DINER: Debiasing Aspect-based Sentiment Analysis with Multi-variable Causal Inference

Jialong Wu[♠♢*] **Linhai Zhang**[♠♢*] **Deyu Zhou**[♠♢†] **Guoqiang Xu**[♡]
[♠] School of Computer Science and Engineering, Southeast University, Nanjing, China
[♢]Key Laboratory of New Generation Artificial Intelligence Technology and Its
Interdisciplinary Applications (Southeast University), Ministry of Education, China
[♡] SANY Group Co., Ltd.
{jialongwu, lzhang472, d.zhou}@seu.edu.cn
xuguoqiang-2012@hotmail.com

Abstract

Though notable progress has been made, neural-based aspect-based sentiment analysis (ABSA) models are prone to learn spurious correlations from annotation biases, resulting in poor robustness on adversarial data transformations. Among the debiasing solutions, causal inference-based methods have attracted much research attention, which can be mainly categorized into causal intervention methods and counterfactual reasoning methods. However, most of the present debiasing methods focus on single-variable causal inference, which is not suitable for ABSA with two input variables (*the target aspect* and *the review*). In this paper, we propose a novel framework based on multi-variable causal inference for debiasing ABSA. In this framework, different types of biases are tackled based on different causal intervention methods. For the review branch, the bias is modeled as indirect confounding from context, where backdoor adjustment intervention is employed for debiasing. For the aspect branch, the bias is described as a direct correlation with labels, where counterfactual reasoning is adopted for debiasing. Extensive experiments demonstrate the effectiveness of the proposed method compared to various baselines on the two widely used real-world aspect robustness test set datasets. [1]

1 Introduction

Aspect-Based Sentiment Analysis (ABSA) aims to classify the polarity of the sentiment (*e.g.*, positive, negative, or neutral) towards a specific aspect of a sentence (*e.g.*, *bugers* in the review "*Tasty bugers, and crispy fries.*") (Hu and Liu, 2004; Jiang et al., 2011; Vo and Zhang, 2015; Zhang et al., 2016, 2022). Most ABSA methods solve the task as an input-output mapping problem based on high-capacity neural networks and pre-trained language models (Wang et al., 2018; Huang and Carley, 2018; Bai et al., 2020). Though remarkable progress has been made, it is demonstrated that these state-of-the-art models are not robust in data transformation where simply reversing the polarity of the target results in over 20% drop in accuracy (Xing et al., 2020).

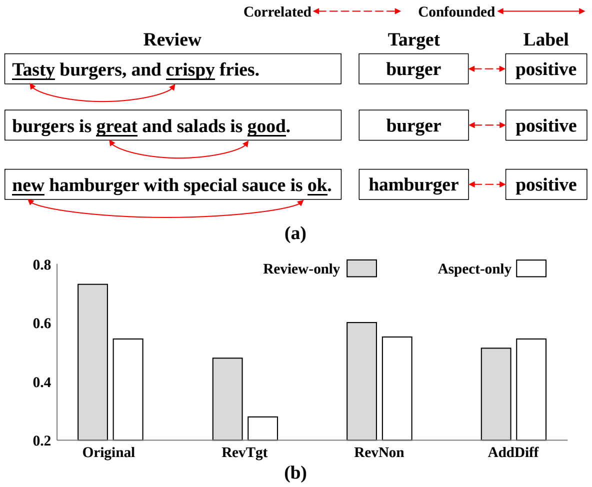

Figure 1: (a) Examples are taken from the SemEval 2014 Restaurant test set. (b) REVTGT denotes reversing the polarity of the target aspect, REVNON denotes reversing the polarity of the non-target aspect, and ADDDIFF denotes adding another non-target aspect with different polarity.

A reasonable explanation is that neural networks trained with the Stochastic Gradient Descent algorithm are vulnerable to annotation biases and learn the shortcuts instead of the underlying task (Xing et al., 2020). As shown in Figure 1 (a), over 50.0% of targets have only one kind of polarity label in the widely used SemEval 2014 Laptop and Restaurant datasets (Pontiki et al., 2014). For 83.9% and 79.6% instances in the test sets, the sentiments of the target aspect and all non-target aspects are the same. Therefore, it is easy for end-to-end neural models to learn such spurious correlations and

[*] Equal Contribution.

[†] Corresponding Author.

[1] Our code and results will be available at https://github.com/callanwu/DINER.

Findings of the Association for Computational Linguistics ACL 2024, pages 3504–3518
August 11-16, 2024 ©2024 Association for Computational Linguistics

make predictions solely based on target aspects or sentiment words describing non-target aspects.

To avoid learning spurious correlations, recent methods focus on debiasing, which can be categorized into argumentation-based methods (Wei and Zou, 2019; Lee et al., 2021), reweight training-based methods (Schuster et al., 2019; Karimi Mahabadi et al., 2020) and causal inference-based methods (Niu et al., 2021; Liu et al., 2022b). Among them, causal inference attracts much research interest for its theoretical-granted property and little modification to the existing learning paradigm. Niu et al. (2021) proposed a debiasing method for the language bias in the vision question answering task by performing counterfactual reasoning. Liu et al. (2022b) employed backdoor adjustment-based intervention for mitigating the context bias in object detection. Recent attempts have been made to solve various biases in natural language processing tasks, including natural language understanding (Tian et al., 2022), implicit sentiment analysis (Wang et al., 2022), and fact verification (Xu et al., 2023).

However, most causal inference-based debiasing methods are based on single-variable causal inference, which is not appropriate for ABSA with two input variables. As shown in Figure 1 (a), there are two types of biases in ABSA. The target aspects A are often directly correlated with the polarity labels L, while the sentiment words for targets in the review R are often indirectly confounded with the non-targets C. To further investigate the difference between aspect-related biases and review-related biases, a simple experiment is conducted by training two probing models with only review or aspect as input. As shown in Figure 1 (b), the aspect-only model has similar performances on the original and the adversarial test set except REVTGT where spurious correlations learned in the training set are flipped, while the review-only model performs differently on four test variants. It might suggest that the biases in the aspect branch are direct and simple, while the biases in the review branch are indirect and complicated, which poses a challenge.

To tackle the above challenge, we propose **D**ebias **IN** Asp**E**ct and **R**eview (**DINER**) based multi-variable causal inference for debiasing ABSA. To be more specific, as illustrated in Figure 2, the unbiased prediction is obtained by calculating the total indirect effect of the target aspect and the review of the polarity label, which is further decomposed and estimated by the hybrid causal intervention method. For the $R \rightarrow L$ branch, a backdoor adjustment intervention is employed to mitigate the indirect confounding between the target sentiment words in the review and the context. For the $A \rightarrow L$ branch, a counterfactual reasoning intervention is employed to remove the direct correlation between the target and the label. Extensive experiments on two widely used real-world robustness test benchmark datasets show the effectiveness of our framework.

