Causal Inference and Counterfactual Text-Debiasing for Multimodal Aspect-Based Sentiment Analysis

Haolong Zheng[1*], Xinyue Wang[2], Tenglong Zheng[3]
[1]School of Communication and Electronic Engineering, Shandong Normal University, Jinan, China
[2]School of Economics, Shandong University of Finance and Economics, Jinan, China
[3]School of Mechanical Engineering, Hebei University of Science and Technology, Shijiazhuang, China
Email: 2023025808@stu.sdnu.edu.cn, zegaln@163.com, 2500511106@stu.hebust.edu.cn

***Abstract*—Multimodal Aspect-Based Sentiment Analysis (MABSA) aims to jointly extract aspect terms from image-text pairs and accurately predict their sentiment polarity. Existing research indicates that the text modality often dominates the sentiment prediction process. However, this modal imbalance tends to induce models to over-rely on statistical co-occurrences between inputs and labels during training, rather than learning true semantic reasoning. Consequently, models erroneously establish spurious correlations between textual features and sentiment labels, which severely impairs their generalization capabilities on out-of-distribution (OOD) data. To address this challenge, we propose a Causal Inference and Counterfactual Text-Debiasing (CCT) framework. The framework first constructs a causal model to dissect the complex causal mechanisms linking multimodal features to sentiment labels. Subsequently, we introduce an auxiliary textual branch combined with an attention mechanism to explicitly model and pinpoint specific textual features that induce spurious correlations. During the inference phase, CCT utilizes counterfactual reasoning to precisely disentangle biases stemming from textual shortcuts from the multimodal Total Effect (TE), while preserving and enhancing valid semantic cues essential for sentiment prediction. Extensive experiments on the Twitter2015 and Twitter2017 datasets demonstrate that the CCT framework can be integrated into existing mainstream models, significantly boosting their generalization performance and robustness.**

***Index Terms*—Multimodal aspect-based sentiment analysis, Causal inference, Counterfactual, Spurious correlation**

I. Introduction

Multimodal Aspect-Based Sentiment Analysis (MABSA) aims to jointly extract aspect terms from text-image pairs while accurately predicting their associated sentiment polarity. Due to its fine-grained complexity, MABSA has emerged as a prominent research focus within the field of multimodal learning. Distinct from traditional unimodal approaches constrained to text,social media has enabled large-scale multimodal content,offering diverse and realistic data sources for aspect-level sentiment understanding. By effectively fusing visual and textual modalities, MABSA enhances recognition robustness, particularly in complex scenarios involving implicit sentiment, sarcasm, or insufficient textual context. As a result, models can more faithfully approximate how humans express and interpret emotions in multimodal interactions.

*Corresponding author

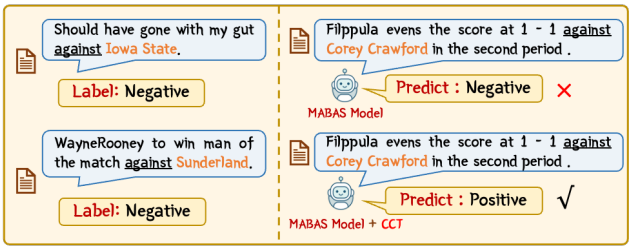


Fig. 1. Examples of Textual Spurious Correlations in MABSA

Current research in MABSA predominantly focuses on text-image modeling and aspect-level sentiment recognition. Early approaches largely relied on traditional machine learning; for instance, Yang et al. [1] employed Support Vector Machines (SVMs) to enhance predictive accuracy. With advances in deep learning, these techniques have been increasingly adopted in MABSA. Contemporary methodologies primarily fall into two categories: Transformer-based and Graph-based approaches. Both paradigms aim to achieve aspect-level sentiment prediction through the joint analysis of text and visual modalities. For example, Zhou et al. [2] integrated an aspect-aware attention module with Graph Convolutional Networks (GCNs) for modeling. Zhang et al. [3] utilized an attention score matrix within GCNs to enhance node representation, thereby optimizing the MABSA task. Despite the efficacy of these solutions, a critical challenge remains: models often tend to capture correlations between inputs and outputs rather than genuine causal relationships. This reliance on spurious correlations severely limits the generalization capabilities of the models. Consequently, integrating causal inference into MABSA is particularly necessary. Existing research indicates that the text modality often dominates MABSA tasks, leading models to establish spurious correlations between textual features and sentiment labels. As illustrated in Fig.1, the word ”against” exhibits a disproportionately high co-occurrence frequency with the ”Negative” label in the Twitter2015 dataset. This statistical bias misleads the model into learning a misconception, incorrectly capturing a dependency relationship. This reliance on shortcuts severely impairs generalization capabilities. During testing, when encountering samples containing ”against” that

