Causal Intervention Improves Implicit Sentiment Analysis

Siyin Wang[†], Jie Zhou[†*], Changzhi Sun[§], Junjie Ye[†], Tao Gui[†], Qi Zhang[†], and Xuanjing Huang[†]

[†]School of Computer Science, Fudan Univerisity
[§]ByteDance AI Lab
{siyinwang20, jie_zhou, jjye19}@fudan.edu.cn
{tgui, qz, xjhuang}@fudan.edu.cn
sunchangzhi@bytedance.com

Abstract

Despite having achieved great success for sentiment analysis, existing neural models struggle with implicit sentiment analysis. This may be due to the fact that they may latch onto spurious correlations ("shortcuts", e.g., focusing only on explicit sentiment words), resulting in undermining the effectiveness and robustness of the learned model. In this work, we propose a CausaL intervention model for implicit sEntiment ANalysis using instrumental variable (CLEAN). We first review sentiment analysis from a causal perspective and analyze the confounders existing in this task. Then, we introduce an instrumental variable to eliminate the confounding causal effects, thus extracting the pure causal effect between sentence and sentiment. We compare the proposed CLEAN model with several strong baselines on both the general implicit sentiment analysis and aspect-based implicit sentiment analysis tasks. The results indicate the great advantages of our model and the efficacy of implicit sentiment reasoning.

1 Introduction

The remarkable success that the field of sentiment analysis has achieved in the past few years has been derived from the use of increasingly high-capacity neural models to extract correlations from data (Peters et al., 2018; Devlin et al., 2018; Liu et al., 2019). Although having reached state-of-the-art results, correlational predictive models can be untrustworthy (Guidotti et al., 2018): they may latch onto spurious correlations ("shortcuts"), leading to poor generalization.

One shortcut might be the explicit sentiment word which is a powerful feature cue. Unfortunately, such a shortcut severely harms the generalization and the robustness of the learned models in implicit sentiment analysis (ISA), where there are no explicit sentiment words about the topic or aspect (Russo et al., 2015). Figure 1 gives a sample of aspect-based sentiment analysis (ABSA) (Zhou et al., 2020b,a), which aims to predict the sentiments of the aspects in the sentence. The aspect "food" has explicit sentiment words "definitely good", but aspect "price" does not. If the model thoughtlessly relies on shortcuts to sentiment words, it may make an incorrect sentiment prediction about the aspect "price". In fact, there are many other kinds of shortcuts that models may learn, for example, the rhetorical question mood expressed by the users (Ranganath et al., 2018) and the co-occurrence of neutral words and sentiment polarities (Wang and Culotta, 2020).

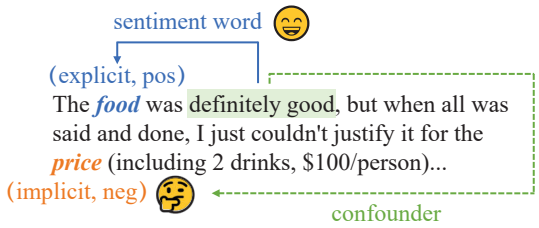


Figure 1: An examples of confounding factors in implicit sentiment analysis for ABSA.

The above shortcomings can potentially be addressed by the causal perspective: knowledge of causal relationships between observations and labels can be used to formalize spurious correlations and alleviate the predictor's dependence on them (Bühlmann, 2020; Veitch et al., 2021; Feder et al., 2021). Motivated by a causal perspective, we incorporate domain knowledge of the causal structure of the data into the learning objective. Specifically, causal intervention is used to curb dependence on shortcuts (e.g., "good → positive") and improve the ability to reason *causal effect* between sentence and sentiment.

In this paper, we rethink the ISA task from a causal perspective and unitize the casual intervention on deep learning. We argue that the causal effect obtained by reasoning directly from the sen-

*Corresponding authors.

Proceedings of the 29th International Conference on Computational Linguistics, pages 6966–6977
October 12–17, 2022.

tence (X) to the sentiment (Y) without relying on other extra prior stereotypical lexical impressions is closer to the original semantic analysis. Our work aims at eliminating the confounding causal effects of $C \rightarrow Y$ and thus extracting the pure causal effect between sentence and sentiment. Inspired by the instrumental variable in causality, we propose a CausaL intervention model for implicit sEntiment ANalysis using instrumental variable (CLEAN). Different from the other work with causal intervention like back-door adjustment (Landeiro and Culotta, 2016), other variables like confounders are not required to be observed. CLEAN contains two-stage learning: (1) In the first stage, we model the relationship between the instrumental variable and sentence; (2) In the second stage, we dismiss the spurious correlation between confounders and sentiment by means of the relationship obtained from the first stage.

To evaluate the effectiveness of our CLEAN, we conduct a series of experiments on both the general implicit sentiment analysis and aspect-based implicit sentiment analysis. In particular, we compare our model with several the mainstream baselines and the results show the great advantages of our model on ISA. We also validate the robustness of the model by adversarial attack and case studies, which proves that our model successfully dismisses the spurious correlation caused by sentiment words and extracts the pure causal effect.

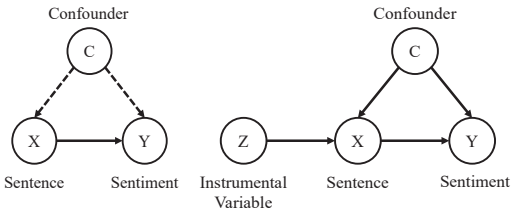
The main contributions are summarized as follows.

- We rethink the implicit sentiment from a causal perspective and proposed a casual intervention model for implicit sentiment analysis (CLEAN).
- To remove the spurious causes of confounders, we incorporate instrumental variable into neural network to enhance its causal reasoning ability.
- We conduct experiments on diverse datasets, including partially implicit and totally implicit sentiment, which shows our effectiveness and rationality to reason implicit sentiment.