Overall, our contributions can be summarized as follows:

- A novel framework is proposed for debiasing ASBA based on multi-variable causal inference. As far as we know, we are the first to uncover and analyze the bias problem in ABSA using multi-variable causal inference.
- A hybrid intervention method is constructed by combining backdoor adjustment and counterfactual reasoning.
- The detailed evaluation demonstrates that the proposed method empirically advances the state-of-the-art baselines.

2 Related Work

Our work is mainly related to two lines of research, described as follows.

2.1 Aspect-Based Sentiment Analysis

ABSA has garnered significant research attention in recent years. Early works focus on feature engineering with manual-construction sentiment lexicons and syntactic features, and rule-based classifiers are adopted to make predictions (Jiang et al., 2011; Kiritchenko et al., 2014). With the development of neural networks and word embedding techniques, neural-based models have dominated the area with architectures such as LSTM, CNN, Attention mechanisms, Capsule Network (Tang et al., 2016a; Wang et al., 2016; Xue and Li, 2018; Jiang et al., 2019). Recent advances in pre-trained language models such as BERT (Devlin et al., 2019) have shifted the paradigm again (Zhang et al., 2022), where most recent models take pre-trained models as backbones (Xu et al., 2019; Hou et al., 2021; Cao et al., 2022). However, ABSA still faces challenges on robustness datasets, and it is precisely such tasks that our approach targets.

2.2 Causal Inference-based Debiasing

Causal inference (Pearl, 1995, 2009) has been widely employed for debiasing in various fields, including computer vision, recommendation, and natural language processing (Niu et al., 2021; Zhang et al., 2021b; Tian et al., 2022). The main methods employed consist of counterfactual reasoning and causal intervention. Niu et al. (2021) proposed to remove the language bias in vision question answering by subtracting the results of a counterfactual language-only model from the results of a vanilla language-vision model. Following this work, counterfactual reasoning is widely applied to debiasing the spurious correlation between input and label in tasks including natural language understanding (Tian et al., 2022), machine reading comprehension (Guo et al., 2023; Zhu et al., 2023) and fact verification (Xu et al., 2023). Liu et al. (2022b) proposed to de-confound the object from the context in object detection with backdoor adjustment, where an inverse probability weight approximation is made to estimate the *do*-operator. Another way to estimate the *do*-operator is known as normalized weighted geometrical mean (NWGM), which is firstly adopted in image caption by Liu et al. (2022a). Following this line of work, backdoor adjustment-based debiasing has widely been explored in tasks including named entity recognition (Zhang et al., 2021a) and multi-modal fake news detection (Chen et al., 2023). Some methods also employ other causal inference techniques, including instrument variable (Wang et al., 2022) and colliding effects (Zheng et al., 2022). However, most of the present debiasing methods focus on debiasing a single input variable, while we are the first to debias two input variables in ABSA simultaneously.

3 Methods

In this section, we will introduce the proposed method, **DINER**, in detail. First, we will define the Structural Causal Model (SCM) of ABSA and derive the formula of causal effect step by step. Then, we will formulate how to estimate the components in the causal effect formula with backdoor adjustment and counterfactual reasoning. Finally, we will introduce the training and inference processes.

3.1 Structural Causal Model of ABSA

The SCM of ABSA, which is formulated as a directed acyclic graph, is shown in Figure 2 (a). The nodes in the SCM denote causal variables, and the edges denote causalities between two nodes (*e.g.*, $X \to Y$ means X causes Y). Then we will discuss the rationale behind how this SCM is built:

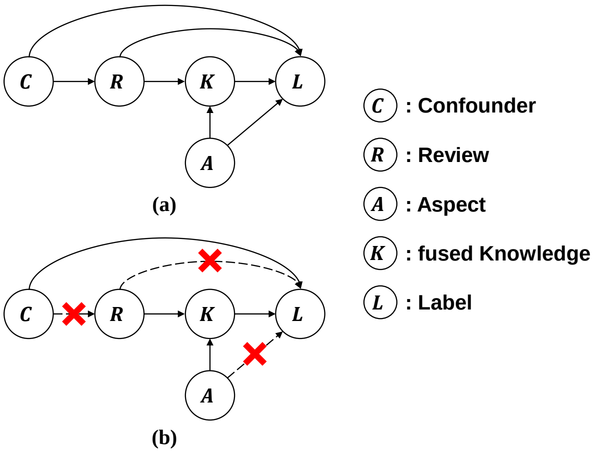

Figure 2: (a) SCM of the proposed method. (b) The desired situation for ABSA, the dotted line means the causalities are blocked.

- $R \to K \leftarrow A$. The prediction of ABSA is dependent on both review R and aspect A. Therefore, a fused knowledge node K is caused by both R and A.
- $K \to L$. The label L is caused by the fused knowledge K, which is the desired causal effect of ABSA.
- $R \to L \leftarrow A$. The label L is also directly affected by review R and aspect A, where the spurious correlation comes from and should be removed.
- $C \to R$ and $C \to L$. The confounder C (the prior context knowledge) caused R and L simultaneously, where the annotation biases come from. For example, most reviews contain positive descriptions for multiple types of food, which will encourage the model to make predictions without identifying the target.

It is worth noticing that we do not add the edge $C \to A$ or $R \to A$. Because we believe the choice of aspect A is made by the annotators and not restricted by the context C or review R.

With the SCM defined, we can derive the formula of causal effect. As shown in Figure 2 (b), the desired situation for ABSA is that the edges that bring biases are all blocked, and the prediction is based on aspect A and review R solely through the

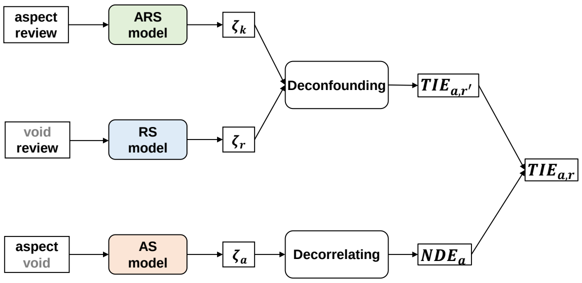

Figure 3: The framework of the proposed method.

fused knowledge K. With the language of causal inference, the prediction should be made on:

$$TIE_{a,r} = TE_{a,r} - NDE_r - NDE_a + IE_{a,r} \quad (1)$$

where $TIE_{a,r}$ denotes the Total Indirect Effect (TIE) from A and R on L, $TE_{a,r}$ denotes the Total Effect (TE), NDE denotes the Natural Direct Effect (NDE), and $IE_{a,r}$ denotes the Interaction Effect (IE) between A and R. The total effect TE contains all causal effects from A and R on L, inducing the biases, while the natural direct effect (NDE) only measures the direct causal effect between two variables, which can be regarded as the bias-only effect. Therefore, subtracting NDE_a and NDE_r from $TE_{a,r}$ will results in the unbiased causal effect from A and R on L, which is the total indirect effect $TIE_{a,r}$. It is worth noticing that since there is no causality between A and R, the value of the interaction effect $IE_{a,r}$ can be set to 0.