actually express positive sentiment, the model erroneously predicts "Negative" polarity by following this spurious path. To mitigate this, early studies attempted to reduce bias at the source by constructing balanced datasets like WEBEmo, but such data-driven approaches incur high annotation costs. Addressing this challenge, scholars have increasingly integrated causal inference paradigms into machine learning. For instance, the CLUE model utilizes causal intervention to disentangle the direct non-causal influence of the text modality, thereby mitigating spurious correlations. However, it is crucial to note that valid semantic clues within the text modality are essential for aspect-level sentiment recognition and must be preserved during the debiasing process.

To address these challenges, we propose a Causal Inference and Counterfactual Text-Debiasing (CCT) framework for MABSA. We construct a Causal Model to dissect the dual pathways of text influence on predictions: i) a direct pathway that induces spurious correlations, acting as a non-causal shortcut; and ii) an indirect pathway that interacts with other modalities to leverage complementary signals and recover more reliable text semantics. During the inference stage, CCT achieves text debiasing via counterfactual reasoning. The core mechanism involves disentangling and subtracting the direct textual bias from the Total Effect (TE), while preserving valid semantic contributions to jointly guide aspect-level prediction. This design mitigates spurious correlation bias and strengthens multimodal complementarity, improving robustness and generalization.

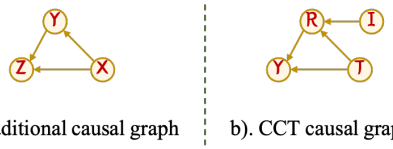
The main contributions of this work are summarized as follows:

1) We investigate spurious correlations between the text modality and sentiment labels in MABSA from a causal perspective.
2) We propose a Causal Inference and Counterfactual Text-Debiasing (CCT) framework. By leveraging counterfactual reasoning mechanisms, CCT effectively mitigates spurious associations between text features and labels, significantly reducing the adverse impact of data bias on model predictions.
3) We conduct extensive comparative experiments on the Twitter2015 and Twitter2017 benchmarks. The empirical results demonstrate the superiority of CCT over prior state-of-the-art methods in both mitigating bias and improving model performance.

II. Related Work

A. Multimodal aspect-based sentiment analysis

Multimodal Aspect-Based Sentiment Analysis (MABSA) aims to extract aspect terms from multimodal data and accurately predict their sentiment polarity by incorporating visual cues. Unlike traditional unimodal approaches, MABSA jointly models text semantics and visual context. This integration offers discriminative cues for complex scenarios involving semantic sparsity, ambiguity, or implicit sentiment. Existing methodologies generally fall into two categories: Transformer-based and Graph-based approaches. Both strive for fine-grained modeling of cross-modal feature interactions. For instance, Wang et al. [4] leveraged attention mechanisms to capture modal correlations. Xiao et al. [5] introduced a text-visual alignment strategy to bridge semantic gaps. In graph-based domains, Zhou et al. [2] employed Graph Convolutional Networks (GCNs) for interaction modeling. Li et al. [6] advanced this via a DualGCN architecture to enhance feature representation. Despite these advancements, significant limitations persist in current research. To address these challenges, we introduce a causal inference paradigm to analysis the complex causal mechanisms linking modal features to sentiment labels. Specifically, we employ counterfactual reasoning to mitigate spurious correlations between text and labels. This approach reduces data bias and enhances predictive performance.



a). Traditional causal graph b). CCT causal graph

Fig. 2. Comparison between a traditional causal graph and our proposed causal graph.