2 Preliminaries

2.1 Structural Causal Model and Causal Effect

In our paper, Structural Causal Model (SCM) (Glymour et al., 2016) is represented as a directed acyclic graphs (DAGs) $G = \{V, E\}$ to reflect causal relationships, where V denotes the set of observational variables and E denotes the direct causal effect (Figure 2(a)). X is a direct cause of Y when variable Y is the child of X.



(a) Causal Graph (b) Instrument Variable for Causal Intervention

Figure 2: Causal Graph

Variable X and Y is called treatment variable and outcome variable respectively when observing the causal relationship between them. The other variables we do not consider their causal relationship are called error terms (ε), also known as exogenous variables. Significantly, total effect between X and Y, denoted as $P(Y \mid X)$, is conceptually different from causal effect of $X \rightarrow Y$, denoted as $P(Y \mid do(X = x))$ because the causal effect only involves the direct path from X to Y, while the total effect involves all paths between X and Y. Based on the ideal hypothesis that none of the error terms will involve in the path between X and Y, people usually treat the total effect and the causal effect as the same. But the actual fact is that a part of error terms (we call it confounder (C)) serves as a common cause of the treatment and outcome, denoted as $X \leftarrow C \rightarrow Y$. Consequently, the total effect is practically different from the causal effect, i.e. $P(Y \mid do(X = x)) \neq P(Y \mid X)$ and treatment-outcome relationship may well be obscured by the spurious correlation between C and Y generated by confounder (Pearl, 2009; Hernán MA, 2020).

2.2 Instrument Variable for Causal Intervention

To recover the gap between total effect $P(Y \mid X)$ and casual effect $P(Y \mid do(X = x))$ and derive pure causal effect, we must "adjust" for potential confounder (C) (Pearl, 1995). Fortunately, applying causal intervention can extract the pure causality from the correlation and therefore overcome the problem of confounding bias. There are four key interventions: randomized controlled trial, backdoor adjustment, front-door adjustment, and instrumental variable estimation. Randomized con-

trolled trials are simply not practicable for natural language, and both the front-door and back-door adjustment require additional observable variables. However, the confounders (e.g., rhetoric confounding word, such as rhetorical questions and sarcasm) are too polymorphic to be observed exhaustively for implicit sentiment analysis. We adopt the instrumental variable to dismiss the spurious correlations instead of directly observing confounders (Figure 2(b)).

Before the intervention, we should find a suitable instrumental variable (Z) that qualifies well the requirements as follows:(Brito and Pearl, 2012)

1. Z is independent of all error terms ε that have an influence on Y which is not mediated by X, $Cov(Z, \varepsilon) = 0$.

2. Z is not independent of X, $Cov(Z, X) \neq 0$.

The intuition behind this definition is that all correlation between Z and Y requires X to act as an intermediary.

Generally, instrument variable estimation contains two stages (Angrist and Pischke, 2008). In the first stage, the coefficient α is obtained by regression estimation of X and Z, denoted as $Cov(Z, X)$. In the second stage, substitute X with the expressions including Z obtained in the first stage into the expression of Y and then regress Y on Z , denoted as $Cov(Z, Y)$. There is no confounding bias between Y and Z owing to the restriction in the definition of Z, i.e. $Cov(Z, \varepsilon) = 0$. A simple linear model for IV estimation consists of 2 equations:

$$X = \alpha Z + \varepsilon_X; Y = \omega X + \varepsilon_Y \tag{1}$$

where Y is the outcome variable (e.g., sentiment), X is the treatment variable (e.g, sentence), Z is the instrumental variable (e.g., stochastic perturbation), and ε_X and ε_Y are error terms including but not limited to confounders(C). Under the conditions above, it can be proved that the equation presents an asymptotically unbiased estimate of the effect of X on Y (Angrist et al., 1996).

$$\omega_{IV} = \frac{\frac{1}{n}\sum_{i=1}^{n}(z_i - \overline{z})(y_i - \overline{y})}{\frac{1}{n}\sum_{i=1}^{n}(z_i - \overline{z})(x_i - \overline{x})} = \frac{Cov(Z, Y)}{Cov(Z, X)} \tag{2}$$

3 Our Approach

In this section, we introduce our CLEAN model for implicit sentiment analysis (Figure 3). We first rethink the ISA from a causal perspective (Section 3.1). Then, we adopt stochastic perturbation as instrumental variable (Section 3.2) and estimate instrumental variable in two stages (Section 3.3 and 3.4).

3.1 Sentiment Analysis from Casual Perspective

Given a sentence X, consisting of a sequence of tokens $(x_1, x_2, ..., x_n)$, our task aims to analyze the polarity Y. For aspect-based sentiment analysis task, we concatenate the sentence and the aspect as the input $X = (x_1, x_2, ..., x_n, [SEP], a_1, a_2, ..., a_m)$. In the current method, a deep neural network is used as a classifier to predict the sentiment polarity label as output and the sentence as input (as Equation 3).

$$y = h(x; w) = W_{xy} \cdot x + \varepsilon_y \tag{3}$$

where ε_y denotes as the error terms including confounders (c) and other errors ($\hat{\varepsilon_y}$).

The prediction above is based on the hypothesis that error terms will not involve in the causal path between X and Y and ignore the influence of error terms mostly. Nevertheless, several research has found that text classification systems based on neural networks are biased towards learning frequent spurious correlations (Leino et al., 2018). It urges us to focus on the longtime unheeded but unavoidable existence of confounder (C) in error terms, which results in the overlooked gap between *total effect* and *causal effect*, denoted as the path $X \leftarrow C \rightarrow Y$. The Equation 3 can be updated in consideration of confounder (c) (Equation 4).

$$y = h(x; w) = W_{xy} \cdot x + W_{cy} \cdot c + \hat{\varepsilon_y} \tag{4}$$

where c and $\hat{\varepsilon_y}$ denotes the confounder and other error terms respectively, W_{cy} denotes the causal effect of $C \rightarrow Y$.