Based on the defintion of TE and NDE (Niu et al., 2021):

$$TE_{a,r} = L_{a,r,k} - L_{a^*,r^*,k^*} \quad (2)$$

$$NDE_a = L_{a,r^*,k^*} - L_{a^*,r^*,k^*} \quad (3)$$

$$NDE_r = L_{a^*,r,k^*} - L_{a^*,r^*,k^*} \quad (4)$$

where L denotes the prediction and x^* denotes variable x is set to be void, we can have:

$$\begin{aligned} TIE_{a,r} &= L_{a,r,k} - L_{a^*,r,k^*} - L_{a,r^*,k^*} + L_{a^*,r^*,k^*} \\ &= TIE_{a,r'} - NDE_a \end{aligned} \quad (5)$$

where $r^{'}$ denotes the debiased review, obtained after the process of deconfounding.

3.2 Deconfounding the Review Branch with Backdoor Adjustment

Based on Eq. (5), we can estimate each component and obtain an unbiased prediction. However, as R and L are indirectly confounded with context C, it is not easy to calculate $L_{a,r,k}$ and L_{a^*,r,k^*}. Therefore, we debias the review R first.

$$L_{a,r,k} = \Psi(\zeta_a, \zeta_{r'}, \zeta_k) \quad (6)$$

where ζ_k denotes the logit of the softmax layer, $\Psi(\cdot)$ denotes the fusion function, specially $\zeta_{r'}$ denotes the debiased output based on R.

There are mainly three types of causal intervention methods based on causal inference, the backdoor adjustment, the front-door adjustment, and the instrument variable adjustment. However, the front-door adjustment requires a mediator variable between input and output, which is not applicable in our SCM (Zhang et al., 2024a,b). The instrument variable adjustment involves building up an extra instrument variable in SCM, which makes the already complex SCM even more complex (Wang et al., 2022). So we choose the backdoor adjustment for debiasing the review branch. Consider the SCM only contains R, C, and L, C satisfies the backdoor criterion, and we can have:

$$\begin{aligned} P(L|do(R)) &= \sum_{\boldsymbol{c}} P(L|R,C)P(C) \\ &= \sum_{\boldsymbol{c}} \frac{P(L,R|C)P(C)}{P(R|C)} \end{aligned} \quad (7)$$

where the $do(R)$ operator denotes a causal intervention that severs the direct effect of R on L.

A common workaround is the application of the Normalized Weighted Geometric Mean (NWGM) (Xu et al., 2015) to approximate the effects of the do-operator. Our approach adopts an Inverse Probability Weighting (IPW) (Pearl, 2009) perspective, which provides a novel lens through which to approximate the infinite sampling of $(l, r)|c$ as shown in Eq. (7).

In a finite dataset, the observable instances of (l, r) for each unique c are limited. Consequently, the number of c values considered in our equation equates to the available data samples rather than the theoretically infinite possibilities of c. Backdoor adjustment bridges the gap between the confounded and de-confounded models, allowing us to treat samples from the confounded model as if they were drawn from the de-confounded scenario. This leads to an approximation:

$$\begin{aligned} P(L|do(R=r)) &\approx \widetilde{P}(L,R|C=c) \\ &\approx \frac{1}{K}\sum_{k=1}^{K} \widetilde{P}(L,R=r^k|C=c) \end{aligned} \quad (8)$$

where $\widetilde{P}$ denotes the inverse weighted probability. We employ a multi-head strategy, inspired by Vaswani et al. (2017) to refine the granularity of our sampling by partitioning the weight and feature dimensions into K groups, r_k denotes the review information in the group $K = k$. For simplicity, subsequent discussions will omit $C = c$, though it is understood that r remains dependent on c. We will employ *TDE* to debias this effect following Liu et al. (2022a) and Tang et al. (2020).

The energy-based model (LeCun et al., 2006) framework underpins our modeling of $\widetilde{P}$, where the softmax-activated probability is proportional to an energy function defined as:

$$\begin{aligned}\widetilde{P}(L = l, R = r^k) &\propto E(l, r^k; w^k) \\ &= \tau \frac{f(l, r^k; w^k)}{g(l, r^k; w^k)}\end{aligned} \tag{9}$$

with τ serving as a scaling factor analogous to the inverse temperature in Gibbs distributions (Geman and Geman, 1984), w^k denotes the weight parameter in the group $K = k$. The numerator $f(l, r^k; w^k)$ represents the unnormalized effect, calculated as logits $(w^k)^\top r^k$, while the denominator $g(l, r^k; w^k)$ serves as a normalization term(or propensity score (Austin, 2011)), which ensures balanced magnitudes of the variables. The denominator, *i.e.*, inverse probability weight, becomes the propensity score under the energy-based model, where the effect is divided into the controlled group $\|w^k\| \cdot \|r^k\|$ and the uncontrolled group $\epsilon \cdot \|r^k\|$.

The computation of logits for $P(L|do(R = r))$ is thus expressed as:

$$P(L|do(R)) = \frac{\tau}{K} \sum_{k=1}^{K} \frac{(w^k)^\top r^k}{(\|w^k\| + \epsilon)\|r^k\|} \tag{10}$$

Now we need to obtain context features given the current samples to force the model to concentrate on the debiased review based on *TDE*. We assume U as a confounder set $\{u_i\}_{i=1}^{N}$, where N is the number of aspects in dataset and u_i is the prototype for the context of class i in feature space. Review features can be linearly or non-linearly represented by the manifolds (Turk and Pentland, 1991; Candès et al., 2011), and so are the context features. Therefore, we model the review-specific context features C of current samples as follows:

$$C = f(r, U) = \sum_{N=1}^{N} P(u_n|r)u_n \tag{11}$$

where $P(u_i|r)$ is the classification probability of the feature r belonging to the context of class i.

The last remaining difficulty is implementing the contextual confounder set U. To obtain more useful contextual information, we employ the lower $\mathcal{K}$ layers of the model on $R \rightarrow L$ branch, which is in early training to model U. It is motivated by three primary considerations: **First**, the acknowledgment of the intrinsic wealth of contextual semantic information harbored within pre-trained language models (Liu et al., 2019b; Devlin et al., 2019) due to their extensive pre-training. **Second**, our requirement is not for highly advanced semantics but rather for contextual information (Zeiler and Fergus, 2014; Liu et al., 2022b); previous empirical studies (Jawahar et al., 2019; Liu et al., 2019a; Geva et al., 2021) have shown that encoder-only models exhibit superior performance in capturing contextual information at lower layers. **Third**, in the initial stages of training, the model's classification capabilities predominantly rely on context.

