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.

Introduction

Most ABSA methods solve the task as an **input-output mapping problem** based on high-capacity neural networks and pre-trained language models. Though remarkable progress has been made, it is demonstrated that these models are not robust in data transformation where simply reversing the polarity of the target results in over 20% drop 😢 in accuracy.

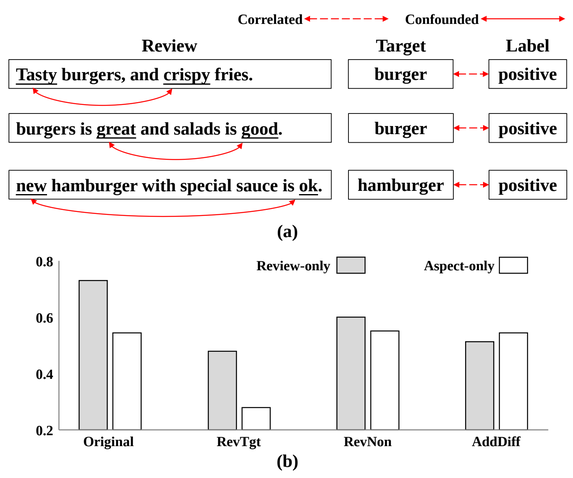

Figure 1. (a) Examples are taken from the SemEval 2014 Restaurant test set. (b) RevTgt denotes reversing the polarity of the target aspect, RevNon denotes reversing the polarity of the non-target aspect, and AddDiff denotes adding another non-target aspect with different polarity.

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

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.

Structural Causal Model of ABSA

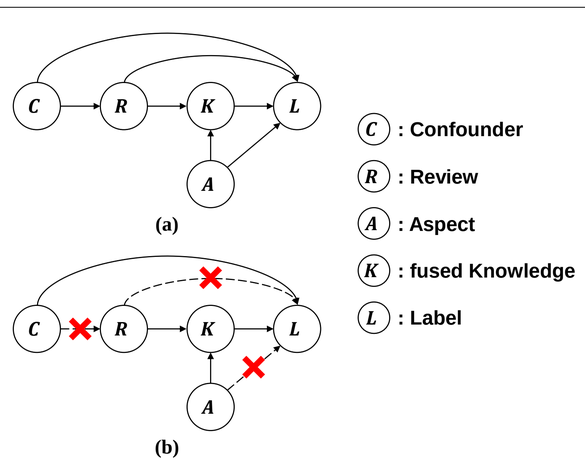

Figure 2. (a) SCM of ABSA. (b) The desired situation for ABSA, the dotted line means the causalities are blocked.

- $R \rightarrow K \leftarrow A$. 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.
- $K \rightarrow L$. The label L is caused by the fused knowledge K, which is the desired causal effect of ABSA.
- $R \rightarrow L \leftarrow A$. The label L is also directly affected by review R and aspect A, where the spurious correlation comes from and should be removed.
- $C \rightarrow R$ and $C \rightarrow L$. 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.

The framework of DINER

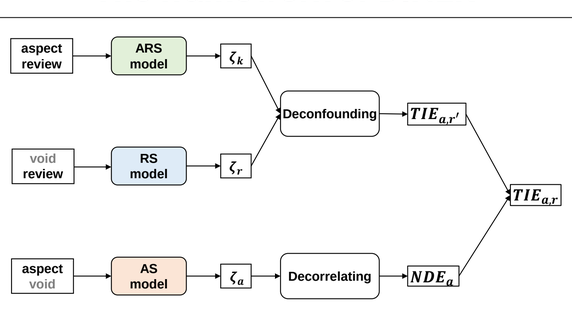

Figure 3. The framework of the proposed method.

