. It explains how biases can arise from spurious correlations and provides a toolkit for causal analysis, which helps identify and remove biases under certain assumptions, is utilized in the essay by the researchers suggest a causal analysis framework that can be efficiently applied to detect, comprehend, and address biases in prompt-based evaluation methods. The paper provides a causal framework diagram, as shown in Fig. 3, which is sufficient to support the experimental data and conclusions of the study. These four causal diagrams reveal the true causes of biases related to prompt preferences, specific word usage, and variations in sample representation in this study—specifically, the existence of backdoor paths that cause the model to confound the effects between variables, leading to biased impacts on performance.

$$P(E|do(M = m), R = r) = \sum_{p\in P}\sum_{x\in X} P(p,x)P(E|m,r,p,x) \quad (2)$$

The paper provides a specific method: first, choose Z = {P, X} as the set of variables. Then, block all backdoor paths between (M, E) in the SCM, which requires achieving this step through backdoor adjustment. Equation (2) provides an intuitive solution, allowing us to derive the overall causal effect between the PLM and the evaluation from the weighted average effect of all valid prompts and detection data.

To clarify the differences between these three applications, two real-world datasets were used in the first application to verify the effectiveness of the causal framework in eliminating psycholinguistic bias and image bias. Therefore, the focus of the first application lies in

utilizing causal interventions to eliminate biases in both text and images, while enhancing the accuracy of fake news detection through multimodal fusion. The second application compares the performance of incorporating causal inference across multiple benchmark datasets, with a focus on verifying the effectiveness of the new model in executing different tasks. The third application primarily investigates the capabilities of PLMs, explores the biases present in prompt-based assessments of PLMs, and proposes methods to eliminate these biases from a causal perspective.

From analyzing the research methods discussed, this paper shows that modern scientific approaches for causal inference and bias removal use techniques like backdoor adjustment, backdoor path blocking, and weighted effects to get the best results.

3.2 Analysis of Experimental Results

Table 1. First experimental result [7]

Dataset	Methods	Accuracy	Fake news			Real news		
			Precision	Recall	F1	Precision	Recall	F1
Twitter	SpotFake+	0.795	0.622	0.607	0.614	0.856	0.864	0.860
	w/CCD	**0.856***	0.750	0.849	**0.797***	0.920	0.860	**0.889***
	MCAN	0.799	0.980	0.401	0.569	0.770	0.996	0.869
	w/CCD	**0.825***	0.829	0.595	**0.692***	0.824	0.939	**0.878***
	HMCAN	0.831	0.955	0.514	0.668	0.804	0.988	0.887
	w/CCD	**0.874***	0.820	0.792	**0.806***	0.899	0.914	**0.906***
Pheme	SpotFake+	0.815	0.711	0.525	0.604	0.840	0.921	0.879
	w/CCD	**0.823***	0.714	0.574	**0.636***	0.854	0.915	**0.883***
	MCAN	0.834	0.716	0.639	0.675	0.872	0.906	0.889
	w/CCD	**0.839***	0.693	0.721	**0.707***	0.896	0.882	**0.889**
	HMCAN	0.848	0.762	0.705	**0.732**	0.881	0.908	0.894
	w/CCD	**0.859***	0.764	0.689	0.724	0.889	0.921	**0.905***

Table 2. Second experimental result [8]

		REVTGT	REVNON	ADDIFF	ORIGINAL
Laptop	Vanilla	62.45	85.93	76.33	80.41
	DINNER	65.02(↑4.12%)	86.67(↑0.86%)	78.06(↑2.27%)	81.19(↑0.97%)
Restaurant	Vanilla	64.06	82.66	83.48	85.18
	DINNER	70.69(↑10.35%)	83.56(↑1.08%)	86.07(↑3.10%)	87.32(↑2.51%)

As shown in the experimental data and conclusions in Table 1 and 2, it is evident that causal inference does play a role in reducing bias in current research. However, when looking at each experiment in detail, the improvement in model performance is relatively small, often only a few percentage points or even fractions of a percent. It is also clear that the improvements achieved through causal inference methods in the specific domain of bias elimination are not as significant as hoped.