B. Causal effect

As a theoretical cornerstone of statistics, causal inference elucidates complex variable interactions through rigorous mathematical formalism and distinct criteria. The Potential Outcome Framework and Structural Causal Models (SCM) have emerged as dominant paradigms for causal analysis, providing the foundational theoretical support for the field. With the advancement of deep learning, there is a growing consensus that the absence of causal modeling capabilities creates a bottleneck, restricting model interpretability and robustness. In this context, Counterfactual Reasoning [1], a core component of causal inference, has garnered significant attention. Counterfactual reasoning has been explored in a range of settings, including visual dialogue [7], adversarial learning [8], scene graph generation, and image sentiment recognition [9].These applications suggest that counterfactual reasoning can enhance model reasoning and generalization in practice.

III. Methodology

A. Task definition

Given a sample $S = \{I, U\}$, where I is an image and $U = (w_1, w_2, ..., w_n)$, is a sentence with n words, MABSA aims to produce an output sequence L that encodes all aspect terms and their associated sentiment polarities. Specifically, $L = [a_1^s, a_1^e, y_1, ..., a_i^s, a_i^e, y_i, ..., a_k^s, a_k^e, y_k]$,where (a_i^s, a_i^e) denotes the start and end indices of the i-th aspect terms in U, y_i is its sentiment polarity, and k is the number of aspects.

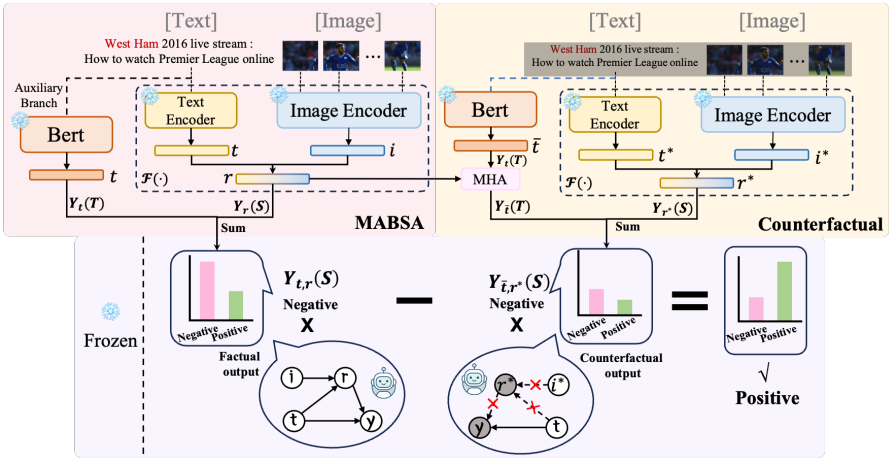


Fig. 3. The overview of our proposed Causal Inference and Counterfactual Text-Debiasing for MABSA

B. Causal inference

Causal models [1] commonly use directed acyclic graphs (DAGs) to represent assumed causal relations among variables in a structured manner. A causal graph provides a compact representation of the variables of interest and the directed relations assumed between them. In a DAG, each variable corresponds to a node, and each directed edge encodes a direct causal influence from parent to child. Accordingly, such a DAG is also referred to as a causal graph. Formally, a causal graph is denoted as $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \{X, Y, Z_1, \ldots, Z_n\}$ is the set of variables (nodes) and $\mathcal{E}$ is the set of directed edges. The edge set $\mathcal{E}$ encodes prior causal assumptions about how variables influence one another. For illustration, Fig.2 a) shows a traditional causal graph with three variables. The effect of X on Z can arise via a direct path $X \rightarrow Z$ and an indirect path mediated by Y, i.e., $X \rightarrow Y \rightarrow Z$. Our model, illustrated in Fig.2 b), specifies a causal graph over the text–image input (T, I), the multimodal representation R, and the sentiment label Y.We analyze the implied causal pathways as follows:

- $T \rightarrow Y$ (Direct path): This path represents the direct dependency of sentiment labels on the text modality. In the context of MABSA, this path often encapsulates spurious correlations, serving as a "shortcut" during inference. While computationally efficient, it may mislead the model and lead to predictions in the wrong direction.
- $(T, I) \rightarrow R \rightarrow Y$ (Indirect path): This path affects the prediction via the multimodal representation R. By aligning text with the corresponding image, this interaction mitigates the impact of text bias, yielding a robust semantic representation R that supports more accurate and generalizable sentiment prediction.

