Chain-of-Thought Prompting with Causal Intervention for Multimodal Aspect-based Sentiment Analysis

Maolin Li[1], Zhuopan Yang[1], Haoran Xie[2], Lap-Kei Lee[3]✉, Fu Lee Wang[3], Yi Yu[4], and Zhenguo Yang[1]✉

[1] Guangdong University of Technology
{gdutlml,zhuopanyang}@gmail.com, yzg@gdut.edu.cn
[2] Lingnan University
hrxie@ln.edu.hk
[3] Hong Kong Metropolitan University
{lklee,pwang}@hkmu.edu.hk
[4] Hiroshima University
yiyu@hiroshima-u.ac.jp

Abstract. Aspect-based sentiment analysis identifies the polarity of targeted words or visual regions expressing latent opinions, typically via large language models (LLMs) on semantic and context understanding. Anyway, LLMs tend to converge on plenty of underlying correlations that are spurious and insignificant. To this end, we propose a chain-of-thought prompting with causal intervention (CPCI) to exploit the causality correlations for multimodal aspect-based sentiment analysis. More specifically, CPCI introduces stochastic perturbations on the original sentences to generate augmented ones as correlation candidates. Furthermore, backdoor adjustment is devised in the causal intervention (CI) module by using the inverse attention matrix of transformer as confounder in the structural causal model. In particular, a random sampling is conducted on the confounder to be integrated with the attention features of the transformer to eliminate the influence of spurious correlations. Finally, a chain-of-thought prompting (CP) module is devised on the augmented sentences without spurious correlations to infer the underlying properties from the latent intent of opinion to implicit sentiment polarity gradually. Extensive experiments conducted on two public datasets show that CPCI outperforms state-of-the-art approaches.

Keywords: Multimodal aspect-based sentiment analysis · Causal intervention · Large language model.

1 Introduction

Multimodal aspect-based sentiment analysis aims to identify the sentiment polarity underlying the specific aspects of sentence-image pairs, while the sentiments of latent opinions are typically expressed in an implicit manner. Recently, the

Chain-of-Thought (CoT) using large language models (LLMs) shows promising and impressive effects on inferring the polarity of implicit sentiment. In the context of CoT with LLMs, eliminating spurious correlations typically overused by LLMs resulting from the powerful associative abilities remains a challenging issue. As shown in Figure 1 (a), the aspect word "GMMusicHub" is recognized to be positive as the LLMs overinterpret the conceivable correlations between "Good job" and "GMMusicHub". To this end, quite a few works focus on debiasing the spurious correlations underlying the uneven distribution of training data.

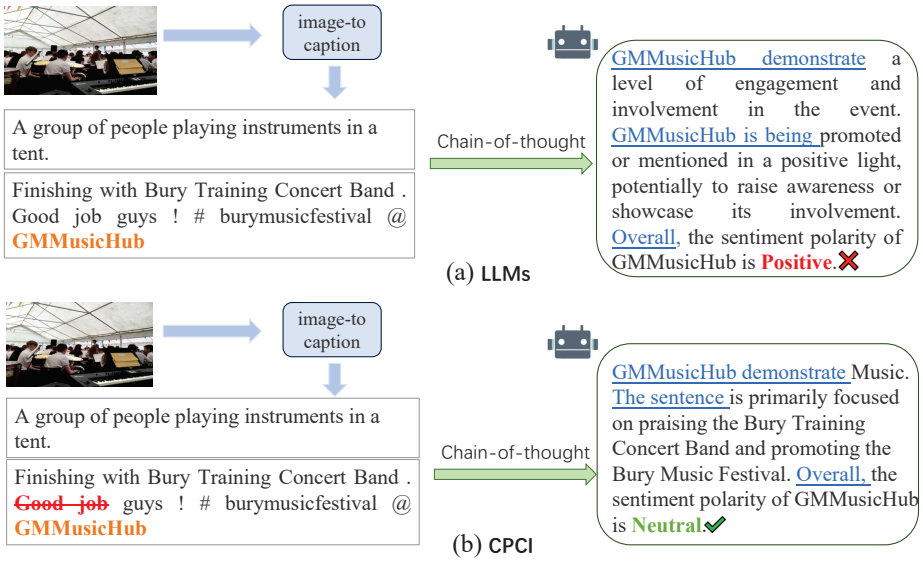

Fig. 1. LLMs tend to rely on shortcuts of the data, while CPCI removes spurious correlations.