In previous studies, gender (Field and Tsvetkov, 2020), age, and address (Landeiro and Culotta, 2016) were found to be confounders in text classification. As for ISA, we focus rather on the naturally existing confounder within the text, i.e., sentiment words. Sentiment words affect the form of the text as a component of the text (i.e., the writer's word choice determines the form of expression) and also affect sentiment expression (Xing et al., 2020). The existence of sentiment words as confounder makes it difficult to distinguish the pure causal effect of $X \rightarrow Y$ and the prediction indiscriminately depends on the spurious correlation

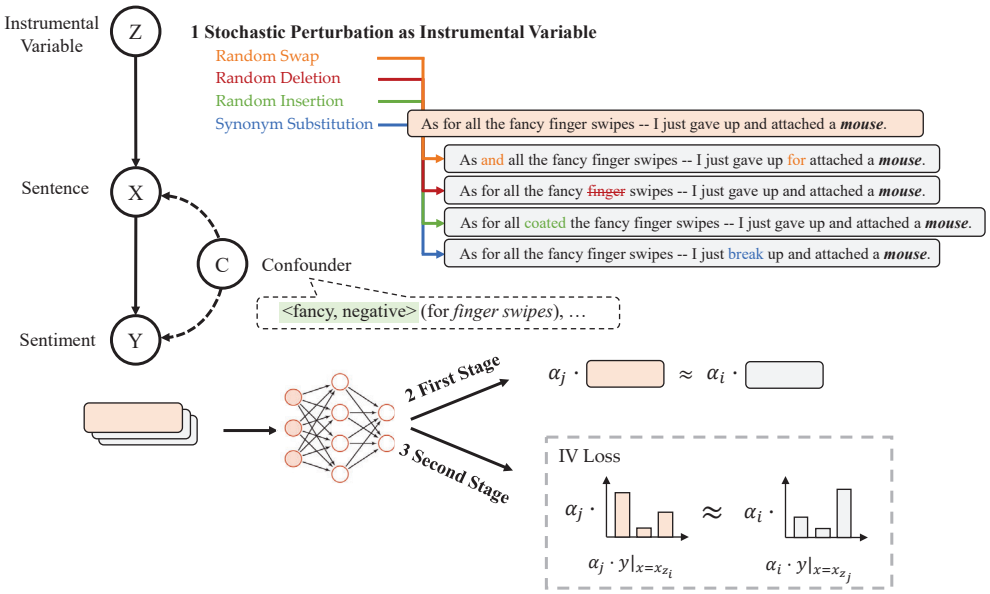


Figure 3: The framework of our CLEAN.

between sentiment words and sentiment will fail in ISA. The main forms are as follows:

• **Inter-aspect Confounding Word** Explicit sentiment words of other aspects with opposite sentiment in the sentence confounds the prediction effect of the current aspect.

• **Inter-clause Confounding Word** In an adversative compound sentence, the other clauses with opposite semantics confound the prediction effect of the current clause.

• **Rhetoric Confounding Word** Sentiment words conveying the opposite of the norm in rhetorical devices such as rhetorical questions and sarcasm confound the prediction effect.

• **Dynamic Neural Confounding Word** Neutral words show dynamic sentiment polarity in different contexts, but the model trained by biased data only learns the spurious correlation of one polarity.

We also provide a detailed analysis of the confounder in case studies (Section 5.2).

Inspired by causal intervention with instrumental variable (Section 2.2), we adopt two-stage instrument variable estimation for ISA to achieve the goal that distinguishes the pure *causal effect* of $X \rightarrow Y$ without any spurious correlations, denoted as $P(Y \mid do(X = x))$.

3.2 Stochastic Perturbation as Instrumental Variable

For text, the two restrictions of instrumental variable could be translated into two basic opinions: (1) instrumental variable Z will not influence the sentiment polarity via any other casual path except through sentence X; (2) instrumental variable Z will influence the format of sentence X. Intuitively, we choose the stochastic perturbation as the instrumental variable of ISA. Inspired by the work of data augmentation (Guohang et al., 2020), we choose random swap, random deletion, random insertion, and synonym substitution as stochastic perturbation: A) **Random Swap**: Swap word randomly; B) **Random Deletion**: Delete word randomly with probability p; C) **Random insertion**: Insert word randomly by word embeddings similarity; D) **Synonym Substitution**: Substitute word by WordNet's synonym. It fortunately meets the requirements of instrumental variable well: (1) stochastic perturbation obviously has no independent effect on sentiment, except through augmentation sentences, i.e. $Cov(Z, \varepsilon) = 0$; (2) stochastic perturbation above will definitely change the sentence into another form, i.e. $Cov(Z, X) \neq 0$.

3.3 The First Stage of CLEAN

Following the traditional pattern of instrumental variable estimation (Section 2.2), the first stage of CLEAN is to establish the causal relationship between stochastic perturbation (Z) and sentence (X), i.e. $Z \rightarrow X$. We use two open source tools[1] to generate augmentation samples x_z from the original sample x and the formal expression is as follows.

$$x_z = f(x, z)$$

where $f(\cdot)$ denotes the different stochastic perturbation on the original sentence. For a specific stochastic perturbation z_i, we have $x_{z_i} = f(x, z_i) \approx \alpha_i \cdot x$.

[1] https://github.com/jasonwei20/eda_nlp
https://github.com/makcedward/nlpaug

Table 1: The statistics information of the datasets. IS means the percent of samples that are implicit sentiment.

