Mitigating Spurious Correlations via Counterfactual Contrastive Learning

Fengxiang Cheng[1], Chuan Zhou[2,3], Xiang Li[4], Alina Leidinger[1], Haoxuan Li[4,3], Mingming Gong[2,3], Fenrong Liu[*5], Robert van Rooij[*1]

[1]University of Amsterdam, [2]The University of Melbourne, [3]MBZUAI,
[4]Peking University, [5]Tsinghua University
fenrong@tsinghua.edu.cn, R.A.M.vanRooij@uva.nl

Abstract

Identifying causal relationships rather than spurious correlations between words and class labels plays a crucial role in building robust text classifiers. Previous studies proposed using causal effects to distinguish words that are causally related to the sentiment, and then building robust text classifiers using words with high causal effects. However, we find that when a sentence has multiple causally related words simultaneously, the magnitude of causal effects will be significantly reduced, which limits the applicability of previous causal effect-based methods in distinguishing causally related words from spuriously correlated ones. To fill this gap, in this paper, we introduce both the probability of necessity (PN) and probability of sufficiency (PS), aiming to answer the counterfactual question that 'if a sentence has a certain sentiment in the presence/absence of a word, would the sentiment change in the absence/presence of that word?'. Specifically, we first derive the identifiability of PN and PS under different sentiment monotonicities, and calibrate the estimation of PN and PS via the estimated average treatment effect. Finally, the robust text classifier is built by identifying the words with larger PN and PS as causally related words, and other words as spuriously correlated words, based on a contrastive learning approach name **CPNS** is proposed to achieve robust sentiment classification. Extensive experiments are conducted on public datasets to validate the effectiveness of our method.

1 Introduction

Distinguishing between spurious correlations and causal relationships in linguistics is crucial for building robust text classifiers (Sridhar et al., 2018; Roberts et al., 2020; Cheng et al., 2025). For example, in the Movies dataset (Maas et al., 2011) containing IMDB movie reviews, *and* is found to have a stronger correlation with positive sentiment than *excellent* (Paul, 2017). However, from the semantics, it should be *excellent* instead of *and* that causes a positive sentiment of a movie review, and the word *and* itself does not necessarily affect the review's sentiment. This motivates the construction of robust text classifiers by identifying and using words that are causally related to sentiment rather than spurious correlated ones (Olteanu et al., 2017).

Table 1: The average ATE of positive and negative sentiment words as treatments on the Kindle dataset, grouped by the difference in the number of positive and negative sentiment words excluding the treatment word.

Positive sentiment words		Negative sentiment words	
# Pos−Neg	ATE	# Neg−Pos	ATE
0	0.547	0	-0.493
1	0.459	1	-0.498
2	0.289	2	-0.325
3	0.239	3	-0.207

To identify words that are causally related to the sentiment, previous methods propose to consider a specific word as the treatment word and estimate the causal effect on the class labels, whereas sentences containing the specific word are considered as belonging to the treatment group and otherwise to the control group. Causal effect estimation methods include text or propensity matching (De Choudhury et al., 2016; Saha et al., 2019), augmented inverse propensity weighting (AIPW) (Pham and Shen, 2017; Sridhar and Getoor, 2019), and representation learning-based methods (Veitch et al., 2020; Wang et al., 2023, 2024). There are also methods relaxing the common assumptions in complex scenarios (Yang et al., 2024; Li et al., 2024; Wang et al., 2025b; Zheng et al., 2025).

However, a critical issue when using causal effects to identify causally related words is that when multiple causally related words appear in the same

*Fenrong Liu and Robert van Rooij are corresponding authors.

sentence, the causal effect of each causal word on sentiment drops dramatically, making it difficult to identify these words. For example, consider a sentence with positive sentiment – *This movie is excellent and marvelous.* When estimating the causal effect of word *excellent* on sentiment, the matched sentences without the word *excellent* may be – *This movie is [token] and marvelous*, in which *[token]* is a word other than *excellent*, and this sentence may also be recognized as positive sentiment. Therefore, the causal effect of word *excellent* on the sentence sentiment will be small because other positive words (e.g., *marvelous*) also appear in the sentence. This poses a great challenge to the effectiveness of previous methods of identifying causally related words by comparing the causal effects of different words on sentence sentiment.

To empirically reveal the limitations of exploiting average treatment effects (ATEs) to identify causally related words, we compute the average ATE of positive and negative sentiment words as treatments on the Kindle dataset (He and McAuley, 2016). As shown in Table 1, each row shows the average ATE with a specific gap between the total positive sentiment words number and the total negative sentiment words number in the sentence without computing the treatment word. Despite the average ATE for positive sentiment words as treatments being positive in each subgroup, we find that the absolute value of the average ATE decreases significantly as more positive words are contained in the sentence, particularly decreasing from 0.547 to 0.239. Similar conclusions also hold for the cases of negative sentiment words as treatments. Importantly, this observation reveals an inherent limitation of using ATE as a proxy to identify the causally related words, which is irrelevant to ATE estimation methods. Consequently, if the absolute value of the ATE for some causally related words as treatments decreases below a certain threshold, the causally related words may be incorrectly identified as spuriously correlated words, thus decreasing the robustness of the text classifier.

To fill this gap, we aim to answer the counterfactual question, i.e., the highest level in the *causal ladder* (Pearl, 2009), 'if a sentence has a certain sentiment in the presence/absence of a word, would the sentiment change in the absence/presence of that word?', instead of the interventional question as in the previous studies, i.e., the second level in the *causal ladder*. We introduce both the probability of necessity (PN) and probability of sufficiency (PS) (Pearl, 2022) and theoretically derive the identifiability results of PN and PS under different sentiment monotonicities. We further propose a novel robust text classification approach, as shown in Figure 1, in which the signs of the estimated ATEs correspond to different sentiment monotonicities, and words with the lowest estimated PN and PS are considered as spuriously correlated words and thus removed to achieve robust text classification. Extensive experiments are conducted on three public datasets, demonstrating the superiority of our proposal on both spurious correlated word identification and robust text classification. The contributions can be summarized as follows.