C. Counterfactual reasoning

Our core idea is to suppress the spurious correlations component along the $T \rightarrow Y$ path while preserving the informative textual contribution, thereby improving the reliability of aspect-level sentiment prediction. Under the causal inference framework, we formulate the causal structure in MABSA as follows:

$$Y_{t,r}(S) = Y(T = t, R = (T = t, I = i)) \tag{1}$$

where, $Y_{t,r}(S)$ denotes the predicted outcome under the observed text $T = t$ and multimodal representation $R = r$, which may be affected by spurious correlations in the text modality.

To quantify the overall causal impact, we define the Total Effect (TE), denoted as E_{TE}, as follows:

$$E_{TE} = Y_{t,r}(S) - Y_{t^*,r^*}(S) \tag{2}$$

where, (t^*, r^*) denotes the corresponding inputs are removed, and the missing components are replaced with a randomly initialized vector.

Subsequently, we analyze the Natural Direct Effect (NDE) caused by spurious correlations in the text modality. Specifically, we aim to extract a spurious text component $Y_{\bar{t}}(T)$ that directly contributes to the text driven prediction $Y_t(T)$.Given the complementarity between text and images, we treat the fused multimodal prediction $Y_r(S)$ as a more reliable proxy of sentiment semantics in the true context. Because the spurious component and the context grounded sentiment representation can diverge substantially, their similarity is expected to be low. Accordingly, we introduce a multi-head attention (MHA) encoder to explicitly capture low similarity regions between text features and multimodal features. The formulation is given as follows:

$$\text{head}_i = \text{Attn}(Q, K, V) \tag{3}$$

$$\text{Attn}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{4}$$

$$\text{Output} = \text{MHA}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^o \tag{5}$$

where, $\mathrm{MHA}(\cdot)$ denotes the multi head attention mechanism. $Q \in \mathbb{R}^{n\times d_q}$, $K \in \mathbb{R}^{n\times d_k}$, and $V \in \mathbb{R}^{n\times d_v}$ are the query, key, and value matrices, respectively. d_q, d_k, and d_v denote their feature dimensions, and h is the number of attention heads.

We use the multimodal representation Y_r as the query and the text representation Y_t as both the key and the value, yielding a pseudo correlated text representation:

$$Y_{\bar{t}}(T) = \mathrm{MHA}(Y_r, Y_t, Y_t) \tag{6}$$

$$Y_{\bar{t},r^*}(S) = Y(T = \bar{t}, R^* = (T = t^*, I = i^*)) \tag{7}$$

The Natural Direct Effect (NDE), denoted as E_{NDE}, is defined as:

$$E_{NDE} = Y_{\bar{t},r^*}(S) - Y_{t^*,r^*}(S) \tag{8}$$

To remove the spurious correlation component captured by the direct text pathway, we subtract NDE from TE to obtain the Total Indirect Effect (TIE), denoted as E_{TIE}:

$$E_{TIE} = Y_{t,r}(S) - Y_{\bar{t},r^*}(S) \tag{9}$$

At inference time, we use E_{TIE} as the debiased signal for prediction, as it excludes spurious correlations.

D. CCT algorithm framework

As illustrated in the Fig.3, the proposed CCT architecture consists of two components: an auxiliary text only predictor $Y_t(T) = \mathrm{BERT}(T = t)$ and a multimodal MABSA predictor $Y_r(S) = \mathcal{F}(T = t, I = i)$, where $\mathcal{F}(\cdot)$ denotes a chosen MABSA backbone (for example, AoM).