To be specific, we encode each r using the aforementioned method, and if r contains a specific aspect, it is then represented as the representation of the corresponding u_n, and we apply the mean feature as the final u_n representation.

Given the modeling of U and C, we are ready for the representations of context bias. We model them as $r_c = \mathcal{F}(r, C)$. Following Liu et al. (2022b), we choose $W \cdot concat(r, M)$ to map since adding more networks to learn how much we need from the context is better.

Now we can debias the impact of C on R ($C \rightarrow R$) based on *TDE*. The final definition of debiased $r^{'}$ is as follows:

$$\zeta_{r'} = \frac{\tau}{K} \sum_{k=1}^{K} \frac{(w^k)^\top}{(\|w^k\| + \epsilon)} \left(\frac{r^k}{\|r^k\|} - \frac{r_c^k}{\|r_c^k\|} \right) \tag{12}$$

3.3 Decorrelating the Aspect Branch with Counterfactual Reasoning

While we have successfully mitigated contextual bias in the $R \rightarrow L$ pathway, the ABSA model, as delineated in Figure 2, remains susceptible to aspect-only bias. This bias persists because the prediction, denoted as $L_{a,r',k}$, is directly influenced by the aspect variable A. To address this, we introduce a counterfactual reasoning approach that estimates the direct causal effect of A on L, effectively isolating the influence of R and K. Figure 3 shows the causal graph of the counterfactual world for ABSA which describes the scenario when A

is set to different values a and a^*. We also set R to its reference value r^*, therefore K would attain the value k^* when $R = r^*$ and $A = a^*$. In this way,the inputs of R and K are blocked, and the model can only rely on the given aspect a for detection. The natural direct effect (NDE) of A on L, which represents the aspect-only bias, is calculated as follows:

$$NDE_a = L_{a,r^*,k^*} - L_{a^*,r^*,k^*} \tag{13}$$

To eliminate this bias, we adjust TE by subtracting NDE, yielding TIE in Eq. (5).

Following the previous studies, we calculate the prediction $L_{a,r,k}$ through a model ensemble with a fusion function:

$$\begin{aligned} L_{a,r,k} &= L(A = a, R = r^{'}, K = k) \\ &= \Psi(\zeta_a, \zeta_{r'}, \zeta_k) \\ &= \zeta_k + \tanh(\zeta_a) + \tanh(\zeta_{r'}) \end{aligned} \tag{14}$$

where $\zeta_{r'}$ is the output of the review-only branch (*i.e.*, $R \to L$), ζ_a is the output of the aspect-only branch (*i.e.*, $A \to L$), and ζ_k is the output of fused features branch (*i.e.*, $K \to L$) as shown in Figure 3. TIE is the debiased result we used for inference.

3.4 Training and Inference

We compute separate losses for each branch during the training stage in line with the methodologies adopted by recent studies (Wang et al., 2021; Niu et al., 2021; Tian et al., 2022; Chen et al., 2023). These branches comprise the fused feature branch (base ABSA, $\mathcal{L}_K$), the aspect-only branch ($\mathcal{L}_A$), and the debiased review-only branch ($\mathcal{L}_R$). The collective minimization of these losses forms a comprehensive multi-task training objective, which serves to optimize the model parameters. The training objective is formally expressed as:

$$\mathcal{L} = \mathcal{L}_K + \alpha\mathcal{L}_A + \beta\mathcal{L}_R \tag{15}$$

where α and β are hyperparameters that control the contribution of each branch to the overall training objective.

The loss component $\mathcal{L}_K$ corresponds to the cross-entropy loss calculated from the predictions of $\Psi(\zeta_a, \zeta_{r'}, \zeta_k)$, as defined in Eq. (14). Similarly, the aspect-only and debiased review-only losses are denoted as $\mathcal{L}_A$ and $\mathcal{L}_R$ respectively.

We use debiased $TIE_{a,r}$ in Eq. 5 for inference.

4 Experiments

4.1 Datasets

We conduct training on the original SemEval 2014 Laptop and Restaurant datasets (Pontiki et al., 2014), and perform testing on the ARTS datasets, as introduced by Xing et al. (2020), to assess the efficacy of the proposed method. Detailed information about the ARTS datasets is shown in Appendix A.

4.2 Baselines

We consider baselines in the ARTS original paper (Xing et al., 2020), which are listed in Appendix B and following strong baselines for comparison:

GraphMerge: Hou et al. (2021) combine multiple dependency trees using a graph-ensemble technique for aspect-level sentiment analysis.

SENTA: Bi et al. (2021) propose a novel Sentiment Adjustment model, employing backdoor adjustment to mitigate confounding effects. And **PT-SENTA** use BERT-PT (Xu et al., 2019) as backbone.

NADS: Cao et al. (2022) apply no-aspect contrastive learning to reduce aspect sentiment bias and improve sentence representations.

ChatGPT: ChatGPT is a conversational version of GPT-3.5 model (Ouyang et al., 2022; OpenAI, 2022). We use the `gpt-3.5-turbo-0125` API from OpenAI[2]. The prompts for this task are presented in Appendix C.

4.3 Implementations

Our method is model-agnostic. In the empirical study, we utilize two types of mainstream encoder-only model, RoBERTa (Liu et al., 2019b)[3] and BERT (Devlin et al., 2019)[4] as the backbone for our experiments. For comprehensive details on the hyperparameters employed in our experiments, refer to Appendix D.

4.4 Evaluation

Following the previous works (Wang et al., 2016; Xue and Li, 2018; Cao et al., 2022), Accuracy (*Acc.*), F1-score and Aspect Robustness Score (ARS) (Xing et al., 2020) are employed as complementary evaluation metrics. ARS considers the