In terms of methodologies, the works on eliminating spurious correlations can be divided into two categories, i.e., data augmentation and causal intervention. More specifically, the works on data augmentation typically augment the data samples to enhance the generalization of the models. For example, Ouyang et al. [5] employed a multi-head attention network to achieve augmentation in the latent space. In contrast, the works on causal intervention aim to curb dependence on shortcuts to eliminate the causal effects. For instance, Zhu et al. [14] designed a causal graph to formulate the position bias as direct to mitigate position bias.

In this paper, we propose a chain-of-thought prompting with causal intervention to intensify causality correlations with data augmentation for multimodal aspect-based sentiment analysis. More specifically, CPCI generates augmented sentences as correlation candidates on original sentences by introducing stochastic perturbations. Moreover, backdoor adjustment is devised in the causal inter-

vention (CI) module, which uses the inverse attention matrix of transformer as the confounder in structural casual model. We conduct a random sampling on the confounder to integrate with the attentions features of transformer, which aims to eliminate the influence of spurious correlations. Finally, we devise a chain-of-thought prompting (CP) module on the augmented data to obtain the potential properties from the latent intent of opinion to implicit sentiment polarity. We evaluate our method on multiple datasets, and the experimental results consistently demonstrate the superior performance of our approach. To the best of our knowledge, it is the first work with causal intervention in the context of multimodal aspect-based sentiment analysis.

The main contributions of this paper are summarized as follows:

- We devise a causal interventions (CI) module which consists of stochastic perturbation and backdoor adjustment, aiming to eliminate the spurious correlations between sentences and the polarity of sentiments.
- We devise a chain-of-thought prompting (CP) strategy on the augmented sentences by eliminating spurious correlations, aiming to infer the latent properties from underlying intent of opinions to implicit sentiment polarity gradually.
- We conduct extensive experiments on two public datasets including Twitter-2015 and Twitter-2017, manifesting the effectiveness of CPCI against state-of-the-art approaches.

2 Preliminary

2.1 Structural Causal Model

As shown in Figure 2 (a),the causal effect is limited to the direct path from T to Y, while the total effect includes all paths from T to Y. In the best-case scenarios, the total effect and the causal effect can be regarded as the same. In practice, confounder serves as a common cause of the treatment and outcome by the backdoor path, denoted as $T \leftarrow C \rightarrow Y$. Consequently, the treatment-outcome relationship may be concealed well by the spurious correlation [6] from T to Y generated by C, distracting the models while training.

2.2 Structural Causal Model

The purpose of the model is to capture the true causality between T and Y. Nevertheless, the information flow between T and Y must pass through the confounder because of the backdoor path $T \leftarrow C \rightarrow Y$. Therefore, we adopt backdoor adjustment to cut off the backdoor path and use $P(Y \mid do(T))$ to replace $P(Y \mid T)$, as shown in Figure 2 (b). The backdoor adjustment assumes that we can observe and stratify the confounder, which exploit the do-calculus on input text by estimating $P_B(Y \mid T) = P(Y \mid do(T))$, where the subscript B denotes the backdoor adjustment on the SCM. Subsequently, the marginal probability is invariant under the intervention, because C will remain unchanged

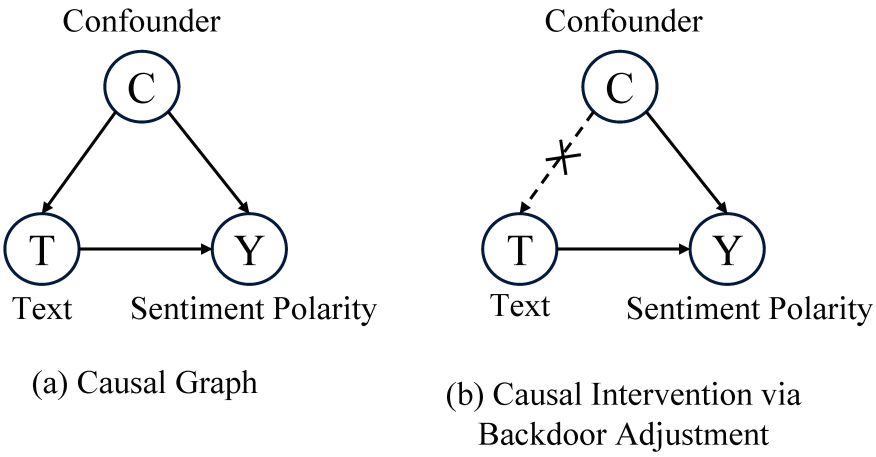

Fig. 2. Causal Intervention Graph.