We then fuse the outputs of the two predictors by additive combination, which leverages complementary information and yields the final prediction score. We then fuse the outputs of the two predictors by additive combination, which leverages complementary information and yields the final prediction score. This fusion produces the factual prediction $Y_{t,r}(S)$ and the counterfactual prediction $Y_{\bar{t},r^*}(S)$, the formulation is given as follows:

$$\begin{aligned} Y_{t,r}(S) &= \mathrm{SUM}\big(Y_t(T) + Y_r(S)\big) \\ &= \log\sigma\big(Y_t(T) + Y_r(S)\big) \end{aligned} \tag{10}$$

$$\begin{aligned} Y_{\bar{t},r^*}(S) &= \mathrm{SUM}\big(Y_{\bar{t}}(T) + Y_{r^*}(S)\big) \\ &= \log\sigma\big(Y_{\bar{t}}(T) + Y_{r^*}(S)\big) \end{aligned} \tag{11}$$

where, $\sigma(\cdot)$ is the sigmoid function.

E. Training phase

During training, we optimize a weighted cross entropy objective:

$$\mathcal{L}_{ce} = \alpha \times CE\big(Y_{t,r}(S), y\big) + \beta \times CE\big(Y_{\bar{t},r^*}(S), y\big) \tag{12}$$

where, α and β are balancing coefficients, and y is the ground truth sentiment label for the corresponding aspect term. We additionally include a KL divergence to regularize the discrepancy between $Y_{t,r}(S)$ and $Y_{\bar{t},r^*}(S)$:

$$\mathcal{L}_{KL} = \mathrm{KL}\big(Y_{t,r}(S), Y_{\bar{t},r^*}(S)\big) \tag{13}$$

The overall objective is:

$$\mathcal{L} = \sum_{(t,i,y)\in R} (\mathcal{L}_{ce} + \mathcal{L}_{KL}) \tag{14}$$

F. Inference phase

At inference time, we use the TIE to infer the sentiment polarity of each aspect term, since it excludes the pseudo correlated component from the direct text pathway:

$$E_{TIE} = Y_{t,r}(S) - Y_{\bar{t},r^*}(S) \tag{15}$$

IV. Experiment

A. Datasets

To assess the performance of our proposed CCT framework on MABSA tasks, we conduct experiments using two benchmark MABSA datasets: Twitter2015 and Twitter2017 [10].

To ensure rigorous evaluation, we adopt two complementary strategies. First, we train and validate the model under the independent and identically distributed (IID) setting for each task to assess performance stability under standard conditions. Second, we evaluate the model on an out of distribution (OOD) test set, where the statistical association between words and sentiment labels is weakened compared with the IID training data. This OOD evaluation allows us to examine whether the CCT framework is robust to spurious correlations and can reliably predict the sentiment polarity of the true aspect terms.

B. Evaluation metrics

Following previous work, we evaluate the model on MABSA using Micro-F1 score (F1), Precision (P), and Recall (R). We use the same metrics for Multimodal Aspect Term Extraction (MATE) . For Multimodal Aspect-oriented Sentiment Classification (MASC), we report Accuracy (Acc) and F1.

C. Baselines

1) Baselines for text-based ABSA: we compare our proposed CCT framework with several classic text-based ABSA methods widely adopted as baselines (SPAN [11], D-GCN [12], RoBERTa [13], BART [14])

2) Baselines for multimodal ABSA: we evaluate CCT against representative MABSA baselines (UMT+TomBERT [15], RpBERT-collapse [16], UMT-collapse [17], OSCGA-collapse [18], JML [19], CapTrRoBERTa [20], VLP-MABSA [21], CMMT [22], MOCOLNet [23], AoM [2])

V. Experimental results

A. Comparison with baselines

To evaluate the proposed CCT framework, we report the results on two datasets for the MABSA task in Table I. Table I summarizes the comparative results under the out of distribution (OOD) setting. As shown in Table I, conventional MABSA models such as JML, CMMT, and AoM exhibit clear performance degradation on the OOD test set. This observation indicates that when the test distribution deviates from the training distribution, these methods become more vulnerable

TABLE I
OOD PERFORMANCE COMPARISON ACROSS BASELINES ON TWITTER2015 AND TWITTER2017 FOR THE MABSA TASK.