- We are the first work to point out the inadequate of ATE or CATE compared with PN and PS for identifying causally related words and spuriously correlated words.
- We design a contrastive-learning probability of necessity and sufficiency (CPNS) to estimate PN and PS in the sentence classification task and achieve more accurate sentence classification via better word identification.
- We conduct extensive experiments on 3 public datasets and 3 backbones, under both cross-domain and in-domain settings to validate the effectiveness of our method.

2 Preliminaries

2.1 Robust Text Classification

In this paper, we consider the task of binary text classification on the dataset $\mathcal{D} = \{(s_1, y_1), ..., (s_n, y_n)\}$. We ignore subscripts for simplicity without ambiguity. For each sentence s consisting of k words, its sentiment label is binary, i.e., $y \in \{0, 1\}$, where 0 denotes negative sentiment and 1 denotes positive sentiment. By exploiting a feature encoder $g : s \mapsto x$, we first transform a sentence s into a dense feature vector x. Finally, we aim to train a binary classifier $f_\theta : x \mapsto \{0, 1\}$ parameterized by θ by minimizing a pre-defined training loss $L(\mathcal{D}; \theta)$, which predicts the sentiment label with each feature vector x.

To enhance the robustness and transferability of the classifier, we consider the more fine-grained word-level relationships to the sentiment label, aiming to distinguish the causally related words from the spuriously correlated words. For instance, the word *and* is spuriously correlated with the positive sentiment label in the IMDB movie reviews,

Table 2: The sentences can be divided into eight strata, with the unobserved values highlighted in red. For each stratum, counterfactual necessity and sufficiency either hold (✓), do not hold (×), or unknown (?).

T	Y	$Y(0)$	$Y(1)$	Necessity	Sufficiency
0	0	0	0	?	×
0	0	0	1	?	✓
0	1	1	0	?	✓
0	1	1	1	?	×
1	0	0	0	×	?
1	0	1	0	✓	?
1	1	0	1	✓	?
1	1	1	1	×	?

but not in the Kindle book reviews. On the contrary, the causally related words have robust relationships with the class label across different domains, upon which we can build a more robust text classifier. Let $\mathcal{W} = \{w_1, w_2, \ldots, w_A\}$ be all the words in the training data, we seek to find the words $\mathcal{W}^{sp} = \{w_1^{sp}, w_2^{sp}, \ldots, w_B^{sp}\} \subseteq \mathcal{W}$ most likely to be spuriously correlated to the sentiment label and $\mathcal{W}^{c} = \{w_1^{c}, w_2^{c}, \ldots, w_C^{c}\} \subseteq \mathcal{W}$ most likely to be causally related to the sentiment label. To achieve robust sentence classification, we remove $\mathcal{W}^{sp}$ and/or $\mathcal{W}^{c}$ from the sentences to train the classifier, formally $f(g(s \setminus \mathcal{W}^{sp}), g(s \setminus \mathcal{W}^{c}), g(s); \theta)$.

2.2 Causal Formulation

We formulate the causally related words identification problem using the Neyman-Rubin causal framework (Imbens and Rubin, 2015). Given a specific word w, the treatment is set to $T = 1$ if w appears in the sentence, otherwise $T = 0$ if w does not appear. Let the sentence removing w be the covariate X, i.e., $X = s \setminus \{w\} \in \mathcal{X}$. Using the Neyman-Rubin causal framework, in addition to the observed sentiment label Y, we denote $Y(0)$ and $Y(1)$ as the potential outcomes when receiving treatment $T = 0$ and $T = 1$, respectively.

Note that for each sentence one can only observe one sentiment label $Y = (1-T)Y(0)+TY(1)$, but not both $Y(0)$ and $Y(1)$, which is also known as the fundamental problem of causal inference (Holland, 1986; Morgan, 2015). We also assume the unconfoundedness that $(Y(0), Y(1)) \perp\!\!\!\perp T \mid X$ and let $0 < \mathbb{P}(T = 1|X = x) < 1$ for all $x \in \mathcal{X}$. That is, given the sentence removing the treatment word, the presence or non-presence of the word w is independent of the potential outcomes, and the probabilities of presence and non-presence of the treatment word are both positive.

The most common estimands for measuring the impact of one specific treatment word on the sentiment label are causal effects. Specifically, the conditional average treatment effect (CATE) with given covariate X is defined as $\mathbb{E}(Y(1) - Y(0) \mid X)$, and the average treatment effect (ATE) is defined as $\mathbb{E}(Y(1) - Y(0))$, which is the average of CATEs over all possible covariate X. Previous works use the causal effects as auxillary metrics to distinguish the causally related words from spuriously related words (Falavarjani et al., 2017; Wood-Doughty et al., 2018; Pryzant et al., 2021)– when a word has a relatively large causal effect on the class label, it is predicted as a causally related word. Oppositely, a word strongly correlated with the class label but not causally related is regarded as a spuriously correlated word.

3 Proposed Method

3.1 PN and PS

When there are more than one positive or negative sentiment words in one sentence, the magnitude of both CATE and ATE will be significantly reduced, which challenges the causally related words identification. In this paper, instead of using population-level causal effect estimation (Wang et al., 2025a; Zhang et al., 2025; Zhou et al., 2025a), we need to identify the causally related words, a counterfactual question on the individual-level (Wu et al., 2025b,a; Zhou et al., 2025b; Li et al., 2025). Specifically, we first theoretically derive the identification results of probability of necessity (PN) and the probability of sufficiency (PS) under different sentiment monotonicities, and further propose a robust text classification algorithm by accurately estimating the PN and PS and removing a certain percentage of words with the lowest estimated PN and PS.