[2] https://platform.openai.com/docs/models/gpt-3-5

[3] https://huggingface.co/FacebookAI/roberta-base

[4] https://huggingface.co/bert-base-uncased

	Laptop			Restaurant		
Model	*Acc.*	F1-score	ARS	*Acc.*	F1-score	ARS
MemNet (Tang et al., 2016b)	-	-	16.93	-	-	21.52
GatedCNN (Xue and Li, 2018)	-	-	10.34	-	-	13.12
AttLSTM (Wang et al., 2016)	-	-	9.87	-	-	14.64
TD-LSTM (Tang et al., 2016a)	-	-	22.57	-	-	30.18
GCN (Zhang et al., 2019)	-	-	19.91	-	-	24.73
BERT-Sent (Xing et al., 2020)	-	-	14.70	-	-	10.89
CapsBERT (Jiang et al., 2019)	-	-	25.86	-	-	55.36
BERT-PT (Xu et al., 2019)	-	-	53.29	-	-	59.29
GraphMerge (Hou et al., 2021)	-	-	52.90	-	-	57.46
NADS (Cao et al., 2022)	-	-	58.77	-	-	64.55
SENTA (Bi et al., 2021)	67.23	-	-	77.30	-	-
PT-SENTA (Bi et al., 2021)	74.16	-	-	80.91	-	-
ChatGPT (Wang et al., 2023)	68.89	56.22	46.39	79.21	61.33	45.01
BERT (Xing et al., 2020)	-	-	50.94	-	-	54.82
BERT†	70.43	66.55	49.53	78.56	69.35	57.86
DINER(BERT-based)	72.56	68.40	53.76	80.69	72.79	62.23
RoBERTa (Ma et al., 2021)	73.57	69.26	-	79.08	72.79	-
RoBERTa†	74.96	72.16	56.27	79.26	70.47	59.96
DINER(RoBERTa-based)	**76.51**	**73.27**	**59.40**	**82.46**	**76.92**	**64.02**

Table 1: We retrained BERT†, RoBERTa† as fair baselines ensuring that comparisons are made under similar training settings, which is crucial for validating **DINER**'s superior performance.

accurate classification of a source example and all its derived variants, produced through the aforementioned three strategies, as a single instance of correctness.

5 Result and Analysis

Table 1 presents a detailed comparison of various models' performance for laptop and restaurant domains of the ARTS datasets, focusing on three key evaluation metrics: *Acc.*, F1-score, and ARS.

Overall, PLMs, on average, perform better than non-PLMs due to the pre-trained knowledge and tasks, making them more robust. Surprisely, ChatGPT does not get perform well in this task, exhibiting ARS scores of only 50.94 in the laptop domain and 54.82 in the restaurant domain, which are even lower than those of most PLMs in Table 1. This underscores ChatGPT's relatively poor robustness in ARTS's variations, despite its otherwise robust performance across various other NLP tasks.

We evaluate **DINER** in two backbones: BERT-based and RoBERTa-based. These configurations are set to evaluate the effectiveness of **DINER** when integrated with different encoder-only PLMs. And **DINER** based on RoBERTa tends to outperform its BERT counterparts, which may be attributed to RoBERTa's more robust pre-training on a larger and more diverse corpus, leading to better generalization capabilities (Liu et al., 2019b). The results are compelling, showing that **DINER**(RoBERTa-based) model achieves the state-of-the-art performance across all metrics in both the laptop and restaurant domains, with a notable *Acc.* of 76.51 and 82.46, F1-scores of 73.27 and 76.92, and ARS of 59.40 and 64.02, respectively. **DINER**(RoBERTa-based) demonstrates superior performance in the Laptop domain, outpacing the baseline RoBERTa† by margins of **1.55, 1.11, and 3.13** in terms of *Acc.*, F1-score, and ARS metrics, respectively. In the Restaurant domain, the model further extends its lead, achieving improvements of **3.20, 6.45, and 4.06** in the same metrics. Similarly, the **DINER**(BERT-based) exhibits empirical enhancements.

5.1 More Detailed Result

We list in detail the performance of each model on the aforementioned three subsets of the ARTS datasets in Table 2. In the Laptop domain, the baseline model RoBERTa† exhibits a RevTgt accuracy of 62.45, as RevTgt is the most challenging subset. It requires the model to pay precise attention to the target sentiment words. In contrast, the DINER framework significantly enhances this metric to 65.02, marking a 4.12% increment. Similarly, for RevNon and AddDiff, DINER outperforms the Vanilla baseline with modest improvements of 0.86% and 2.27%, respectively. The Restaurant do-

		RevTgt	RevNon	AddDiff	Original
Laptop	`Vanilla`	62.45	85.93	76.33	80.41
	DINER	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
	DINER	70.69(↑ 10.35%)	83.56(↑ 1.08%)	86.07(↑ 3.10%)	87.32(↑ 2.51%)

Table 2: We use RoBERTa as the backbone. `Vanilla` refers to RoBERTa† in Table 1. We compare the *Acc.* on the `Vanilla` and our **DINER** framework. We also calculate the *change* of accuracy.

Methods	Laptop	Restaurant
`Vanilla`	74.96	79.26
$R \rightarrow L$ branch		
+ Causal Intervention(NWGM)	75.44	81.02
+ Causal Intervention(IPW)	75.50	81.19
+*TDE*	75.92	81.78
$A \rightarrow L$ branch		
+ Counterfactual Inference	75.23	80.51
DINER	**76.51**	**82.46**

Table 3: Ablation studies on two branches of our method. Experiments are based on RoBERTa backbone, *Acc.* are reported.

Fusion Strategy	Laptop	Restaurant
`MUL-Vanilla`	53.01	65.52
`MUL-sigmoid`	63.72	76.35
`MUL-tanh`	52.10	61.36
`SUM-Vanilla`	74.32	80.14
`SUM-sigmoid`	75.97	81.76
`SUM-tanh`	**76.51**	**82.46**

Table 4: Impact of Different Fusion Strategies.

main further underscores the efficacy of the DINER framework, where a remarkable 10.35% improvement is observed in the RevTgt task, elevating the accuracy from 64.06 to 70.69. The framework also exhibits gains in RevNon and AddDiff tasks by 1.08% and 3.10%, respectively. The significant improvement observed in the restaurant domain underscores the effectiveness of our methods, particularly given the inherently challenging nature of the test set data in this domain, as highlighted by Xing et al. (2020). Specifically, the more challenging the dataset, the greater the improvement our framework offers. Interestingly, our method also gives a slight improvement on the Original test set, illustrating the fact that we have also debiased the robust test data on the Original test set.

5.2 Effects of the Two Branches in DINER

We delve into the empirical evaluation of the dual-branch architecture underpinning the DINER framework, specifically examining its constituent elements through ablation studies. The studies are shown in Table 3, offering insights into the incremental benefits conferred by each branch.

For the $R \rightarrow L$ branch, NWGM (Xu et al., 2015) yields a marginal improvement in accuracy across both domains. The method of IPW (Pearl, 2009) further enhances performance, suggesting the efficacy of backdoor adjustment intervention in the methods, and IPW has a more precise approximation compared to NWGM (Xu et al., 2015). We further debias context based on TDE, as described in Section 3.2, and performance is further enhanced upon the application of Counterfactual Reasoning.

Parallel to this, the $A \rightarrow L$ branch investigates the impact of Counterfactual Inference. After conducting Counterfactual Inference at this branch, *Acc.* in the Laptop and Restaurant domains improved by 0.27 and 1.25, respectively. In the Restaurant domain, the bias associated with aspects is more pronounced.