Model	Twitter2015			Twitter2017		
	P	R	F1	P	R	F1
Text-based						
SPAN [11]	53.7	53.9	53.8	59.6	61.7	60.6
D-GCN [12]	58.3	58.8	59.4	64.1	64.2	64.1
RoBERTa [13]	61.8	65.3	63.5	65.5	66.9	66.2
BART [14]	62.9	65.0	63.9	65.2	65.6	65.4
Multimodal						
TomBERT+UMT [15]	58.4	61.3	59.8	62.3	62.4	62.4
RpBERT-collapse [16]	49.3	46.9	48.0	57.0	55.4	56.2
UMT-collapse [17]	60.4	61.6	61.0	60.0	61.7	60.8
OSCGA-collapse [18]	63.1	63.7	63.2	63.5	63.5	63.5
VLP-MABSA [21]	65.1	68.3	66.6	66.9	69.2	68.0
MOCOLNet [23]	66.3	67.8	67.1	67.2	68.7	67.9
CapTrRoBERTa [20]	60.6	66.1	63.2	67.1	67.4	67.3
JML [19]	65.0	63.2	64.1	66.5	65.5	66.0
JML+CCT	**68.2**	**67.9**	**67.3**	**69.1**	**68.3**	**68.8**
CMMT [22]	64.6	68.7	66.5	67.6	69.4	68.5
CMMT+CCT	**67.4**	**70.2**	**69.7**	**69.9**	**71.3**	**70.0**
AoM [2]	67.9	69.3	68.6	68.4	71.0	69.7
AoM+CCT	**69.4**	**71.6**	**70.2**	**72.5**	**74.1**	**72.8**

TABLE II
OOD PERFORMANCE COMPARISON ACROSS BASELINES ON TWITTER2015 AND TWITTER2017.

Methods	Twitter2015			Twitter2017		
	P	R	F1	P	R	F1
UMT-collapse	77.8	81.7	79.7	86.7	86.8	86.7
OSCGA-collapse	81.7	82.1	81.9	90.2	90.7	90.4
VLP-MABSA	83.6	87.9	85.7	90.8	92.6	91.7
JML	83.6	81.2	82.4	92.0	90.7	91.4
JML+CCT	**85.3**	**83.1**	**84.7**	**92.4**	**91.2**	**92.5**
CMMT	83.9	88.1	85.9	92.2	**93.9**	**93.1**
CMMT+CCT	**85.5**	**89.7**	**87.1**	**92.7**	93.4	92.9
AoM	84.6	87.9	86.2	91.8	**92.8**	92.3
AoM+CCT	**86.3**	**89.2**	**88.5**	**92.3**	92.5	**92.4**

to spurious associations between textual cues and sentiment labels, which in turn harms generalization. In contrast, incorporating CCT consistently improves performance under the OOD setting, suggesting stronger robustness under distribution shift. This gain can be attributed to CCT, which decomposes the influence of the text modality from a causal perspective, suppresses the spurious component, and preserves informative textual semantics that are beneficial for aspect-level sentiment prediction.

B. Ablation study

To analyze the contribution of each module in CCT to MABSA, we conduct a series of ablation studies on two datasets, with results reported in Table III. We adopt JML, CMMT, and AoM as backbone models and construct ablated variants on top of each backbone for comparison.

Removing CCT leads to consistent performance drops across backbones, indicating that CCT plays a key role in improving robustness. This suggests that the causal based debiasing mechanism suppresses spurious text driven shortcuts while preserving informative textual semantics that are beneficial for aspect level sentiment prediction.

Moreover, the full model consistently outperforms its ablated variants, demonstrating the effectiveness of the overall framework design. These results further highlight the importance of multimodal information for aspect level sentiment analysis, where complementary cues from images and text provide richer evidence and support more reliable predictions.

When the image modality is removed, all three backbones exhibit a clear performance decrease, confirming that visual information makes a prominent contribution to aspect-level sentiment recognition.

TABLE III
ABLATION STUDY ON TWITTER2015 AND TWITTER2017.