Definition 3.1 (Probability of Necessity (Pearl, 2022))**.** *The probability of necessity is the probability that sentiment $Y = y$ would not occur in the absence of word (denoted as $T = 0$), in the case where the word and sentiment $Y = y$ did occur, i.e.,* $\mathbb{P}(Y(0) = 1 - y \mid T = 1, Y = y, X)$.

Definition 3.2 (Probability of Sufficiency (Pearl, 2022))**.** *The probability of sufficiency is the probability of the capacity of a word to produce sentiment $Y = 1 - y$, in the case where the word is absent (denoted as $T = 0$) with sentiment $Y = y$, i.e.,* $\mathbb{P}(Y(1) = 1 - y \mid T = 0, Y = y, X)$.

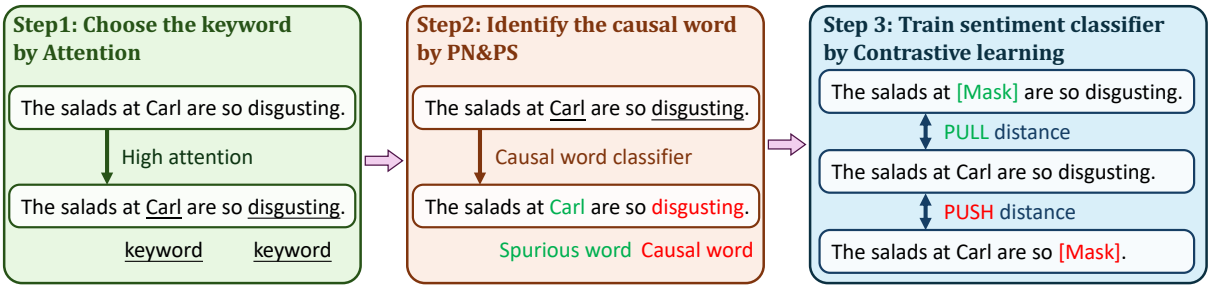


Figure 1: A three-step process consists of selecting keywords, identifying the causal words, and reweighting the keywords in the training of sentence sentiment classifier.

Algorithm 1: Robust text classification using words with high probability of necessity and sufficiency

Input: hyperparameters $\alpha, \beta, k > 0$, training data $\mathcal{D} = \{(s_1, y_1), \ldots, (s_n, y_n)\}$;

Train an initial classifier $f(x; \theta)$ on training data $\mathcal{D}$;

Extract from $f(x; \theta)$ the words $\{w_1, \ldots, w_M\}$ that are most strongly associated with each class according to the initial classifier;

for $m \in \{1, \ldots, M\}$ **do**

  Estimate $\hat{\mathbb{P}}(Y \mid T = 0, X)$ and $\hat{\mathbb{P}}(Y \mid T = 1, X)$;

  Estimate average treatment effect $\hat{\tau}_m$ of word w_m;

  if $\hat{\tau}_m \geq 0$ **then**

    $\mathrm{PN}_m \leftarrow 1 + \frac{1}{n_{\text{pos}}} \sum_{i:y_i=1} \frac{\hat{\mathbb{P}}(Y=0|T=0,X)-1}{\hat{\mathbb{P}}(Y=1|T=1,X)}$;

    $\mathrm{PS}_m \leftarrow 1 + \frac{1}{n_{\text{neg}}} \sum_{i:y_i=0} \frac{\hat{\mathbb{P}}(Y=1|T=1,X)-1}{\hat{\mathbb{P}}(Y=0|T=0,X)}$;

  else

    $\mathrm{PN}_m \leftarrow 1 + \frac{1}{n_{\text{neg}}} \sum_{i:y_i=0} \frac{\hat{\mathbb{P}}(Y=1|T=0,X)-1}{\hat{\mathbb{P}}(Y=0|T=1,X)}$;

    $\mathrm{PS}_m \leftarrow 1 + \frac{1}{n_{\text{pos}}} \sum_{i:y_i=1} \frac{\hat{\mathbb{P}}(Y=0|T=1,X)-1}{\hat{\mathbb{P}}(Y=1|T=0,X)}$;

Rank the words ascendingly by $\alpha\mathrm{PN} + \beta\mathrm{PS}$;

Classify the words ranked at final $K\%$ as $\mathcal{W}^c$, and others as $\mathcal{W}^{sp}$;

Train a robust f using the loss as Eq (3.3);

Output: robust transferable text classifier $f(x; \theta)$.

3.2 Identification and Estimation

Based on the definition of PN and PS, we can analyze the necessity and sufficiency of the treatment word for the sentiment of the sentence, as Table 2 shows. Since PN and PS are at the counterfactual level, we require one more assumption than standard causal inference for treatment effects.

Assumption 3.1 (Monotonicity)**.** *For each word as treatment, either the word is positively monotonic to the class label* $Y(1) \geq Y(0)$ *or negatively monotonic* $Y(1) \leq Y(0)$.

We argue that this assumption is not strong since it only requires the sentiment of a word would be either positive or negative across different sentence contexts, but can with varying causal effect values. For example, the causal effect of the word *excellent* to the positive sentiment may change according to different sentence contexts, but barely be negative. Next, we derive the identifiability of PN and PS under different sentiment monotonicities as follows.