Dataset	Postive		Neural		Negative		IS (%)	
	Train	Test	Train	Test	Train	Test	Train	Test
Restaurant	2164	728	805	196	633	196	28.59	23.84
Laptop	987	341	866	128	460	169	30.87	27.27
CLIPEval	435	144	205	72	640	155	100.00	100.00

To obtain the accurate value of α which can well represent the relationship between x and x_z, a neural network was constructed based on the BERT and α was set as a self-learning parameter.

$$\alpha_i = \min_{\alpha} \sum_{x_{z_i}=f(x,z_i)} \parallel \mathcal{M}(x_{z_i}) - \alpha \cdot \mathcal{M}(x) \parallel$$

where $\mathcal{M}$ denotes a text encoder (e.g., BERT).

3.4 The Second Stage of CLEAN

Substituting the relation above between original sample x and augmentation sample x_{z_i} into Equation 4, we will get the $y \mid_{x=x_{z_i}}$ with different proportionality coefficient α obtained from the first stage.

$$\begin{aligned} y \mid_{x=x_{z_i}} &= W_{xy} \cdot x_{z_i} + W_{cy} \cdot c + \hat{\varepsilon_y} \\ &= \alpha_i \cdot W_{xy} \cdot x + W_{cy} \cdot c + \hat{\varepsilon_y} \\ &= \alpha_i \cdot y \mid_{do(x=x)} + W_{cy} \cdot c + \hat{\varepsilon_y} \end{aligned} \tag{5}$$

where $y \mid_{do(x=x)}$ denotes the prediction only along the path $X \rightarrow Y$.

As the neural network is not totally linear, we slightly generalize the usage of two stages in instrumental variable. We set the dismission of influence of the confounder as a regularization function instead of directly calculating the effect between X and Y as a single value (Equation 2), which is obviously more fit for deep learning method.

$$\mathcal{L}_{IV} = \sum_{i \neq j} \parallel \alpha_j \cdot y \mid_{x=x_{z_i}} - \alpha_i \cdot y \mid_{x=x_{z_j}} \parallel$$

The reason we just model the prediction $y \mid_{x=x_{z_i}}$ and unitize the regularization loss on it is that the essence of the $\mathcal{L}_{IV}$ is to force the model to suppress the confounding effect caused by sentiment words. It can be easily proved by substituting the $y \mid_{x=x_{z_i}}$ with Equation 5 obtained by two-stage learning. The benefit is obviously that we can suppress the confounding effect without directly observing the confounders (c).

$$\mathcal{L}_{IV} = \sum_{i \neq j} \parallel (\alpha_i - \alpha_j) \cdot (W_{cy} \cdot c + \hat{\varepsilon_y}) \parallel$$

In addition, the model should not go to the other extreme, i.e., ignore sentiment words entirely, which would be inconsistent with the normal process of natural language understanding. We set a hyperparameter β to achieve balance and combine the causal regularization loss function $\mathcal{L}_{IV}$ with the conventional cross-entropy loss $\mathcal{L}_{CE}$ and the influence of the β is analyzed in Section 5.3.

$$\mathcal{L}_{ALL} = \mathcal{L}_{CE} + \beta \mathcal{L}_{IV}$$

4 Experiment

4.1 Datasets and Metrics

Implicit Sentiment Analysis To show our model's better performance in understanding implicit sentiment, we evaluate the implicit polarity prediction on a total implicit dataset, CLIPEval from SemEval 2015 task 9 (Russo et al., 2015), which consists of self-reported entity reviews collected from psychological research with 1,280 sentences for the training and 371 for the test.

Aspect-based Implicit Sentiment Analysis As our aim to dismiss the spurious correlation between explicit sentiment words and polarity, we also conducted experiments on both explicit and implicit datasets, Laptop and Restaurant review from SemEval 2014 task 4 (Pontiki et al., 2014). The segmentation of explicit sentiment (ESE) and implicit sentiment (ISE) is based on the work of (Li et al., 2021b) based on the annotation of opinion words (Fan et al., 2019).

We adopt two widely used metrics accuracy and macro-F1 to evaluate the performance of our model and the baselines.

4.2 Baselines

To investigate the effectiveness of our CLEAN model, we compare it with several typical baseline models for implicit sentiment analysis and aspect-based implicit sentiment analysis.

Implicit Sentiment Analysis We select four popular baselines for implicit sentiment analysis, which are listed as follows. SHELIFBK (Dragoni,

Table 2: The main results for aspect-based sentiment analysis. For ESE and ISE, we provide the F1 score. We use the results reported in (Li et al., 2021b). The baselines with † are our implementation.