By comparing the performance improvements at both branches, we can also discern that the bias and shortcuts from $R \rightarrow L$ branch are more pronounced, and our approach has effectively addressed these issues.

5.3 Impact of Different Fusion Strategies

Following prior studies (Wang et al., 2021; Niu et al., 2021; Chen et al., 2023), we devise several differentiable arithmetic binary operations for the fusion strategy in Eq. (16):

$$\begin{cases} \texttt{MUL-Vanilla}: L_{a,r',k} = \zeta_a \cdot \zeta_{r'} \cdot \zeta_k, \\ \texttt{MUL-sigmoid}: L_{a,r',k} = \zeta_k \cdot \sigma(\zeta_a) \cdot \sigma(\zeta_{r'}), \\ \texttt{MUL-tanh}: L_{a,r',k} = \zeta_k \cdot \tanh(\zeta_a) \cdot \tanh(\zeta_{r'}), \\ \texttt{SUM-Vanilla}: L_{a,r',k} = \zeta_a + \zeta_{r'} + \zeta_k, \\ \texttt{SUM-sigmoid}: L_{a,r',k} = \zeta_k + \sigma(\zeta_a) + \sigma(\zeta_{r'}), \\ \texttt{SUM-tanh}: L_{a,r',k} = \zeta_k + \tanh(\zeta_a) + \tanh(\zeta_{r'}) \end{cases} \tag{16}$$

The *Acc.* performance of six distinct different fusion strategies are reported in Table 4. From the

Type	Examples(Target Aspect: food)	Gold	Baseline	DINER
Original	The **food** is top notch, the service is attentive, and the atmosphere is great.	Positive	Positive ✓	Positive ✓
RevTgt	The **food** is nasty, but the service is attentive, and the atmosphere is great.	Negative	Negative ✓	Negative ✓
RevNon	The **food** is top notch, the service is heedless, but the atmosphere is not great.	Positive	Negative ✗	Positive ✓
AddDiff	The **food** is top notch, the service is attentive, and the atmosphere is great, but music is too heavy, waiters is angry and staff is arrogant.	Positive	Negative ✗	Positive ✓

Table 5: Examples of case study. The corresponding gold labels and the predictions for each example are presented.

table, we can find that the MUL fusion, regardless of the activation function, consistently underperforms in comparison to its SUM counterparts. Apparently, SUM fusion strategies are more stable and robust, and more suitable for the ASBA task. The superior performance of SUM fusion strategies, particularly with the tanh activation, underscores the effectiveness of the additive strategy in capturing the nuanced interplay of features pertinent to the ABSA task.

5.4 Case Study

To demonstrate the efficacy of the proposed method, we present a case study featuring a sample and its three adversarial variants in Table 5. We compare our proposed method based on RoBERTa with the baseline RoBERTa†.

From the table, the results clearly demonstrate that our method **DINER**, exhibits enhanced robustness compared to the baseline approach. Specifically, Original and RevTgt types, where either no changes or direct negative changes were made to the targeted aspect, both methods perform equally well.

However, the distinction in performance is evident in more complex adversarial examples. In the RevNon type, where distractors are introduced in non-target aspects (*e.g.*, service and atmosphere), the baseline fails to maintain its accuracy, misclassifying the overall sentiment as Negative. In contrast, DINER successfully recognizes the sentiment as Positive, reflecting its ability to isolate the influence of perturbations to non-target aspects. The AddDiff type further complicates the scenario by adding multiple negative aspects unrelated to the target. Despite these challenges, DINER continues to accurately assess the sentiment towards the food as Positive, whereas the baseline erroneously shifts to a Negative prediction.

The resilience of our method to adversarial conditions suggests it is well-suited for real-world environments where reliable sentiment analysis is crucial.

6 Conclusion

In this paper, to debias the target and review in ABSA simultaneously, a novel debiasing framework, DINER, is proposed with multi-variable causal inference. Specifically, the aspect is assumed to have a direct correlation with the label, so a counterfactual reasoning-based intervention is employed to debias the aspect branch. In the meantime, the sentiment words towards the target in the review are assumed to be indirectly confounded with the context, where a backdoor adjustment-based intervention is employed to debias the review branch. Extensive experiments show the effectiveness of the proposed method in debiasing ABSA compared to normal state-of-the-art ABSA methods and debiasing methods.

Limitations

Though achieving promising results in the experiments, our work still has the following limitations.

- Though the proposed method is based on multi-variable causal inference, the causal effects of the target aspect and the review are assumed to be independent, which means no interaction between the target and the review is modeled or considered.
- The proposed method is only evaluated on two robustness testing datasets for ABSA. More real-world datasets and more data transformation methods should be evaluated for future work.
- The general ABSA task includes the joint extraction of aspect and sentiment polarity, while the proposed method restricts the task to a given aspect. Future work should be considered for more generalized ABSA tasks.

Acknowledgement

The authors would like to thank the anonymous reviewers for their insightful comments. This work is funded by the National Natural Science Foundation of China (62176053). This work is supported by the Big Data Computing Center of Southeast University.

References

Peter C Austin. 2011. An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivariate behavioral research, 46(3):399–424.

Xuefeng Bai, Pengbo Liu, and Yue Zhang. 2020. Investigating typed syntactic dependencies for targeted sentiment classification using graph attention neural network. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29:503–514.

Zhen Bi, Ningyu Zhang, Ganqiang Ye, Haiyang Yu, Xi Chen, and Huajun Chen. 2021. Interventional aspect-based sentiment analysis. arXiv preprint arXiv:2104.11681.

Emmanuel J Candès, Xiaodong Li, Yi Ma, and John Wright. 2011. Robust principal component analysis? Journal of the ACM (JACM), 58(3):1–37.

Jiahao Cao, Rui Liu, Huailiang Peng, Lei Jiang, and Xu Bai. 2022. Aspect is not you need: No-aspect differential sentiment framework 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, pages 1599–1609.

Ziwei Chen, Linmei Hu, Weixin Li, Yingxia Shao, and Liqiang Nie. 2023. 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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Stuart Geman and Donald Geman. 1984. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. IEEE Transactions on pattern analysis and machine intelligence, (6):721–741.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key-value memories. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 5484–5495, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Wangzhen Guo, Qinkang Gong, Yanghui Rao, and Hanjiang Lai. 2023. Counterfactual multihop QA: A cause-effect approach for reducing disconnected reasoning. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 4214–4226, Toronto, Canada. Association for Computational Linguistics.

Xiaochen Hou, Peng Qi, Guangtao Wang, Rex Ying, Jing Huang, Xiaodong He, and Bowen Zhou. 2021. Graph ensemble learning over multiple dependency trees for aspect-level sentiment classification. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 2884–2894.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, pages 168–177.