Theorem 3.1 (Identifiability Under Monotonicity)**.** *Under Assumption 3.1 that* $Y(1) \geq Y(0)$*, the PN and PS are identifiable:*

$$\begin{aligned}
&\mathbb{P}(Y(0) = 0 \mid T = 1, Y = 1, X) \\
=&1 + \frac{\mathbb{P}(Y = 0 \mid T = 0, X) - 1}{\mathbb{P}(Y = 1 \mid T = 1, X)}, \\
&\mathbb{P}(Y(1) = 1 \mid T = 0, Y = 0, X) \\
=&1 + \frac{\mathbb{P}(Y = 1 \mid T = 1, X) - 1}{\mathbb{P}(Y = 0 \mid T = 0, X)}.
\end{aligned}$$

Under Assumption 3.1 that $Y(1) \leq Y(0)$*, the PN*

and PS are identifiable:

$$\begin{aligned}&\mathbb{P}(Y(0)=1 \mid T=1, Y=0, X)\\ &=1+\frac{\mathbb{P}(Y=1 \mid T=0, X)-1}{\mathbb{P}(Y=0 \mid T=1, X)},\\ &\mathbb{P}(Y(1)=0 \mid T=0, Y=1, X)\\ &=1+\frac{\mathbb{P}(Y=0 \mid T=1, X)-1}{\mathbb{P}(Y=1 \mid T=0, X)}.\end{aligned}$$

Proof. Without loss of generality, we only prove the identification of $\mathbb{P}(Y(0)=0 \mid T=1, Y=1, X)$ under the sentiment monotonicity $Y(1) \geq Y(0)$ in below:

$$\begin{aligned}&\mathbb{P}(Y(0)=0 \mid T=1, Y=1, X)\\ &=\frac{\mathbb{P}(Y(0)=0, Y=1 \mid T=1, X)}{\mathbb{P}(Y=1 \mid T=1, X)}\\ &=\frac{\mathbb{P}(Y(0)=0, Y(1)=1 \mid T=1, X)}{\mathbb{P}(Y=1 \mid T=1, X)}\\ &=\frac{\mathbb{P}(Y(0)=0, Y(1)=1 \mid X)}{\mathbb{P}(Y=1 \mid T=1, X)},\end{aligned} \tag{1}$$

where the first equality holds directly from the definition of conditional probability, the second equality is from the consistency assumption, and the third equality is from the strong ignorability assumption.

For the $\mathbb{P}(Y(0)=0, Y(1)=1 \mid X)$ term in the numerator, we have the following results:

$$\begin{aligned}&\mathbb{P}(Y(0)=0, Y(1)=1 \mid X)\\ &=\big(\mathbb{P}(Y(0)=0, Y(1)=1 \mid X)+\mathbb{P}(Y(0)=1, Y(1)=1 \mid X)\big)\\ &+\big(\mathbb{P}(Y(0)=0, Y(1)=0 \mid X)+\mathbb{P}(Y(0)=0, Y(1)=1 \mid X)\big)\\ &-\big(\mathbb{P}(Y(0)=0, Y(1)=0 \mid X)+\mathbb{P}(Y(0)=0, Y(1)=1 \mid X)\\ &+\underbrace{\mathbb{P}(Y(0)=1, Y(1)=0 \mid X)}_{\text{equals to 0 because } Y(1)\geq Y(0)}+\mathbb{P}(Y(0)=1, Y(1)=1 \mid X)\big)\\ &=\mathbb{P}(Y(1)=1 \mid X)+\mathbb{P}(Y(0)=0 \mid X)-1\\ &=\mathbb{P}(Y=1 \mid T=1, X)+\mathbb{P}(Y=0 \mid T=0, X)-1.\end{aligned} \tag{2}$$

Combining Eq. (1) and Eq. (2) identifies PN as:

$$\begin{aligned}&\mathbb{P}(Y(0)=0 \mid T=1, Y=1, X)\\ &=1+\frac{\mathbb{P}(Y=0 \mid T=0, X)-1}{\mathbb{P}(Y=1 \mid T=1, X)}.\end{aligned} \tag{3}$$

The rest of the identifiability results can be obtained by following a similar argument. □

The theoretical results above not only provide the proof of identifiability for PN and PS, but also explicitly show plausible estimators with observed variables. For example, the RHS of Eq. (3) is only about the distribution of observed data (Y, T, X).

Table 3: Summary statistics of datasets.

Dataset	Food	IMDB	SST-2
Samples	17,273	35,000	67,349
Positive samples	13,618	17,540	37,569
Negative samples	3,656	17,461	29,781

3.3 Robust Sentiment Classifier Training

We exploit the identification results and propose an algorithm for robust sentiment classifier training as shown in Algorithm 1. From Theorem 3.1, we note that the identification results under $Y(1) \leq Y(0)$ (negative sentiment words) and $Y(1) \geq Y(0)$ (positive sentiment words) are different. This motivates us to first determine whether $Y(1) \geq Y(0)$ or $Y(1) \leq Y(0)$, which is obtained by the sign of the estimated ATE $\hat{\tau}_m$ (line 6), then estimate the PN and PS for each treatment word. To reduce computational cost, with the training data $\mathcal{D}$, we first train an initial classifier $f(x;\theta)$ to find the candidate words $\{w_1, \ldots, w_M\}$ which are mostly correlated with the class label (lines 1 to 2). Then we take each candidate word $w_m, m \in \{1, 2, \ldots, M\}$ as the treatment word and estimate its PN and PS (lines 3 to 11). Then we compute the aggregation $\alpha\text{PN} + \beta\text{PS}$ for each candidate word as $Agg_1, Agg_2, \ldots, Agg_M$. Denote the upper $k\%$ quantile of these aggregations as $Agg(k\%)$, then causally related words are identified as:

$$\mathcal{W}^c = \{w_m : m \in \mathbb{N}, 1 \leq m \leq M, Agg_m \geq Agg(k\%)\}.$$

And the spuriously correlated words are oppositely identified as the complement set:

$$\mathcal{W}^{sp} = \{w_m : m \in \mathbb{N}, 1 \leq m \leq M, Agg_m < Agg(k\%)\}.$$

Let $\mathcal{L}_{ce}$ be the cross-entropy loss for the sentence classification task, and $\mathcal{L}_{con}$ be the contrastive loss ensuring that the sentence feature encoding aligns with causally related and spuriously correlated word identification:

$$\mathcal{L}_{con} = sim(g(s), g(s \setminus \mathcal{W}^{sp})) - sim(g(s), g(s \setminus \mathcal{W}^{c})),$$

where $sim(\cdot, \cdot)$ means the cosine similarity. To obtain a robust sentence classifier $f(g(\cdot))$, we finally use the following training loss:

$$\mathcal{L} = \mathcal{L}_{ce} + \lambda \mathcal{L}_{con},$$

where $\lambda > 0$ is a hyperparameter.