		Restaurant				Laptop			
		Acc	F1	ESE	ISE	Acc	F1	ESE	ISE
NN	ATAE-LSTM	76.90	62.64	84.16	53.71	65.37	62.92	75.69	37.86
	IAN	76.88	67.71	86.52	46.07	67.24	63.72	75.86	44.25
	RAM	80.23	70.80	85.11	55.81	74.49	71.35	75.86	44.25
	MGAN	81.25	71.94	85.18	60.04	75.39	72.47	76.16	56.31
GNN	TransCap	79.55	71.41	86.52	59.93	73.87	70.10	77.16	60.34
	ASGCN	80.77	72.02	84.29	62.91	75.55	71.05	75.46	57.77
	BiGCN	81.97	73.48	87.19	59.05	74.59	71.84	79.53	62.64
	CDT	82.30	74.02	88.79	65.87	77.19	72.99	77.53	68.90
	RGAT	83.30	76.08	89.45	61.05	77.42	73.76	80.17	65.52
BERT	BERT-SPC	83.57	77.16	89.21	65.54	78.22	73.45	81.47	69.54
	CapsNet+BERT	85.09	77.75	91.68	64.04	78.21	73.34	82.33	67.24
	BERT-PT	84.95	76.96	92.15	64.79	78.07	75.08	81.47	71.27
	BERT-ADA	**87.14**	80.05	**94.14**	65.92	78.96	74.18	**82.76**	70.11
	R-GAT+BERT	86.60	81.35	92.73	67.79	78.21	74.07	82.44	72.99
	TransEncAsp	77.10	57.92	86.97	48.96	65.83	59.53	74.31	43.20
	TransEncAsp+SCAPT	83.39	74.53	88.04	68.55	77.17	73.23	78.70	72.82
	BERT-SPC†	85.09	77.19	91.68	64.04	77.90	73.50	80.99	69.71
	BERT-SPC† (Aug4)	84.20	76.55	90.50	64.04	76.65	70.86	81.64	63.43
	BERT-SPC† (Aug8)	80.98	67.77	90.39	50.94	75.71	71.62	77.97	69.71
	BERT-SPC† (Aug16)	77.59	67.44	85.35	52.81	74.61	69.92	77.97	65.71
Ours	CLEAN	87.05	**81.40**	92.50	**69.66**	**80.41**	**77.25**	81.21	**78.29**

2015), ATTLSTM (Lin et al., 2017), MTL (Zheng et al., 2018), BERT-SPC (Xu et al., 2019).

Aspect-based Implicit Sentiment Analysis The commonly used baselines can be split into three parts, neural network, graph neural network, and BERT-based models, which are given as follows.

Neural Network: ATAE-LSTM (Wang et al., 2016), IAN (Ma et al., 2017), RAM (Chen et al., 2017), MGAN (Fan et al., 2018).

Graph Neural Network: TransCap (Chen and Qian, 2019), ASGCN (Zhang et al., 2019), BiGCN (Zhang and Qian, 2020), CDT (Sun et al., 2019), RGAT (Wang et al., 2020).

BERT-based Models: BERT-SPC (Xu et al., 2019), CapsNet+BERT (Jiang et al., 2019), BERT-ADA (Rietzler et al., 2020), R-GAT+BERT (Wang et al., 2020), TransEncAsp (Li et al., 2021b), TransEncAsp+SCAPT (Li et al., 2021b).

Moreover, to explore the influence of the augmentation sentences, we add them into the training dataset for our basic model (BERT-SPC). For example, BERT-SPC (Aug4) means we add four augmentation samples for each example.

4.3 Implementation Details

We implement CLEAN with PyTorch based on Hugging Face Transformer [2] and run them on the GPU(NVIDIA GTX 2080ti). During training, we set the coefficient λ of $\mathcal{L}_2$ regularization item is 0.01, 10^{-5} and dropout rate is 0.1. The learning rate is set as 2e-5 and the batch size is set as 16. Adam optimizer (Kingma and Ba, 2014) is used to update all the parameters.

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

5 Experimental Analysis

5.1 Main Results

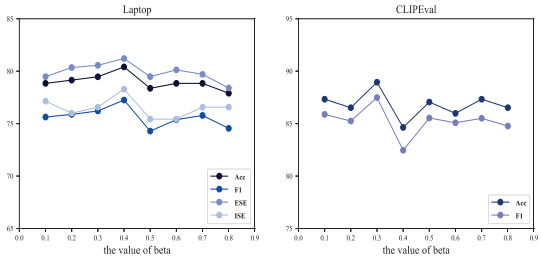
To evaluate the performance of our CLEAN model, we compare it with several mainstream baseline models for both the implicit sentiment analysis and aspect-based implicit sentiment analysis (Table 3 and Table 2). We find the following observation from these tables. **First**, our model outperforms all the baselines in most cases. Particularly, our model obtains the best F1 scores over all the three datasets of two tasks. **Second**, our CLEAN strategy significantly improves the performance of the baseline. CLEAN improves more than two points in terms of F1 over all the datasets compared with BERT-SPC, which is the base of our model. **Third**, our model can improve the performance of implicit sentiment analysis effectively. For example, compared with the BERT-SPC†, we obtain more than five points improvement on ISE over both Restaurant and Laptop. Also, we obtain the best results of implicit sentiment analysis over CLIPEval. **Forth**, the model that regards the augmentation sentence as a data augmentation (e.g., BERT-SPC† (Aug4)) performs even worse than the one without augmentation as noise may exist. This shows that our

	Sentence Example	Target	BERT-SPC	ISAIV
E1	The food was definitely good , but when all was said and done, I just couldn't justify it for the **price** (including 2 drinks, $100/person)...	price	positive ✗	negative ✓
E2	And as for all the fancy finger swipes -- I just gave up and attached a **mouse**.	mouse	negative ✗	neutral ✓
E3	I was a little concerned about the **touch pad** based on reviews, but I've found it fine to work with.	touch pad	negative ✗	positive ✓
E4	How can hope to stay in business with **service** like this?	service	positive ✗	negative ✓
E5	The **steak** melted in my mouth.	steak	negative ✗	positive ✓
E6	15% **gratuity** automatically added to the bill.	gratuity	positive ✗	positive ✗

Figure 4: Some examples of case studies.

Table 3: The main results for implicit sentiment analysis. We use the results reported in (Xiang et al., 2021). The baselines with † are our implementation.

Method	CLIPEval	
	Acc	F1
SHELLFBK	56.00	54.00
ATTLSTM	82.43	82.21
MTL	82.94	83.17
BERT-SPC†	87.06	84.74
BERT-SPC† (Aug4)	85.71	83.56
BERT-SPC† (Aug8)	86.52	85.30
BERT-SPC† (Aug16)	85.44	84.54
CLEAN	**88.95**	**87.49**



(a) Laptop (b) CLIPEval

Figure 5: The influence of β

CLEAN algorithm can help learn the implicit sentiment reasoning behind the data.