Binxuan Huang and Kathleen Carley. 2018. Parameterized convolutional neural networks for aspect level sentiment classification. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 1091–1096, Brussels, Belgium. Association for Computational Linguistics.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.

Long Jiang, Mo Yu, Ming Zhou, Xiaohua Liu, and Tiejun Zhao. 2011. Target-dependent Twitter sentiment classification. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pages 151–160, Portland, Oregon, USA. Association for Computational Linguistics.

Qingnan Jiang, Lei Chen, Ruifeng Xu, Xiang Ao, and Min Yang. 2019. A challenge dataset and effective models for aspect-based sentiment analysis. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 6280–6285, Hong Kong, China. Association for Computational Linguistics.

Rabeeh Karimi Mahabadi, Yonatan Belinkov, and James Henderson. 2020. End-to-end bias mitigation by modelling biases in corpora. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 8706–8716, Online. Association for Computational Linguistics.

Svetlana Kiritchenko, Xiaodan Zhu, Colin Cherry, and Saif Mohammad. 2014. NRC-Canada-2014: Detecting aspects and sentiment in customer reviews. In Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pages 437–442, Dublin, Ireland. Association for Computational Linguistics.

Yann LeCun, Sumit Chopra, Raia Hadsell, M Ranzato, and Fujie Huang. 2006. A tutorial on energy-based learning. Predicting structured data, 1(0).

Minwoo Lee, Seungpil Won, Juae Kim, Hwanhee Lee, Cheoneum Park, and Kyomin Jung. 2021. Crossaug: A contrastive data augmentation method for debiasing fact verification models. In Proceedings of the 30th ACM International Conference on Information & Knowledge Management, CIKM '21, page 3181–3185, New York, NY, USA. Association for Computing Machinery.

Bing Liu, Dong Wang, Xu Yang, Yong Zhou, Rui Yao, Zhiwen Shao, and Jiaqi Zhao. 2022a. Show, deconfound and tell: Image captioning with causal inference. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 18041–18050.

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019a. Linguistic knowledge and transferability of contextual representations. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 1073–1094, Minneapolis, Minnesota. Association for Computational Linguistics.

Ruyang Liu, Hao Liu, Ge Li, Haodi Hou, Ting-Hao Yu, and Tao Yang. 2022b. Contextual debiasing for visual recognition with causal mechanisms. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 12755–12765.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

Ilya Loshchilov and Frank Hutter. 2018. Decoupled weight decay regularization. In International Conference on Learning Representations.

Fang Ma, Chen Zhang, and Dawei Song. 2021. Exploiting position bias for robust aspect sentiment classification. In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 1352–1358, Online. Association for Computational Linguistics.

Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. 2021. Counterfactual vqa: A cause-effect look at language bias. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 12700–12710.

OpenAI. 2022. Introducing ChatGPT.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems, 35:27730–27744.

Judea Pearl. 1995. Causal diagrams for empirical research. Biometrika, 82(4):669–688.

Judea Pearl. 2009. Causal inference in statistics: An overview.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Haris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. SemEval-2014 task 4: Aspect based sentiment analysis. In Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pages 27–35, Dublin, Ireland. Association for Computational Linguistics.

Tal Schuster, Darsh Shah, Yun Jie Serene Yeo, Daniel Roberto Filizzola Ortiz, Enrico Santus, and Regina Barzilay. 2019. Towards debiasing fact verification models. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 3419–3425, Hong Kong, China. Association for Computational Linguistics.

Duyu Tang, Bing Qin, Xiaocheng Feng, and Ting Liu. 2016a. Effective LSTMs for target-dependent sentiment classification. In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 3298–3307, Osaka, Japan. The COLING 2016 Organizing Committee.

Duyu Tang, Bing Qin, and Ting Liu. 2016b. Aspect level sentiment classification with deep memory network. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 214–224, Austin, Texas. Association for Computational Linguistics.

Kaihua Tang, Jianqiang Huang, and Hanwang Zhang. 2020. Long-tailed classification by keeping the good and removing the bad momentum causal effect. Advances in neural information processing systems, 33:1513–1524.

Bing Tian, Yixin Cao, Yong Zhang, and Chunxiao Xing. 2022. Debiasing nlu models via causal intervention and counterfactual reasoning. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 36, pages 11376–11384.

Matthew A Turk and Alex P Pentland. 1991. Face recognition using eigenfaces. In Proceedings. 1991 IEEE computer society conference on computer vision and pattern recognition, pages 586–587. IEEE Computer Society.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. Advances in neural information processing systems, 30.

Duy-Tin Vo and Yue Zhang. 2015. Target-dependent twitter sentiment classification with rich automatic features. In Twenty-fourth international joint conference on artificial intelligence.

Shuai Wang, Sahisnu Mazumder, Bing Liu, Mianwei Zhou, and Yi Chang. 2018. Target-sensitive memory networks for aspect sentiment classification. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 957–967, Melbourne, Australia. Association for Computational Linguistics.

Siyin Wang, Jie Zhou, Changzhi Sun, Junjie Ye, Tao Gui, Qi Zhang, and Xuanjing Huang. 2022. Causal intervention improves implicit sentiment analysis. In Proceedings of the 29th International Conference on Computational Linguistics, pages 6966–6977, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Wenjie Wang, Fuli Feng, Xiangnan He, Hanwang Zhang, and Tat-Seng Chua. 2021. Clicks can be cheating: Counterfactual recommendation for mitigating clickbait issue. In Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 1288–1297.

Yequan Wang, Minlie Huang, Xiaoyan Zhu, and Li Zhao. 2016. Attention-based LSTM for aspect-level sentiment classification. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 606–615, Austin, Texas. Association for Computational Linguistics.

Zengzhi Wang, Qiming Xie, Zixiang Ding, Yi Feng, and Rui Xia. 2023. Is chatgpt a good sentiment analyzer? a preliminary study. arXiv preprint arXiv:2304.04339.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 6382–6388, Hong Kong, China. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 38–45, Online. Association for Computational Linguistics.

Xiaoyu Xing, Zhijing Jin, Di Jin, Bingning Wang, Qi Zhang, and Xuanjing Huang. 2020. Tasty burgers, soggy fries: Probing aspect robustness in aspect-based sentiment analysis. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 3594–3605, Online. Association for Computational Linguistics.

Hu Xu, Bing Liu, Lei Shu, and Philip Yu. 2019. BERT post-training for review reading comprehension and aspect-based sentiment analysis. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 2324–2335, Minneapolis, Minnesota. Association for Computational Linguistics.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In International conference on machine learning, pages 2048–2057. PMLR.

Weizhi Xu, Qiang Liu, Shu Wu, and Liang Wang. 2023. Counterfactual debiasing for fact verification. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 6777–6789, Toronto, Canada. Association for Computational Linguistics.