Through the analysis of the aforementioned experimental results, this paper identifies several reasons why causal inference methods perform poorly in the practical application of bias elimination. Firstly, there may be issues with the quality of experimental data in the study,

which might necessitate the introduction of external or more realistic data to explore the model's robustness under real-world conditions. Additionally, experimental data may suffer from sample bias, leading to poorer data conditions. Causal algorithms might not meet the data requirements and may lack sufficient discernment ability [10], which could result in not achieving the desired outcomes. This necessitates modeling and evaluation that incorporates prior knowledge. Next, the experiment suggests that researchers need to establish more reasonable and accurate standards. The research indicates that evaluating the effectiveness of methods in practice is quite challenging. Finally, the application scenarios of the methods in this study are relatively specific and fixed. Future research and practical applications could consider exploring their use in broader tasks. This paper also suggests that research on bias elimination can be conducted through methods such as awareness priors and capsule networks.

4 Conclusion

Reviewing the entire text, this paper summarizes the current status and research progress of causal inference in the elimination of biases. Looking at current technology, most methods for removing bias focus on adjusting data and algorithms. However, they often rely on correlation analysis and can't fully get rid of bias in models. Causal inference offers a new perspective by constructing causal models to fundamentally reveal and eliminate the causes of bias. Despite its theoretical advantages, causal inference still faces challenges in practical applications, especially with large-scale datasets and complex models, where determining and adjusting causal relationships is not straightforward. Additionally, this paper highlights the generalization issues of causal inference across different language processing tasks and the challenges of applying this method to multimodal datasets.

Future research can focus on several directions: firstly, further optimizing the efficiency of causal inference models to meet the demands of large-scale data processing; Secondly, research how to integrate causal inference with existing bias mitigation methods to form a comprehensive solution. Finally, exploring the potential applications of causal inference in other AI fields, such as computer vision and recommendation systems, could be valuable. As research advances, causal inference is expected to provide a more solid foundation for the fairness and transparency of artificial intelligence, helping to create fairer and more trustworthy AI systems. It will also continue to offer new ideas and momentum for the progress in bias elimination within natural language processing models.

References

1. Z. B. Xie, and T. Lukasiewicz, An Empirical Analysis of Parameter-Efficient Methods for Debiasing Pre-Trained Language Models. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 15730–15745, Toronto, Canada. Association for Computational Linguistics. (2023)
2. D. Oba, M. Kaneko, and D. Bollegala. In-Contextual Gender Bias Suppression for Large Language Models. In Findings of the Association for Computational Linguistics: EACL 2024, pages 1722–1742, St. Julian’s, Malta. Association for Computational Linguistics. (2024)
3. L. J. Wang, Y. Y. Li, T. Miller, S. Bethard, and G. Savova, Two-Stage Fine-Tuning for Improved Bias and Variance for Large Pretrained Language Models. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 15746–15761, Toronto, Canada. Association for Computational Linguistics. (2023)

4. F. Zhou, Y. Z. Mao, L. Yu, Y. Yang, and 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), pages 4227–4241, Toronto, Canada. Association for Computational Linguistics. (2023)
5. J. Z. Zhu, S. J. Wu, X. W. Zhang, Y. X. Hou, and Z. Y. Feng. Causal Intervention for Mitigating Name Bias in Machine Reading Comprehension. In Findings of the Association for Computational Linguistics: ACL 2023, pages 12837–12852, Toronto, Canada. Association for Computational Linguistics. (2023)
6. J. Pearl, and D. Mackenzie, The book of why: The new science of cause and effect, (Basic book, New York, 2020)
7. Z. W. Chen, L. M. Hu, W. X. Li, Y. X. Shao, and L. Q. Nie. Causal Intervention and Counterfactual Reasoning for Multi-modal Fake News Detection. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 627–638, Toronto, Canada. Association for Computational Linguistics (2023)
8. J. L. Wu, L. H. Zhang, D. Y. Zhou, and G. Q. Xu, DINER: Debiasing Aspect-based Sentiment Analysis with Multi-variable Causal Inference. In Findings of the Association for Computational Linguistics ACL 2024, pages 3504–3518 August 11-16 (2024)
9. B. X. Cao, H. Y. Lin, X. P. Han, F. C. Liu, and L. Sun, Can Prompt Probe Pretrained Language Models? Understanding the Invisible Risks from a Causal View. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 5796–5808, Toronto, Canada. Association for Computational Linguistics (2023)
10. N. Jain, D. J. Zhang, W. U. Ahmad, Z. J. Wang, F. Nan, X. P. Li, M. Tan, Ramesh Nallapati, Baishakhi Ray, Parminder Bhatia, Xiaofei Ma, and Bing Xiang. ContraCLM: Contrastive Learning For Causal Language Model. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 6436–6459, Toronto, Canada. Association for Computational Linguistics (2023)