when cutting the link between C and T, denoted as $P(C) = P_B(C)$. T and C are independent after backdoor adjustment, which are denoted as $P_B(C \mid T) = P_B(C)$. Based on these conclusions, we can formulate as below,

$$\begin{aligned} P(Y \mid do(T)) &= P_B(Y \mid T) \\ &= P_B(Y \mid T, C) P_B(C \mid T) \\ &= P_B(Y \mid T, C) P_B(C) \\ &= P(Y \mid T, C) P(C) \end{aligned} \tag{1}$$

where $P(C)$ denotes the prior probability of C.

3 Methodology

3.1 Overview of the Framework

As shown in Figure 3, we propose a chain-of-thought prompting with causal intervention (CPCI) framework for MABSA, which consists of two modules, i.e., Causal Intervention (CI) and Chain-of-thought Prompting (CP) modules. Specifically, the CI module conducts stochastic perturbations on the original sentences to generate augmented data, and uses the inverse attention matrix of the transformer to devise a backdoor adjustment as a confounder. Furthermore, we select the sentence with reduction of spurious correlation by evaluating the perturbation effect as input for LLMs. Finally, we design a CP module to gradually infer the underlying properties from the latent intent of opinions to implicit sentiment polarity.

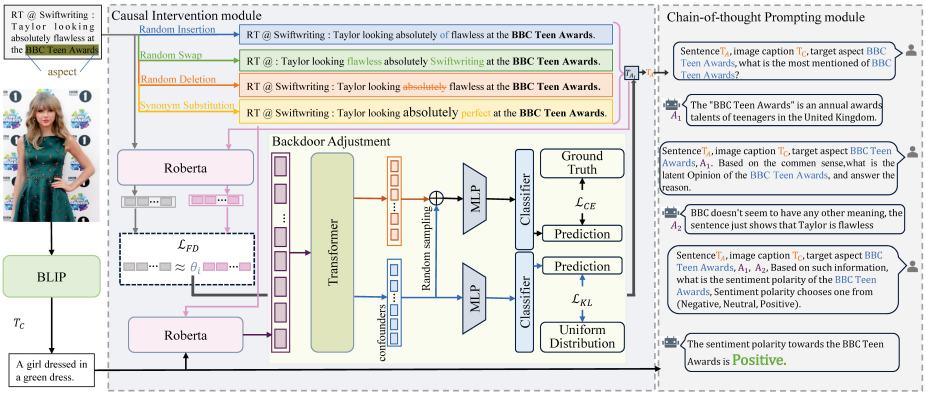

Fig. 3. Overview of the CPCI framework.

3.2 Causal Intervention Module

Stochastic Perturbations Sentiment words show the natural sentiment intensity and polarity of sentences, thus the unavoidable confounders within the text for sentiment polarity would confuse models for predictions. To this end, we use nlpaug[5] to design stochastic perturbations for data augmentation to decrease the spurious correlations in text. More specifically, stochastic perturbations consist of inserting words randomly, swapping words randomly, deleting words randomly with probability m, and substituting the words by using WordNet synonyms. In particular, all the stochastic perturbations do not include the operation of the aspect sentiment words as follows,

$$T_{A_i} = F(T, S_i) \tag{2}$$

where T_{A_i} denotes the augmented text candidate, $F(\cdot)$ denotes the different stochastic perturbation, i denotes the different perturbation methods. We formulize the evaluations of the perturbation effect for original data and augmentation data. More specifically, we extract the features of the original text and the augmented text, and devise a self-learning parameter layer to build a relationship between the features as follows,

$$T \approx \theta_i \cdot T_{A_i} \tag{3}$$

where θ_i denotes the learning parameter corresponding to different perturbations. To obtain the accurate value of θ_i, we use the pre-training model Roberta [4] to extract the features as follows,

$$\theta_i = argmin \sum || \theta \cdot f_{Roberta}(T) - f_{Roberta}(T_{A_i}) || \tag{4}$$

where $f_{Roberta}(\cdot)$ denotes the text encoder of the pre-training model Roberta, θ is a self-learning parameter, $||\cdot||$ denotes the absolute value symbol. In addition,

[5] https://github.com/makcedward/nlpaug

we devise a feature distance (FD) loss to remain the semantics of original text, bounding the randomness of stochastic perturbations as follows,

$$\mathcal{L}_{FD} = \sum || f_{Roberta}(T) - \theta_i \cdot f_{Roberta}(T_{A_i}) ||$$ (5)

where $\mathcal{L}_{FD}$ denotes the feature distance loss.