5.2 Case Studies

We present five samples in Figure 4 to explain the four main types of confounders (Section 3.1), which shows the effectiveness and rationality of our model to reason implicit sentiment. (1) **Inter-aspect Confounding Word.** In example E1, "definitely good" is the sentiment words of aspect *food*, implying positive sentiment but confounds the prediction of aspect *price*. In E2, the user expresses a negative sentiment towards aspect *finger swipes* with opinion word "fancy", which confounds the prediction of aspect *mouse*. (2) **Inter-clause Confounding Word.** In E3, the first and second clauses form an adversative relation, and the true meaning of the expression is that the *mouse* works well, but the sentiment word "a little concerned" in the first clause confounds the prediction. (3) **Rhetoric Confounding Word.** In E4, the customer used the rhetorical device of a rhetorical question to express that the restaurant's service was terrible, but the existence of the word "hope" confounds the prediction, (4) **Dynamic Neural Confounding Word.** In E5, the word "melted" is absolutely a neutral word, but when we directly count the proportion of aspect-level sentiment polarity that co-occur with "melted", we surprisingly find that 83.33% aspect polarity is negative, which well explains why the previous model predicts "negative" strangely. Due to the unbalanced distribution of training data, the model tends to tag the neural word with specific sentiment polarity and predict based on this spurious correlation learned superficially before.

5.3 Further analysis

Influence of Augmentation Sample Number. We explore the influence of augmentation sample number here (Table 4). The influence of sample number on model performance depends on two conflicting factors: the degree of deviation from the original sentence and the chance to find more potential confounders. With the increase in sample number, the model has a greater chance of finding more potential confounders and adjusting for them. On the other hand, a larger generation samples number means that more samples deviating from the original sentence are involved in the learning procedure, and therefore the accuracy of prediction decreases. Over Restaurant, the two conflicting factors reach a better balance at 8; while on Laptop, the negative effect of semantic deviation outweighs the positive

Table 4: The influence of augmentation sample number.

#Num	Restaurant 0.6				Laptop 0.4			
	Acc	F1	ESE	ISE	Acc	F1	ESE	ISE
4	86.88	80.99	92.50	68.91	**80.41**	**77.25**	**81.21**	**78.29**
8	**87.05**	**81.40**	**92.50**	**69.66**	78.68	75.23	79.70	76.00
16	85.09	77.75	91.21	65.54	78.06	75.04	79.05	75.43

Table 5: The results of robustness.

Model	Restaurant (Trans.)		Laptop (Trans.)	
	Acc	F1	Acc	F1
BERT-SPC	57.04	44.43	51.05	41.01
CLEAN	58.77	48.56	59.45	43.24

effect of correction for more confounders, and the best result is achieved at 4.

Influence of β. Either emphasizing sentiment words only or completely ignoring them is not reasonable. The purpose of our hyper-parameter beta is to strike a balance between these two terms (Figure 5). In Laptop and CLIPEval, performance is best at 0.4 and 0.3 relatively, and both show a trend of high in the middle and low on both sides, indicating that our hypothesis is rational.

Robustness. We also analyze the robustness of our proposed CLEAN (Table 5). We test our model on a robustness testing dataset, Revnon of TextFlint (Wang et al., 2021), which reverses the sentiment of the non-target aspects with originally the same sentiment as target. Our model outperforms the model BERT-SPC without causal intervention, which means CLEAN can also improve the robustness by learning the implicit sentiment reasoning.

Limitation. We also analyze wrong samples and find the model may fail when encountering the expression with unusual knowledge. In Figure 4 E6, due to the lack of prior knowledge about “gratuity", “automatically added" is likely to be perceived as a good thing. Admittedly, our work mainly focuses on the reasoning ability and doesn’t integrate external corpus and knowledge and therefore lacks abundant prior knowledge. The better combination of prior knowledge and causal inference is also an intriguing and worth exploring field.

6 Related Work

6.1 Implicit Sentiment Analysis

Implicit sentiment analysis (ISA) task plays an important role in sentiment analysis field (Liu, 2012; Zhou et al., 2019, 2020c). Early studies mainly trained machine learning models based on handcrafted features or explicit characterization of implicit feature information. Some studies argued that seemingly neural words actually contain emotional content and then construct a lexicon (Feng et al., 2013; Castelló and Stede, 2017). Label propagation was used to judge the affective polarity of the words automatically (Ding and Riloff, 2016; Li et al., 2021a). Moreover, Balahur et al. (2011) proposed to build a commonsense knowledge base (EmotiNet) with the concept of affective value and the sentiment.

Recent efforts (He et al., 2018; Tang et al., 2020) used syntax information from dependency trees to enhance attention-based models. Using syntactic analysis tree and CNN, Liao et al. (2019) analyzed fact-implied implicit sentiment by fusing multi-level semantic information, including sentiment target, sentence, and context semantic. A lot of works (Zhang et al., 2019; Sun et al., 2019; Wang et al., 2020) incorporated tree-structured syntactic information via graph neural networks to capture aspect-aware information in text. Another method is to utilize external corpus and pre-trained knowledge to enhance semantic awareness of models (Xu et al., 2019; Rietzler et al., 2020; Dai et al., 2021; Li et al., 2021b; Zhou et al., 2020b).

The existing methods mainly improve the ISA by integrating external corpus and knowledge. However, the knowledge is always not complete which will influence the models’ performance. In this paper, we solve it via causal intervention to learn the reasoning behind the sentiment classification.