Notice that the proposed algorithm does not require accurate estimations of PN, PS, or ATE. For

Table 4: Model Performance in **cross-domain** scenario. For example, Food → IMDB means training sentiment classification model in the Food dataset and evaluating such model in the IMDB dataset.

Backbone	Method	Accuracy					
		Food → IMDB	**Food → SST-2**	**IMDB → Food**	**IMDB → SST-2**	**SST-2 → Food**	**SST-2 → IMDB**
BERT	Vanilla	76.8992 ± 0.1397	74.7920 ± 0.5986	84.0926 ± 0.0989	83.8881 ± 0.3337	81.3165 ± 0.2504	83.0483 ± 0.1395
	IPS	75.0492 ± 0.1646	75.2224 ± 0.5952	84.5926 ± 0.1228	83.5868 ± 0.4895	79.6684 ± 0.2207	82.0075 ± 0.1421
	Matching	76.9100 ± 0.1153	76.3974 ± 0.6026	86.2862 ± 0.1063	85.6815 ± 0.4246	82.7475 ± 0.2751	83.9517 ± 0.1098
	DR	77.5504 ± 0.1311	76.0091 ± 0.5712	86.1915 ± 0.1325	85.4389 ± 0.4277	83.8937 ± 0.2582	82.1909 ± 0.1259
	TarNet	77.8892 ± 0.1298	77.2310 ± 0.5669	86.3300 ± 0.1678	85.7963 ± 0.4352	82.4399 ± 0.2635	83.4027 ± 0.1326
	CPNS	**78.3375 ± 0.1251**	**77.7762 ± 0.5424**	**86.4512 ± 0.1293**	**86.7145 ± 0.3221**	**84.3810 ± 0.2598**	**84.8827 ± 0.1242**
RoBERTa	Vanilla	85.1983 ± 0.1499	80.8895 ± 0.4482	88.8889 ± 0.0934	84.3469 ± 0.2468	85.0392 ± 0.2680	86.5325 ± 0.1875
	IPS	85.4346 ± 0.1327	81.3142 ± 0.4886	90.3285 ± 0.1082	85.4532 ± 0.6639	86.0248 ± 0.1937	85.8912 ± 0.1729
	Matching	85.2917 ± 0.1234	81.0364 ± 0.5674	**90.7609 ± 0.1314**	85.7532 ± 0.6639	85.5943 ± 0.2395	86.0975 ± 0.1754
	DR	85.3427 ± 0.1416	81.3728 ± 0.5126	90.8327 ± 0.0927	85.3785 ± 0.2451	86.2873 ± 0.1812	86.4731 ± 0.1847
	TarNet	85.5827 ± 0.1658	**81.8789 ± 0.5966**	89.2054 ± 0.1038	86.2841 ± 0.5746	86.1724 ± 0.1863	**87.2145 ± 0.1639**
	CPNS	**85.6083 ± 0.1139**	81.6786 ± 0.5013	90.7088 ± 0.0830	**86.8732 ± 0.6080**	**89.9024 ± 0.2067**	86.6217 ± 0.1671
ALBERT	Vanilla	81.4591 ± 0.4322	81.4431 ± 0.6104	85.2347 ± 0.0877	85.1937 ± 0.5085	83.1485 ± 0.2108	84.6483 ± 0.1391
	IPS	80.6608 ± 0.1454	81.5638 ± 0.7462	84.8704 ± 0.2109	87.2066 ± 0.3064	82.6841 ± 0.2253	84.3176 ± 0.1417
	Matching	82.3450 ± 0.1952	81.8508 ± 0.5676	81.0421 ± 0.2459	86.5423 ± 0.5586	83.2167 ± 0.2021	84.4798 ± 0.1473
	DR	80.9235 ± 0.1762	80.4723 ± 0.7311	86.4012 ± 0.1938	86.8742 ± 0.4661	83.4182 ± 0.1895	84.0061 ± 0.1406
	TarNet	83.0824 ± 0.1387	82.3452 ± 0.7705	85.6414 ± 0.2317	86.8192 ± 0.4821	83.4821 ± 0.2014	**85.2145 ± 0.1536**
	CPNS	**84.5300 ± 0.1219**	**82.6858 ± 0.6798**	**87.7744 ± 0.2413**	**87.8680 ± 0.4071**	**83.9764 ± 0.1612**	85.1917 ± 0.1172

PN (PS), we only need to make sure the upper $k\%$ words have larger aggregated PN and PS estimates than the lower $1-k\%$ words. While for ATE, the only requirement is that the sign of $\hat{\tau}_m$ is correct. Not relying on the accurate ATE estimation further enhances the robustness of our algorithm in addition to the advantages of the metrics PN and PS themselves over the widely adopted causal effects.

4 Experiments

In this section, we conduct extensive experiments on our proposed method, aiming to answer the following research questions (RQs):

- **RQ1**: Can CPNS effectively eliminate spurious correlations and perform better in **cross-domain** settings?
- **RQ2**: Can our method maintain its advantage in **in-domain** settings?
- **RQ3**: Does CPNS outperforms traditional causal effect estimation methods in **identifying causal words**?
- **RQ4**: How **sensitive** is the model's performance to changes in its **hyperparameters**?

4.1 Experimental Setup

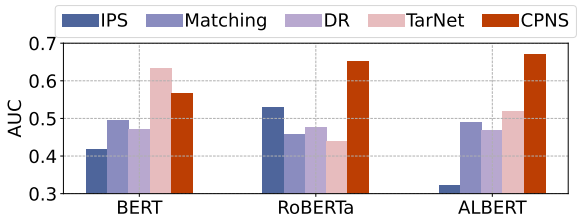
Datasets. We conduct the sentiment analysis experiments on three widely-used datasets: FineFood (Food) (McAuley and Leskovec, 2013), IMDB movie reviews (IMDB) (Maas et al., 2011), and Stanford Sentiment Treebank (SST-2) (Socher et al., 2013). The summary statistics are shown in Table 3, where positive sample means the sentence with positive sentiment, and negative sample means the sentence with negative sentiment.