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.

$$TIE_{a,r} = TE_{a,r} - NDE_r - NDE_a + IE_{a,r} \quad (1)$$
$$TE_{a,r} = L_{a,r,k} - L_{a^*,r^*,k^*} \quad (2)$$
$$NDE_a = L_{a,r^*,k^*} - L_{a^*,r^*,k^*} \quad (3)$$
$$NDE_r = L_{a^*,r,k^*} - L_{a^*,r^*,k^*} \quad (4)$$
$$TIE_{a,r} = L_{a,r,k} - L_{a^*,r,k^*} - L_{a,r^*,k^*} + L_{a^*,r^*,k^*} \quad (5)$$
$$= TIE_{a,r'} - NDE_a \quad (6)$$

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.

Deconfounding the Review Branch with Backdoor Adjustment

$$L_{a,r,k} = \Psi(\zeta_a, \zeta_{r'}, \zeta_k) \quad (7)$$

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.

Consider the SCM only contains R, C, and L, C satisfies the backdoor criterion, and we can have:

$$P(L|do(R)) = \sum_{\boldsymbol{c}} P(L|R,C)P(C) = \sum_{\boldsymbol{c}} \frac{P(L,R|C)P(C)}{P(R|C)} \quad (8)$$

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

$$P(L|do(R=r)) \approx \widetilde{P}(L,R|C=c) \approx \frac{1}{K}\sum_{k=1}^{K} \widetilde{P}(L,R=r^k|C=c) \quad (9)$$

where $\widetilde{P}$ denotes the inverse weighted probability.

$$\widetilde{P}(L=l,R=r^k) \propto E(l,r^k;w^k) = \tau\frac{f(l,r^k;w^k)}{g(l,r^k;w^k)} \quad (10)$$

with τ serving as a scaling factor analogous to the inverse temperature in Gibbs distributions, w^k denotes the weight parameter in the group $K=k$. The computation of logits for $P(L|do(R=r))$ is thus expressed as:

$$P(L|do(R)) = \frac{\tau}{K}\sum_{k=1}^{K}\frac{(w^k)^\top r^k}{(\|w^k\|+\epsilon)\|r^k\|} \quad (11)$$

Therefore, we model the review-specific context features C of current samples as follows:

$$C = f(r,U) = \sum_{N=1}^{N} P(u_n|r)u_n \quad (12)$$

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

$$\zeta_{r'} = \frac{\tau}{K}\sum_{k=1}^{K}\frac{(w^k)^\top}{(\|w^k\|+\epsilon)}\left(\frac{r^k}{\|r^k\|} - \frac{r_c^k}{\|r_c^k\|}\right) \quad (13)$$

Decorrelating the Aspect Branch with Counterfactual Reasoning

The NDE of A on L, which represents the aspect-only bias, is calculated as follows:

$$NDE_a = L_{a,r^*,k^*} - L_{a^*,r^*,k^*} \quad (14)$$

We calculate the prediction $L_{a,r,k}$ through a model ensemble with a fusion function:

$$L_{a,r,k} = L(A=a,R=r',K=k) = \Psi(\zeta_a,\zeta_{r'},\zeta_k) = \zeta_k + \tanh(\zeta_a) + \tanh(\zeta_{r'}) \quad (15)$$

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

Experimental Result

	Laptop			Restaurant		
Model	*Acc.*	F1-score	ARS	*Acc.*	F1-score	ARS
MemNet	-	-	16.93	-	-	21.52
GatedCNN	-	-	10.34	-	-	13.12
AttLSTM	-	-	9.87	-	-	14.64
TD-LSTM	-	-	22.57	-	-	30.18
GCN	-	-	19.91	-	-	24.73
BERT-Sent	-	-	14.70	-	-	10.89
CapsBERT	-	-	25.86	-	-	55.36
BERT-PT	-	-	53.29	-	-	59.29
GraphMerge	-	-	52.90	-	-	57.46
NADS	-	-	58.77	-	-	64.55
SENTA	67.23	-	-	77.30	-	-
PT-SENTA	74.16	-	-	80.91	-	-
ChatGPT	68.89	56.22	46.39	79.21	61.33	45.01
BERT	-	-	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	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.

Case Study

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 ✓