6.2 Causality for NLP

Recently, some researchers are beginning to combine causality and NLP to create more robust and interpretable models (Wood-Doughty et al., 2018; Tang et al., 2021). Most of the papers integrated backdoor and counterfactual into NLP tasks. Particularly, Landeiro and Culotta (2016) applied the back-door adjustment to text classification by controlling the artificially predetermined confounding variable. Feng et al. (2021) introduced counterfactual reasoning into the model learning process by generating representative counterfactual samples

and comparing the counterfactual and factual samples. Veitch et al. (2021) utilized distinct regularization schema for distinct causal structure to induce counterfactual invariance. Niu et al. (2021) utilized the counterfactual inference on VQA models by subtracting the language bias as direct language effect from the total causal effect.

Different from these studies, we explore the causal graph for ISA and incorporate it using the causal intervention.

7 Conclusion

In this paper, we proposed a causal intervention model for implicit sentiment analysis using instrument variable (CLEAN). Given that the current model indiscriminately focuses on the correlation between sentiment and sentiment words and consequently performs poorly in implicit sentiment analysis as the explicit sentiment words disappear, we rethink the implicit sentiment analysis from a causal perspective and analyze the four main forms of sentiment words as potential confounders. Inspired by the instrumental variable of causal intervention, we adopt stochastic perturbation as instrumental variable and construct a model with two-stage learning. Across three different datasets, including general implicit sentiment analysis and aspect-based sentiment analysis, our CLEAN shows great advantages in implicit sentiment.

Acknowledge

The authors wish to thank the reviewers for their helpful comments and suggestions. This work was partially funded by the National Natural Science Foundation of China (No. 61976056, 62076069), Shanghai Municipal Science and Technology Major Project (No.2021SHZDZX0103).

References

Joshua D Angrist, Guido W Imbens, and Donald B Rubin. 1996. Identification of causal effects using instrumental variables. *Journal of the American statistical Association*, 91(434):444–455.

Joshua D Angrist and Jörn-Steffen Pischke. 2008. Mostly harmless econometrics. In *Mostly Harmless Econometrics*. Princeton university press.

Alexandra Balahur, Jesús M Hermida, and Andrés Montoyo. 2011. Detecting implicit expressions of sentiment in text based on commonsense knowledge. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA 2.011)*, pages 53–60.

Carlos Brito and Judea Pearl. 2012. Generalized instrumental variables. *arXiv preprint arXiv:1301.0560*.

Peter Bühlmann. 2020. Invariance, causality and robustness. *Statistical Science*, 35(3):404–426.

Núria Bertomeu Castelló and Manfred Stede. 2017. Extracting word lists for domain-specific implicit opinions from corpora. In *IWCS 2017 - 12th International Conference on Computational Semantics - Long papers*.

Peng Chen, Zhongqian Sun, Lidong Bing, and Wei Yang. 2017. Recurrent attention network on memory for aspect sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 452–461. Association for Computational Linguistics.

Zhuang Chen and Tieyun Qian. 2019. Transfer capsule network for aspect level sentiment classification. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 547–556. Association for Computational Linguistics.

Junqi Dai, Hang Yan, Tianxiang Sun, Pengfei Liu, and Xipeng Qiu. 2021. Does syntax matter? a strong baseline for aspect-based sentiment analysis with RoBERTa. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1816–1829, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *north american chapter of the association for computational linguistics*.

Haibo Ding and Ellen Riloff. 2016. Acquiring knowledge of affective events from blogs using label propagation. In *Thirtieth AAAI Conference on Artificial Intelligence*.

Mauro Dragoni. 2015. SHELLFBK: an information retrieval-based system for multi-domain sentiment analysis. In *Proceedings of the 9th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2015, Denver, Colorado, USA, June 4-5, 2015*, pages 502–509. The Association for Computer Linguistics.

Feifan Fan, Yansong Feng, and Dongyan Zhao. 2018. Multi-grained attention network for aspect-level sentiment classification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 3433–3442. Association for Computational Linguistics.

Zhifang Fan, Zhen Wu, Xinyu Dai, Shujian Huang, and Jiajun Chen. 2019. Target-oriented opinion words extraction with target-fused neural sequence labeling. *north american chapter of the association for computational linguistics*.

Amir Feder, Katherine A Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, Justin Grimmer, Roi Reichart, Margaret E Roberts, et al. 2021. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *arXiv preprint arXiv:2109.00725*.

Fuli Feng, Jizhi Zhang, Xiangnan He, Hanwang Zhang, and Tat-Seng Chua. 2021. Empowering language understanding with counterfactual reasoning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2226–2236, Online. Association for Computational Linguistics.

Song Feng, Jun Seok Kang, Polina Kuznetsova, and Yejin Choi. 2013. Connotation lexicon: A dash of sentiment beneath the surface meaning. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1774–1784, Sofia, Bulgaria. Association for Computational Linguistics.

Anjalie Field and Yulia Tsvetkov. 2020. Unsupervised discovery of implicit gender bias. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 596–608. Association for Computational Linguistics.

Madelyn Glymour, Judea Pearl, and Nicholas P Jewell. 2016. *Causal inference in statistics: A primer*. John Wiley & Sons.

Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A survey of methods for explaining black box models. *ACM Computing Surveys*.

Liu Guohang, Zhang Shi-bin, Tang Haozhe, Yang Lu, Jiazhong Lu, and Huang Yuanyuan. 2020. Easy data augmentation method for classification tasks. *active media technology*.

Ruidan He, Wee Sun Lee, Hwee Tou Ng, and Daniel Dahlmeier. 2018. Effective attention modeling for aspect-level sentiment classification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1121–1131, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Robins JM Hernán MA. 2020. *Causal Inference: What If*. Boca Raton: Chapman & Hall/CRC.

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 2019, Hong Kong, China, November 3-7, 2019*, pages 6279–6284. Association for Computational Linguistics.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Virgile Landeiro and Aron Culotta. 2016. Robust text classification in the presence of confounding bias. In *Thirtieth AAAI Conference on Artificial Intelligence*.