Backbone Models. We use standard PLMs including BERT (Devlin, 2018), RoBERTa (Liu, 2019), and ALBERT (Lan, 2019) as the backbone models for obtaining the embeddings of sentences and top words. In addition, for our method, we use the two-layer MLP as the backbone model for learning a balanced representation.

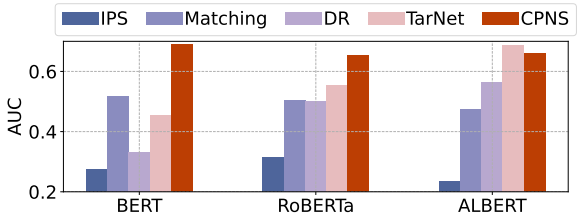
Baselines. We compare several causal effect estimation methods to achieve robust sentiment classification. Specifically, **IPS** (Saha et al., 2019) estimates the causal effects using the conditional probability of receiving a treatment given confounders (i.e., sentences), named propensity scores. The inverse of the propensities are used to reweight the observed samples, enabling unbiased causal effect estimation under accurate propensity scores. **DR** (Sridhar and Getoor, 2019) models both the treatment assignment and the outcome. It has the desirable property that the effect estimate remains unbiased as long as either the propensity or outcome model is unbiased. **Matching** (Wang and Culotta, 2020) aims to match individuals with similar features but alternative treatment to impute their counterfactual outcomes. For example, it groups sentences based on representations and then estimates the causal effect within each group. **Tar-**

Table 5: The accuracy performance of the model under the **in-domain** scenarios.

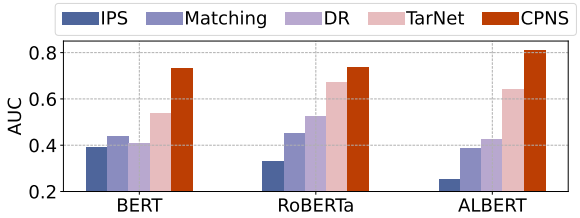
Backbone	Method	Accuracy		
		Food	IMDB	SST-2
BERT	Vanilla	95.1397 ± 0.1007	89.3167 ± 0.1062	90.5638 ± 0.3618
	IPS	95.0212 ± 0.0964	88.9050 ± 0.1031	91.1908 ± 0.4217
	Matching	95.9865 ± 0.1186	89.1433 ± 0.0751	92.0803 ± 0.3443
	DR	95.9529 ± 0.0822	89.2021 ± 0.0712	92.7632 ± 0.3684
	TarNet	95.3283 ± 0.0777	89.4058 ± 0.0688	92.1024 ± 0.3729
	CPNS	**96.2037** ± 0.0968	**89.6367** ± 0.0749	**93.6744** ± 0.3874
RoBERTa	Vanilla	96.1145 ± 0.1399	89.8967 ± 0.0696	93.4911 ± 0.2485
	IPS	95.8731 ± 0.0925	89.4278 ± 0.0687	93.2385 ± 0.3614
	Matching	96.3636 ± 0.0731	90.7992 ± 0.0762	93.4290 ± 0.3839
	DR	96.7013 ± 0.0852	90.5126 ± 0.0663	93.4290 ± 0.3839
	TarNet	97.0589 ± 0.0635	90.0200 ± 0.0751	93.7654 ± 0.3341
	CPNS	**97.0943** ± 0.0872	**91.6723** ± 0.0634	**94.0890** ± 0.3195
ALBERT	Vanilla	95.7710 ± 0.0900	88.9367 ± 0.0552	89.9208 ± 0.3422
	IPS	95.5783 ± 0.1088	89.0858 ± 0.1275	90.1685 ± 0.3821
	Matching	96.0943 ± 0.0768	88.9017 ± 0.0852	90.7076 ± 0.3684
	DR	96.0427 ± 0.0821	89.0347 ± 0.0733	90.8653 ± 0.3957
	TarNet	95.8421 ± 0.0762	89.0925 ± 0.0815	90.9615 ± 0.3627
	CPNS	**96.1094** ± 0.0720	**89.4275** ± 0.0727	**91.1047** ± 0.3445



(a) The performance on the Food dataset.



(b) The performance on the IMDB dataset.



(c) The performance on the SST-2 dataset.

Figure 2: Comparison between causal effect-based methods and the CPNS for causal word identification.

Net (Shalit et al., 2017) learns a balanced representation to estimate causal effects. We also include a **Vanilla** approach, which directly uses the backbone model without removing spurious correlations.

Implement Details. We utilize a setup of 8 NVIDIA 3090 GPUs, supported by 300GB random access memory (RAM). It takes approximately 5 hours to train for 10 epochs on the Food dataset.

4.2 Cross-Domain Performance (RQ1)

We conduct experiments across three benchmark datasets using three widely adopted backbone models. We compare our method, CPNS, with several representative causal inference baselines. In particular, we focus on the cross-domain sentiment classification setting, where the model is trained on one dataset (source domain) and evaluated on another (target domain), a scenario that is more challenging due to domain shift and spurious correlations. We report classification accuracy on the target domain as the primary evaluation metric.

As shown in Table 4, all causal baselines generally outperform the vanilla backbone model, validating the necessity of identifying causally related words and removing spuriously correlated ones in sentiment analysis. Among these baselines, IPS and Matching are consistently outperformed by more advanced methods such as DR and TarNet.

Our proposed method CPNS achieves the best or second-best accuracy in all scenarios, consistently outperforming existing approaches. This improvement stems from a key limitation in traditional causal effect estimation: when multiple sentiment-related words appear in a sentence, the magnitudes of average treatment effect (ATE) and conditional average treatment effect (CATE) are often diluted, which compromises the identification of truly causal words.