Backdoor Adjustment We fit the deep learning model in a parameterized method by implementing Eq. (1). Given the image I, we use the pre-training model BLIP [2] to generate the caption T_C for the image. The original text T_{A_i}, the image caption T_C and the aspect word T_W consist of a sequence of tokens $(t_1, t_2, \ldots, t_n)$, $(b_1, b_2, \ldots, b_n)$ and $(a_1, a_2, \ldots, a_n)$, respectively. Furthermore, we concatenate them as the input $X = (T_{A_i}, [SEP], T_C, [SEP], T_W)$ and extract them features by using the pre-trained model Roberta as follows,

$$F_x = f_{Roberta}(X)$$ (6)

where $f_{Roberta}(\cdot)$ denotes the text encoder of the pre-training model Roberta. Moreover, we use zero-initialize $Q_0 \in \mathbb{R}^{k \times h}$ as the queries in the cross-attention module of the transformer, where k is the number of categories, h is the feature hidden size. Each decoder layer L updates the querier Q_{L-1} from its previous layer. Furthermore, we use the inverse attention matrix to obtain the confounders as follows,

$$Q_L = softmax(\frac{\bar{Q}_{L-1}\bar{F}_x}{\sqrt{h}})F_x$$ (7)

$$\widehat{Q}_L = (1 - softmax(\frac{\bar{Q}_{L-1}\bar{F}_x}{\sqrt{h}}))F_x$$ (8)

where Q_L, $\widehat{Q}_L$ denote the causal features and confounding features respectively, $\bar{Q}_{L-1}$ and $\bar{F}_x$ denote the position encodings of F_x and Q_{L-1}, respectively. The proposed backdoor adjustment aims to making confounding features irrelevant of classification. In particular, we use a predefined uniform distribution to push its prediction equally to all categories using a KL-Divergence (KL) loss as follows,

$$y_c = \varphi(MLP(\widehat{Q}_L))$$ (9)

$$\mathcal{L}_{KL} = -\frac{1}{D}\sum KL(y_{unif}, y_c)$$ (10)

where φ denotes the classifier, y_c denotes the prediction, D denotes the number of training data. $KL(\cdot)$ denotes the KL-Divergence, y_{unif} denotes a predefined uniform distribution.

A critical issue of backdoor adjustment is to combine confounding and causal features, and we approximate backdoor adjustment by a random sample operation. More specifically, we stratify the confounding features, randomly add them to the causal features, and feed them into MLP and classifier to obtain the predictions as follows,

$$\widetilde{p_c} = \varphi(MLP(Q_L) + MLP(\widetilde{Q_L}))$$ (11)

where $\widetilde{Q_L}$ denotes the stratified features obtained from Q_L, $\widetilde{p_c}$ denotes the prediction. Furthermore, we adopt cross-entropy (CE) loss guided by causal inference to minimize the effects of confounders as follows,

$$\mathcal{L}_{CE} = \widetilde{p_c} log(p) + (1 - \widetilde{p_c}) log(1 - p) \tag{12}$$

where p denotes the ground truth label, $\mathcal{L}_{CE}$ denotes the cross-entropy loss. In terms of the loss terms, CPCI consists of cross-entropy loss, KL- Divergence loss and feature distance loss. In summary, the objective function of the proposed CPCI approach is as follows,

$$\mathcal{L}_{ALL} = \mathcal{L}_{CE} + \alpha \mathcal{L}_{KL} + \beta \mathcal{L}_{FD} \tag{13}$$

where α, β are hyper-parameters.

3.3 Chain-of-thought Prompting Module

It is a challenge to predict implicit sentiment polarity due to the opinion cues appear implicitly and obscurely. Consequently, recognizing implicit sentiment necessitates the capacity to use common sense and multi-hop reasoning to deduce the hidden purpose of opinion. Inspired by the recent progress on chain-of-thought, we design a three-step prompting module step-by-step to obtain the sentiment polarity.