Methods	Twitter2015			Twitter2017		
	P	R	F1	P	R	F1
JML+CCT	**68.2**	**67.9**	**67.3**	**69.1**	**68.3**	**68.8**
w/o CCT	66.7	65.9	65.2	67.4	66.1	67.6
w/o Image	64.9	64.4	63.8	65.2	64.7	65.0
w/o MABSA model	52.4	51.9	51.6	53.8	53.2	53.7
CMMT+CCT	**67.4**	**70.2**	**69.7**	**69.9**	**71.3**	**70.0**
w/o CCT	65.8	67.4	66.1	66.3	68.5	67.9
w/o Image	64.2	67.8	66.3	67.7	68.3	66.7
w/o MABSA model	54.6	55.7	54.9	53.1	55.4	54.2
AoM+CCT	**69.4**	**71.6**	**70.2**	**72.5**	**74.1**	**72.8**
w/o CCT	68.1	70.4	68.9	71.3	72.8	71.6
w/o Image	66.3	68.7	67.2	69.5	71.8	69.3
w/o MABSA model	54.7	56.2	54.9	56.4	57.3	56.5

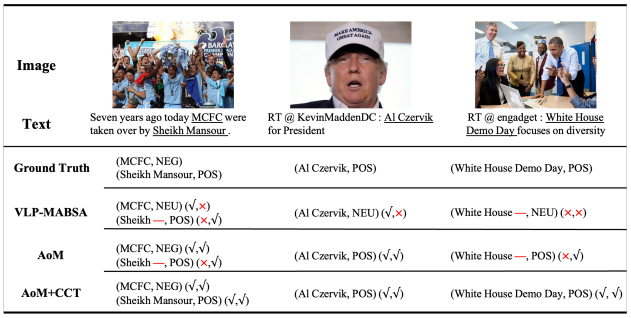


Fig. 4. Three case studies comparing predictions from VLP-MABSA, AoM, and AoM+CCT.

C. Case Study

To further investigate how the model works, we present three representative examples in Fig.4, comparing the ground truth with predictions from VLP-MABSA, AoM, and AoM+CCT. In the first case, VLP-MABSA misidentifies the sentiment polarity of ”MCFC” and fails to extract the second aspect term, while AoM also struggles with aspect extraction, indicating that multimodal fusion alone is insufficient to alleviate text bias. In the second case, VLP-MABSA extracts the correct aspect but produces an incorrect sentiment prediction due to spurious textual correlations. In the third case, both VLP-MABSA and AoM only partially extract the complex

aspect term "White House," leading to degraded sentiment prediction. Overall, these cases demonstrate that the CCT framework enables AoM+CCT to achieve more stable and accurate predictions by suppressing spurious correlations while preserving genuinely useful semantic information, thereby improving robustness in challenging scenarios.

VI. Conclusion

This paper proposes a Causal Inference and Counterfactual Text-Debiasing (CCT) framework to address spurious correlations between the text modality and sentiment labels, thereby reducing their adverse impact on model predictions. CCT incorporates causal and counterfactual reasoning to suppress spurious correlations while preserving and strengthening informative textual semantics that support aspect-level sentiment prediction. This design improves the model ability to capture sentiment signals under distribution shift and diverse data conditions, leading to more robust classification performance. We validate CCT through extensive experiments on the Twitter2015 and Twitter2017 benchmarks. Results show that CCT yields consistent gains under both standard evaluation and out of distribution settings, indicating improved robustness. CCT improves over conventional MABSA baselines and alleviates the impact of dataset bias on generalization.

This work focuses on the causal pathways linking the text modality to sentiment labels. Future work will extend counterfactual analysis to the image modality to characterize richer causal mechanisms across modalities.

References

[1] Q. YANGal and C. LIUb, "Application of twin objective function svm in sentiment analysis," *Machine Learning and Artificial Intelligence: Proceedings of MLIS 2020*, vol. 332, p. 221, 2020.

[2] R. Zhou, W. Guo, X. Liu, S. Yu, Y. Zhang, and X. Yuan, "AoM: Detecting aspect-oriented information for multimodal aspect-based sentiment analysis," in *Findings of the Association for Computational Linguistics: ACL 2023*. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 8184–8196.

[3] Z. Zhang, Z. Zhou, and Y. Wang, "Ssegcn: Syntactic and semantic enhanced graph convolutional network for aspect-based sentiment analysis," in *Proceedings of the 2022 conference of the North American Chapter of the association for computational linguistics: human language technologies*, 2022, pp. 4916–4925.