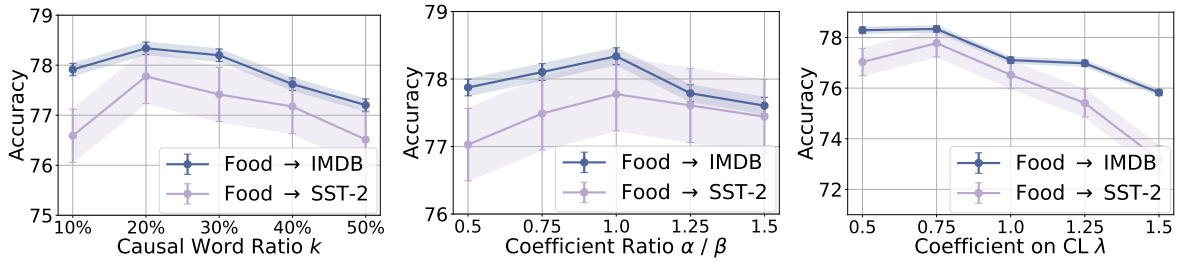


Figure 3: The variation in the model's performance under different parameters.

4.3 In-Domain Performance (RQ2)

To evaluate the in-domain effectiveness of CPNS, we train and test the model on the same dataset across three domains: Food, IMDB, and SST-2. As shown in Table 4, CPNS consistently achieves the highest or competitive accuracy across all backbones and datasets.

The superior in-domain performance of CPNS can be attributed to its better identification of causal words and its ability to suppress spurious correlations without losing semantically meaningful information. Unlike methods that rely purely on global statistical adjustments, CPNS accurately identifies causally relevant words using PN and PS, ensuring that essential sentiment cues are preserved.

In addition, our adoptation of contrastive learning further strengthens this capability: the model is trained to down-weight the influence of spuriously correlated words while simultaneously enhancing the representations of truly causally related words. This dual mechanism allows CPNS to eliminate misleading associations while maintaining or even reinforcing genuine sentiment signals, which is crucial even when there is no domain shift.

4.4 Causal Word Identification (RQ3)

As shown in Figure 2, CPNS consistently achieves the highest AUC scores across all datasets and backbone architectures, clearly outperforming all baselines. In this work, the labels for causal words were generated by an LLM (GPT-o1)[1] and verified manually. For example, on the SST-2 dataset, CPNS achieves an AUC above 0.8 when using ALBERT as the backbone, while other methods fall significantly behind. This demonstrates the superiority of CPNS in distinguishing causally related words from spuriously correlated ones.

These results validate the effectiveness of our CPNS framework for causal word identification and highlight its robustness across different datasets and backbones, making it a more reliable foundation for real-world sentiment analysis tasks.

[1] https://openai.com/

4.5 Parameter Sensitivity Analysis (RQ4)

To assess the parameter sensitivity of CPNS, we conduct experiments using five values for each of three key hyperparameters, as shown in Figure 3.

Causal Word Ratio k. Keywords are ranked based on a linear combination of their calculated PN and PS scores, with the top $k\%$ treated as causal. The results show that performance peaks at $k = 20\%$, suggesting that a moderate proportion of high-confidence causal words strikes a good balance between retaining meaningful information and avoiding spurious correlations.

Coefficient Ratio α/β. The best accuracy is obtained when PN and PS are equally weighted ($\alpha/\beta = 1.0$). Performance drops slightly as the ratio deviates from 1.0, indicating that both PN and PS contribute equally to causal word identification.

Contrastive Loss Coefficient λ. Performance is maximized at $\lambda = 0.75$. Smaller values weaken the effect of contrastive learning, while larger values cause the model to overemphasize contrastive loss, harming classification accuracy.

5 Conclusion

This paper proposes a novel method for distinguishing causally related and spuriously correlated words in sentiment classification by leveraging the probability of necessity (PN) and the probability of sufficiency (PS), aiming to eliminate linguistic spurious correlations. Theoretically, we derive the identifiability of PN and PS under different sentiment monotonicity assumptions. Empirically, we conduct extensive experiments across multiple datasets and backbone models, covering both cross-domain and in-domain scenarios, to demonstrate the effectiveness of our method in causal word identification and sentence sentiment classification.

Limitations

One possible limitation of this paper is that the monotonicity assumption may be violated for a few sentences with the presence of negation words. Specifically, we assume that adding positive (negative) sentiment words will monotonically increase (decrease) the probability of getting a positive label. While convenient for identification, this assumption can be violated by natural language constructs. For example, negation words such as *not*, *never*, and *hardly* will invert or attenuate sentiment and break monotonicity. When such non-monotonic examples arise, the exact identifiability of PN and PS breaks down, yielding only bounded or partial estimates. Addressing this issue may require relaxing the monotonicity assumption, for instance via partial monotonic models or by deriving PN/PS under weaker conditions (e.g., bounded identification).

Acknowledgments

HL was supported by the National Natural Science Foundation of China (623B2002). MG was supported by ARC DP240102088 and WIS-MBZUAI 142571. FL was supported by the Beijing Natural Science Foundation (No. L257007) and the Tsinghua University Initiative Scientific Research Program. RvR was supported by the Dutch Research Council (NWO) (406.18.TW.007).

References

Fengxiang Cheng, Haoxuan Li, Fenrong Liu, Robert Van Rooij, Kun Zhang, and Zhouchen Lin. 2025. Empowering llms with logical reasoning: A comprehensive survey. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*.

Munmun De Choudhury, Emre Kiciman, Mark Dredze, Glen Coppersmith, and Mrinal Kumar. 2016. Discovering shifts to suicidal ideation from mental health content in social media. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 2098–2110.

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Seyed Mirlohi Falavarjani, Hawre Hosseini, Zeinab Noorian, and Ebrahim Bagheri. 2017. Estimating the effect of exercising on users' online behavior. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11, pages 734–738.

Ruining He and Julian McAuley. 2016. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *Proceedings of the 25th International Conference on World Wide Web*, pages 507–517.