1) Step 1. Giving the text T_A selected by the minimum θ_i among the different augmented text, the image caption T_C, and the aspect word T_w, we resort to LLMs by constructing the following template,

Sentence T_A ,image caption T_C, target aspect T_w , what is the most mentioned of T_w ?

2) Step 2. We resort to LLMs to answer the latent intent of opinion towards the aspect T_w by constructing the following template,

Sentence T_A , image caption T_C, target T_w, A_1. Based on the commen sense, what is the latent opinion of the T_w and answer the reason.

3) Step 3. We resort to LLMs to infer the sentiment polarity as the answer by constructing the following template,

Sentence T_A, image caption T_C, target T_w, A_1, A_2. Based on such information,what is the sentiment polarity of the T_w, Sentiment polarity chooses one from (Negative, Neutral, Positive).

4 Experiments

4.1 Main Results

Table 1 presents the performance of different approaches on Twitter-2015 and Twitter-2017, from which we can observe that our proposed CPCI performs the best in terms of all metrics on both two datasets. More specifically, CPCI outperforms M2DF with the second best results by 0.85% and 1.10% on accuracy,

Table 1. The experimental results on two MABSA datasets. The best results are marked in bold.

Methods	Twitter-15		Twitter-17	
	Acc	F1	Acc	F1
MIMN[8]	71.84	65.69	65.88	62.99
ESAFN[11]	73.38	67.37	67.83	64.22
SaliencyBERT[7]	77.03	72.36	69.69	67.19
TomBert[10]	77.15	71.75	70.34	68.03
EF-CapTrBERT[1]	78.01	73.25	69.77	68.42
FITE[9]	78.49	73.90	70.90	68.70
FITE-DE-Large[9]	78.80	74.80	73.90	73.00
ITM[12]	78.27	74.19	72.61	71.97
VLP-MABSA[3]	78.60	73.80	73.80	71.80
M2DF[13]	78.90	74.80	74.30	73.00
CPCI (ours)	**79.75**	**75.90**	**78.36**	**76.43**

as well as 4.06% and 3.43% points on F1 score. The reason is that CPCI infers implicit sentiment polarity by reducing spurious correlations. Meanwhile, CPCI with the chain-of-thought has the ability to infer the implicit sentiment polarity. In particular, CPCI improves the performance significantly on Twitter-2017, which can be explained that quite a few sentences in Twitter-2017 require external knowledge as support, and LLMs are able to compensate for this limitation owing to the availability of prior knowledge.

4.2 Ablation Study

Effects of the causal intervention module. As mentioned in Section 3.2, we develop the CI module to eliminate the spurious correlations of text, and adopt CP module to infer sentiment polarity. We denote CPCI without CI module as w/o CI. As shown in Table 2, we can observe that CPCI outperforms CPCI without CI module on all the metrics. The reason is that the CI module provides the CP module with the ability to eliminate spurious correlation text, which is beneficial to the understanding of LLMs on the real relationship between aspect words and sentiment polarity. Especially on the Twitter-2017 dataset, CPCI outperforms CPCI without CI module by 4.70% on accuracy and 3.44% on F1 score due to a large amount of prior knowledge in Twitter-2017 dataset, which needs to reduce the plenty of spurious and insignificant underlying correlations.

Effectiveness of the loss functions of CPCI. As mentioned in Section 3.2, we develop three loss functions including CE, KL and FD in the CI module to eliminate the influence of spurious correlations. More specifically, CE denotes the cross-entropy loss as depicted in Eq. (12), KL denotes the KL-Divergence loss as depicted in Eq. (10) and FD denotes the feature distance loss as depicted in Eq. (5). As shown in Table 3, we can observe that CPCI with CE loss is sig-

nificantly beneficial to eliminating spurious correlations. Meanwhile, by utilizing the KL and FD losses, CPCI is able to capture the causal correlations, hence obtaining a significant performance improvement.

Table 2. Ablation study on CPCI. The best results are marked in bold.

Methods	Twitter-15		Twitter-17	
	Acc	F1	Acc	F1
w/o CI	77.09	72.12	73.66	72.99
CPCI	**79.75**	**75.90**	**78.36**	**76.43**

Table 3. Effectiveness of different loss terms of CPCI. The best results are marked in bold.