Wei Xue and Tao Li. 2018. Aspect based sentiment analysis with gated convolutional networks. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2514–2523, Melbourne, Australia. Association for Computational Linguistics.

Matthew D Zeiler and Rob Fergus. 2014. Visualizing and understanding convolutional networks. In Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13, pages 818–833. Springer.

Chen Zhang, Qiuchi Li, and Dawei Song. 2019. Aspect-based sentiment classification with aspect-specific graph convolutional networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 4568–4578, Hong Kong, China. Association for Computational Linguistics.

Congzhi Zhang, Linhai Zhang, Jialong Wu, Deyu Zhou, and Yulan He. 2024a. Causal prompting: Debiasing large language model prompting based on front-door adjustment.

Congzhi Zhang, Linhai Zhang, and Deyu Zhou. 2024b. Causal walk: Debiasing multi-hop fact verification with front-door adjustment. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 38, pages 19533–19541.

Meishan Zhang, Yue Zhang, and Duy-Tin Vo. 2016. Gated neural networks for targeted sentiment analysis. In Proceedings of the AAAI conference on artificial intelligence, volume 30.

Wenkai Zhang, Hongyu Lin, Xianpei Han, and Le Sun. 2021a. De-biasing distantly supervised named entity recognition via causal intervention. 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), pages 4803–4813, Online. Association for Computational Linguistics.

Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2022. A survey on aspect-based sentiment analysis: Tasks, methods, and challenges. IEEE Transactions on Knowledge and Data Engineering.

Yang Zhang, Fuli Feng, Xiangnan He, Tianxin Wei, Chonggang Song, Guohui Ling, and Yongdong Zhang. 2021b. Causal intervention for leveraging popularity bias in recommendation. In Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 11–20.

Junhao Zheng, Zhanxian Liang, Haibin Chen, and Qianli Ma. 2022. Distilling causal effect from miscellaneous other-class for continual named entity recognition. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 3602–3615, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jiazheng Zhu, Shaojuan Wu, Xiaowang Zhang, Yuexian Hou, and Zhiyong Feng. 2023. 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.

	REVTGT	REVNON	ADDDIFF	ALL
Laptop	466	135	638	1877
Restaurant	846	444	1120	3530

Table 6: The statistics of datasets being evaluated.

A Dataset Example

ARTS datasets employ three distinct strategies to rigorously test the model's robustness: REVTGT is to generate sentences that reverse the original sentiment of the target aspect. REVNON is to change the target sentiment. ADDDIFF investigate if adding more nontarget aspects can confuse the model. We provide concrete instances of how each strategy is applied to manipulate aspect sentiment within the dataset in Table 7. Detailed statistics of the test sets are provided in Table 6.

B Baselines

TD-LSTM: Tang et al. (2016a) use dual LSTMs to encode context around a target aspect, combining final states for sentiment classification.

AttLSTM: Wang et al. (2016) introduce an Attention-based LSTM that merges aspect and word embeddings for each token.

GatedCNN: Xue and Li (2018) utilize a Gated CNN with a Tanh-ReLU mechanism, integrating aspect embeddings with CNN-encoded text.

MemNet: Tang et al. (2016b) employ memory networks, using sentences as external memory to compute attention based on the target aspect.

GCN: Zhang et al. (2019) apply a GCN to the sentence's syntax tree, followed by an aspect-specific masking layer.

BERT: Xu et al. (2019) use a BERT-based (Devlin et al., 2019) baseline and takes as input the concatenation of the aspect and the review.

BERT-Sent: (Xu et al., 2019) BERT-Sent takes as input reviews without aspect.

BERT-PT: Xu et al. (2019) enhance BERT's capabilities through post-training on additional review datasets.

CapsBERT: Jiang et al. (2019) use BERT to encode sentences and aspect terms, then utilize Capsule Networks for polarity prediction.

C ChatGPT Prompt

We conduct 3-shot prompting experiments on the ARTS datasets following (Wang et al., 2023). We set the decoding temperature as 0 to increase ChatGPT's determinism. The prompts are presented in Table 8.

D Model Hyper Parameters

The model parameters are optimized by AdamW (Loshchilov and Hutter, 2018), with a learning rate of 5e-5 and weight decay of 0.01. The batch size is 256, and a dropout probability of 0.1 is used. The number of training epochs is 20. We explore the hyperparameters α and β, setting their values to $\{0.6, 0.8, 1, 1.2, 1.4\}$ for each, respectively. The optimal values for α and β are 0.8 and 1.0, respectively. We set $\mathcal{K}$ in set $\{3,6,9\}$ in accordance with the theoretical principles discussed in (Geva et al., 2021). Our implementation leverages the *PyTorch*[5] framework and *HuggingFace Transformers*[6] library (Wolf et al., 2020). Our experiments are carried out with an NVIDIA A100 80GB GPU.

[5] https://github.com/pytorch/pytorch

[6] https://github.com/huggingface/transformers

Type	Review
Original	Tasty **burgers**, and crispy fries. (Target Aspect: **burgers**)
RevTgt	Terrible **burgers**, but crispy fries.
RevNon	Tasty **burgers**, but soggy fries.
AddDiff	Tasty **burgers**, crispy fries, but poorest service ever!

Table 7: The adversarial examples of the original sentence. Each example is annotated with the **Target Aspect**, and altered sentence parts.

Dataset	**Prompt**
Laptop	Sentence: The **screen** almost looked like a barcode when it froze. What is the sentiment polarity of the aspect **screen** in this sentence? Label: negative Sentence: Screen, keyboard, and mouse: If you cant see yourself spending the extra money to jump up to a Mac the beautiful screen, responsive **island backlit keyboard**, and fun multi-touch mouse is worth the extra money to me alone. What is the sentiment polarity of the aspect **island backlit keyboard** in this sentence? Label: positive Sentence: Size: I know 13 is small (especially for a desktop replacement) but with an **external monitor**, who cares. What is the sentiment polarity of the aspect **external monitor** in this sentence? Label: neutral Sentence: {sentence} What is the sentiment polarity of the {aspect} in this sentence?
Restaurant	Sentence: Our **server** was very helpful and friendly. What is the sentiment polarity of the aspect **server** in this sentence? Label: positive Sentence: We had **reservations** at 9pm, but was not seated until 10:15pm. What is the sentiment polarity of the aspect **reservation** in this sentence? Label: negative Sentence: It's the perfect restaurant for NY life style, it got cool design, awsome drinks and food and lot's of good looking people eating and hanging at the pink **bar**... What is the sentiment polarity of the aspect **bar** in this sentence? Label: neutral Sentence: {sentence} What is the sentiment polarity of the {aspect} in this sentence?

Table 8: The prompts used for prompting ChatGPT for each domain.