Paul W Holland. 1986. Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960.

Guido W Imbens and Donald B Rubin. 2015. *Causal inference in statistics, social, and biomedical sciences*. Cambridge university press.

Zhenzhong Lan. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Haoxuan Li, Zeyu Tang, Zhichao Jiang, Zhuangyan Fang, Yue Liu, Zhi Geng, and Kun Zhang. 2025. Fairness on principal stratum: A new perspective on counterfactual fairness. In *International Conference on Machine Learning*.

Haoxuan Li, Chunyuan Zheng, Sihao Ding, Peng Wu, Zhi Geng, Fuli Feng, and Xiangnan He. 2024. Be aware of the neighborhood effect: Modeling selection bias under interference. In *International Conference on Learning Representations*.

Yinhan Liu. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 364.

Andrew Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150.

Julian John McAuley and Jure Leskovec. 2013. From amateurs to connoisseurs: modeling the evolution of user expertise through online reviews. In *Proceedings of the 22nd International Conference on World Wide Web*, pages 897–908.

SL Morgan. 2015. *Counterfactuals and causal inference*. Cambridge University Press.

Alexandra Olteanu, Onur Varol, and Emre Kiciman. 2017. Distilling the outcomes of personal experiences: A propensity-scored analysis of social media. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pages 370–386.

Michael Paul. 2017. Feature selection as causal inference: Experiments with text classification. In *Proceedings of the 21st Conference on Computational Natural Language Learning*, pages 163–172.

Judea Pearl. 2009. *Causality*. Cambridge university press.

Judea Pearl. 2022. Probabilities of causation: three counterfactual interpretations and their identification. In *Probabilistic and Causal Inference: The Works of Judea Pearl*, pages 317–372.

Thai T Pham and Yuanyuan Shen. 2017. A deep causal inference approach to measuring the effects of forming group loans in online non-profit microfinance platform. *arXiv preprint arXiv:1706.02795*.

Reid Pryzant, Dallas Card, Dan Jurafsky, Victor Veitch, and Dhanya Sridhar. 2021. Causal effects of linguistic properties. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4095–4109.

Margaret E Roberts, Brandon M Stewart, and Richard A Nielsen. 2020. Adjusting for confounding with text matching. *American Journal of Political Science*, 64(4):887–903.

Koustuv Saha, Benjamin Sugar, John Torous, Bruno Abrahao, Emre Kıcıman, and Munmun De Choudhury. 2019. A social media study on the effects of psychiatric medication use. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, pages 440–451.

Uri Shalit, Fredrik D Johansson, and David Sontag. 2017. Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning*, pages 3076–3085. PMLR.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642.

Dhanya Sridhar and Lise Getoor. 2019. Estimating causal effects of tone in online debates. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 1872–1878.

Dhanya Sridhar, Aaron Springer, Victoria Hollis, Steve Whittaker, and Lise Getoor. 2018. Estimating causal effects of exercise from mood logging data. In *IJCAI/ICML Workshop on CausalML*.

Victor Veitch, Dhanya Sridhar, and David Blei. 2020. Adapting text embeddings for causal inference. In *Conference on Uncertainty in Artificial Intelligence*, pages 919–928. PMLR.

Fan Wang, Chaochao Chen, Weiming Liu, Tianhao Fan, Xinting Liao, Yanchao Tan, Lianyong Qi, and Xiaolin Zheng. 2024. CE-RCFR: Robust counterfactual regression for consensus-enabled treatment effect estimation. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3013–3023.

Hao Wang, Zhichao Chen, Zhaoran Liu, Xu Chen, Haoxuan Li, and Zhouchen Lin. 2025a. Proximity matters: Local proximity enhanced balancing for treatment effect estimation. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, pages 2927–2937.

Hao Wang, Jiajun Fan, Zhichao Chen, Haoxuan Li, Weiming Liu, Tianqiao Liu, Quanyu Dai, Yichao Wang, Zhenhua Dong, and Ruiming Tang. 2023. Optimal transport for treatment effect estimation. *Advances in Neural Information Processing Systems*.

Haotian Wang, Haoxuan Li, Hao Zou, Haoang Chi, Long Lan, Wanrong Huang, and Wenjing Yang. 2025b. Effective and efficient time-varying counterfactual prediction with state-space models. In *International Conference on Learning Representations*.

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*, volume 2018, page 4586.

Anpeng Wu, Haoxuan Li, Chunyuan Zheng, Kun Kuang, and Kun Zhang. 2025a. Classifying treatment responders: Bounds and algorithms. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1*, pages 1611–1622.

Peng Wu, Haoxuan Li, Chunyuan Zheng, Yan Zeng, Jiawei Chen, Yang Liu, Ruocheng Guo, and Kun Zhang. 2025b. Learning counterfactual outcomes under rank preservation. *Advances in Neural Information Processing Systems*.

Wenjing Yang, Haotian Wang, Haoxuan Li, Hao Zou, Ruochun Jin, Kun Kuang, and Peng Cui. 2024. Your neighbor matters: Towards fair decisions under networked interference. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3829–3840.

Zhiheng Zhang, Haoxiang Wang, Haoxuan Li, and Zhouchen Lin. 2025. Active treatment effect estimation via limited samples. In *International Conference on Machine Learning*.

Chunyuan Zheng, Haocheng Yang, Haoxuan Li, and Mengyue Yang. 2025. Unveiling extraneous sampling bias with data missing-not-at-random. *Advances in Neural Information Processing Systems*.

Chuan Zhou, Yaxuan Li, Chunyuan Zheng, Haiteng Zhang, Min Zhang, Haoxuan Li, and Mingming Gong. 2025a. A two-stage pretraining-finetuning framework for treatment effect estimation with unmeasured confounding. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1*, page 2113–2123.

Chuan Zhou, Lina Yao, Haoxuan Li, and Mingming Gong. 2025b. Counterfactual implicit feedback modeling. *Advances in Neural Information Processing Systems*.