[4] K. Wang, W. Shen, Y. Yang, X. Quan, and R. Wang, "Relational graph attention network for aspect-based sentiment analysis," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, Jul. 2020, pp. 3229–3238.

[5] L. Xiao, X. Wu, J. Xu, W. Li, C. Jin, and L. He, "Atlantis: Aesthetic-oriented multiple granularities fusion network for joint multimodal aspect-based sentiment analysis," *Information Fusion*, vol. 106, p. 102304, 2024.

[6] R. Li, H. Chen, F. Feng, Z. Ma, X. Wang, and E. Hovy, "Dual graph convolutional networks for aspect-based sentiment analysis," in *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)*, 2021, pp. 6319–6329.

[7] X. Zhang, F. Zhang, and C. Xu, "Multi-level counterfactual contrast for visual commonsense reasoning," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 1793–1802.

[8] H. Ni, J. Song, X. Zhu, F. Zheng, and L. Gao, "Camera-agnostic person re-identification via adversarial disentangling learning," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 2002–2010.

[9] D. Yang, K. Yang, M. Li, S. Wang, S. Wang, and L. Zhang, "Robust emotion recognition in context debiasing," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 12 447–12 457.

[10] J. Yu, J. Jiang, and R. Xia, "Entity-sensitive attention and fusion network for entity-level multimodal sentiment classification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 429–439, 2019.

[11] M. Hu, Y. Peng, Z. Huang, D. Li, and Y. Lv, "Open-domain targeted sentiment analysis via span-based extraction and classification," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 537–546.

[12] G. Chen, Y. Tian, and Y. Song, "Joint aspect extraction and sentiment analysis with directional graph convolutional networks," in *Proceedings of the 28th international conference on computational linguistics*, 2020, pp. 272–279.

[13] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[14] H. Yan, J. Dai, T. Ji, X. Qiu, and Z. Zhang, "A unified generative framework for aspect-based sentiment analysis," in *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)*, 2021, pp. 2416–2429.

[15] J. Yu, J. Jiang, L. Yang, and R. Xia, "Improving multimodal named entity recognition via entity span detection with unified multimodal transformer." Association for Computational Linguistics, 2020.

[16] L. Sun, J. Wang, K. Zhang, Y. Su, and F. Weng, "Rpbert: a text-image relation propagation-based bert model for multimodal ner," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, no. 15, 2021, pp. 13 860–13 868.

[17] J. Yu, J. Jiang, and R. Xia, "Entity-sensitive attention and fusion network for entity-level multimodal sentiment classification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 429–439, 2019.

[18] Z. Wu, C. Zheng, Y. Cai, J. Chen, H.-f. Leung, and Q. Li, "Multimodal representation with embedded visual guiding objects for named entity recognition in social media posts," in *Proceedings of the 28th ACM International conference on multimedia*, 2020, pp. 1038–1046.

[19] X. Ju, D. Zhang, R. Xiao, J. Li, S. Li, M. Zhang, and G. Zhou, "Joint multi-modal aspect-sentiment analysis with auxiliary cross-modal relation detection," in *Proceedings of the 2021 conference on empirical methods in natural language processing*, 2021, pp. 4395–4405.

[20] Z. Khan and Y. Fu, "Exploiting bert for multimodal target sentiment classification through input space translation," in *Proceedings of the 29th ACM international conference on multimedia*, 2021, pp. 3034–3042.

[21] Y. Ling, J. Yu, and R. Xia, "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)*. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 2149–2159.

[22] L. Yang, J.-C. Na, and J. Yu, "Cross-modal multitask transformer for end-to-end multimodal aspect-based sentiment analysis," *Information Processing & Management*, vol. 59, no. 5, p. 103038, 2022.

[23] J. Mu, F. Nie, W. Wang, J. Xu, J. Zhang, and H. Liu, "Mocolnet: A momentum contrastive learning network for multimodal aspect-level sentiment analysis," *IEEE Transactions on Knowledge and Data Engineering*, vol. 36, no. 12, pp. 8787–8800, 2023.