Klas Leino, Emily Black, Matt Fredrikson, Shayak Sen, and Anupam Datta. 2018. Feature-wise bias amplification. *arXiv preprint arXiv:1812.08999*.

Qizhi Li, Xianyong Li, Yajun Du, and Xiaoliang Chen. 2021a. Iswr: An implicit sentiment words recognition model based on sentiment propagation. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 248–259. Springer.

Zhengyan Li, Yicheng Zou, Chong Zhang, Qi Zhang, and Zhongyu Wei. 2021b. Learning implicit sentiment in aspect-based sentiment analysis with supervised contrastive pre-training. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 246–256, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jian Liao, Suge Wang, and Deyu Li. 2019. Identification of fact-implied implicit sentiment based on multi-level semantic fused representation. *Knowledge-Based Systems*, 165:197–207.

Zhouhan Lin, Minwei Feng, Cícero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. 2017. A structured self-attentive sentence embedding. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Bing Liu. 2012. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167.

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

Dehong Ma, Sujian Li, Xiaodong Zhang, and Houfeng Wang. 2017. Interactive attention networks for aspect-level sentiment classification. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, pages 4068–4074. ijcai.org.

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.

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

Judea Pearl. 2009. *Causality*, 2 edition. Cambridge University Press.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *north american chapter of the association for computational linguistics*.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. Semeval-2014 task 4: Aspect based sentiment analysis. *international conference on computational linguistics*.

Suhas Ranganath, Xia Hu, Jiliang Tang, Suhang Wang, and Huan Liu. 2018. Understanding and identifying rhetorical questions in social media. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 9(2):1–22.

Alexander Rietzler, Sebastian Stabinger, Paul Opitz, and Stefan Engl. 2020. Adapt or get left behind: Domain adaptation through BERT language model finetuning for aspect-target sentiment classification. In *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 4933–4941. European Language Resources Association.

Irene Russo, Tommaso Caselli, and Carlo Strapparava. 2015. Semeval-2015 task 9: Clipeval implicit polarity of events. *SemEval-2015*, page 443.

Kai Sun, Richong Zhang, Samuel Mensah, Yongyi Mao, and Xudong Liu. 2019. Aspect-level sentiment analysis via convolution over dependency tree. 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 2019, Hong Kong, China, November 3-7, 2019*, pages 5678–5687. Association for Computational Linguistics.

Hao Tang, Donghong Ji, Chenliang Li, and Qiji Zhou. 2020. Dependency graph enhanced dual-transformer structure for aspect-based sentiment classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6578–6588, Online. Association for Computational Linguistics.

Kaihua Tang, Mingyuan Tao, and Hanwang Zhang. 2021. Adversarial visual robustness by causal intervention. *arXiv preprint arXiv:2106.09534*.

Victor Veitch, Alexander D'Amour, Steve Yadlowsky, and Jacob Eisenstein. 2021. Counterfactual invariance to spurious correlations: Why and how to pass stress tests. *arXiv preprint arXiv:2106.00545*.

Kai Wang, Weizhou Shen, Yunyi Yang, Xiaojun Quan, and Rui Wang. 2020. Relational graph attention network for aspect-based sentiment analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3229–3238, Online. Association for Computational Linguistics.

Xiao Wang, Qin Liu, Tao Gui, Qi Zhang, Yicheng Zou, Xin Zhou, Jiacheng Ye, Yongxin Zhang, Rui Zheng, Zexiong Pang, et al. 2021. Textflint: Unified multilingual robustness evaluation toolkit for natural language processing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 347–355.

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.

Zhao Wang and Aron Culotta. 2020. Identifying spurious correlations for robust text classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3431–3440.

Zach Wood-Doughty, Ilya Shpitser, and Mark Dredze. 2018. Challenges of using text classifiers for causal inference. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2018, page 4586. NIH Public Access.

Chunli Xiang, Junchi Zhang, and Donghong Ji. 2021. A message-passing multi-task architecture for the implicit event and polarity detection. *PloS one*, 16(3):e0247704.

Xiaoyu Xing, Zhijing Jin, Di Jin, Bingning Wang, Qi Zhang, and Xuan-Jing 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.

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.

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 2019, Hong Kong, China, November 3-7, 2019, pages 4567–4577. Association for Computational Linguistics.

Mi Zhang and Tieyun Qian. 2020. Convolution over hierarchical syntactic and lexical graphs for aspect level sentiment analysis. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 3540–3549. Association for Computational Linguistics.

Renjie Zheng, Junkun Chen, and Xipeng Qiu. 2018. Same representation, different attentions: Shareable sentence representation learning from multiple tasks. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, pages 4616–4622. ijcai.org.

Jie Zhou, Qin Chen, Jimmy Xiangji Huang, Qinmin Vivian Hu, and Liang He. 2020a. Position-aware hierarchical transfer model for aspect-level sentiment classification. *Information Sciences*, 513:1–16.

Jie Zhou, Jimmy Xiangji Huang, Qin Chen, Qinmin Vivian Hu, Tingting Wang, and Liang He. 2019. Deep learning for aspect-level sentiment classification: survey, vision, and challenges. *IEEE access*, 7:78454–78483.

Jie Zhou, Jimmy Xiangji Huang, Qinmin Vivian Hu, and Liang He. 2020b. Sk-gcn: Modeling syntax and knowledge via graph convolutional network for aspect-level sentiment classification. *Knowledge-Based Systems*, 205:106292.

Jie Zhou, Junfeng Tian, Rui Wang, Yuanbin Wu, Wenming Xiao, and Liang He. 2020c. Sentix: A sentiment-aware pre-trained model for cross-domain sentiment analysis. In *Proceedings of the 28th international conference on computational linguistics*, pages 568–579.