Methods	Twitter-15		Twitter-17	
	Acc	F1	Acc	F1
CE	77.53	72.59	74.23	72.70
CE + KL	79.17	75.45	77.92	74.19
CE + FD	78.59	73.46	75.31	71.09
CE + KL + FD	**79.75**	**75.90**	**78.36**	**76.43**

5 Conclusion

In this paper, we propose a chain-of-thought prompting with causal intervention to eliminate the spurious correlations for MABSA. More specifically, we design stochastic perturbations in the CI module to generate an augmented ones as correlation candidates on original sentences. In order to remove the confounder effect in the structural causal model, we design a backdoor adjustment that uses the inverse attention matrix as confounder of the transformer. Furthermore, a random sampling is conducted on the confounder to be integrated with the attention features of the transformer. In particular, we design a chain-of-thought prompting (CP) module that gradually infers the underlying properties from the latent purpose of opinion to implicit sentiment polarity, utilizing the best augmented sentence that spurious correlation elimination.

6 Acknowledgments

This work is partly supported by the the Guangdong Basic and Applied Basic Research Foundation (No.2024A1515010237), and RGC of HKSAR, China, under Grant UGC/FDS16/E13/23.

References

1. Khan, Z., Fu, Y.: Exploiting bert for multimodal target sentiment classification through input space translation. In: Proceedings of the 29th ACM international conference on multimedia. pp. 3034–3042 (2021)
2. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International conference on machine learning. pp. 12888–12900 (2022)
3. Ling, Y., Yu, J., Xia, R.: Vision-language pre-training for multimodal aspect-based sentiment analysis. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022. pp. 2149–2159 (2022)
4. Liu, Z., Lin, W., Shi, Y., Zhao, J.: A robustly optimized BERT pre-training approach with post-training. In: Li, S., Sun, M., Liu, Y., Wu, H., Liu, K., Che, W., He, S., Rao, G. (eds.) Chinese Computational Linguistics - 20th China National Conference, CCL. vol. 12869, pp. 471–484 (2021)
5. Ouyang, J., Feng, S., Wang, B., Yang, Z.: Pseudo dense counterfactual augmentation for aspect-based sentiment analysis. Neurocomputing **561**, 126869 (2023)
6. Pearl, J.: Interpretation and identification of causal mediation. Psychological methods **19**(4), 459 (2014)
7. Wang, J., Liu, Z., Sheng, V., Song, Y., Qiu, C.: Saliencybert: Recurrent attention network for target-oriented multimodal sentiment classification. In: Pattern Recognition and Computer Vision: 4th Chinese Conference, PRCV 2021. pp. 3–15 (2021)
8. Xu, N., Mao, W., Chen, G.: Multi-interactive memory network for aspect based multimodal sentiment analysis. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 371–378 (2019)
9. Yang, H., Zhao, Y., Qin, B.: Face-sensitive image-to-emotional-text cross-modal translation for multimodal aspect-based sentiment analysis. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. pp. 3324–3335 (2022)
10. Yu, J., Jiang, J.: Adapting BERT for target-oriented multimodal sentiment classification. In: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019. pp. 5408–5414 (2019)
11. Yu, J., Jiang, J., Xia, R.: Entity-sensitive attention and fusion network for entity-level multimodal sentiment classification. IEEE/ACM Transactions on Audio, Speech, and Language Processing **28**, 429–439 (2019)
12. Yu, J., Wang, J., Xia, R., Li, J.: Targeted multimodal sentiment classification based on coarse-to-fine grained image-target matching. In: IJCAI. pp. 4482–4488 (2022)
13. Zhao, F., Li, C., Wu, Z., Ouyang, Y., Zhang, J., Dai, X.: M2DF: multi-grained multi-curriculum denoising framework for multimodal aspect-based sentiment analysis. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023. pp. 9057–9070 (2023)
14. Zhu, J., Wu, L., Wu, S., Zhang, X., Hou, Y., Feng, Z.: Causal MRC: mitigating position bias based on causal graph. In: Abbadi, A.E., Dobbie, G., Feng, Z., Chen, L., Tao, X., Shao, Y., Yin, H. (eds.) Database Systems for Advanced Applications. DASFAA 2023 International Proceedings. vol. 13922, pp. 251–266 (2023)