Revealing the Limitations of Exploiting Causal Effects to Resolve Linguistic Spurious Correlations

Fengxiang Cheng[1], **Haoxuan Li**[2], **Alina Leidinger**[1], **Robert van Rooij**[1]

[1] Institute for Logic, Language and Computation, University of Amsterdam,
[2] Center for Data Science, Peking University
f.cheng@uva.nl, hxli@stu.pku.edu.cn, a.j.leidinger@uva.nl, 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 effect 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 spurious 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 removing a certain percentage of words with the lowest estimated PN and PS. Extensive experiments are conducted on public datasets to validate the effectiveness of our method.

Introduction

Distinguishing between spurious correlations and causal relationships in linguistics is crucial for building robust text classifiers (Sridhar et al. 2018; Roberts, Stewart, and Nielsen 2020). 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, Varol, and Kiciman 2017).

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. Typical 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 (Johansson, Shalit, and Sontag 2016; Veitch, Sridhar, and Blei 2020; Wang et al. 2024). These methods also demonstrate impressive performance in domains such as recommender systems (Schnabel et al. 2016; Li et al. 2023a,b) and computer vision (Hu et al. 2022; Duan et al. 2023).

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

Table 1: The average ATE of positive and negative sentiment words as treatments on the Kindle dataset (He and McAuley 2016), grouped by the difference in the number of positive and negative sentiment words excluding the treatment word.

However, a critical issue when using causal effects to identify causally related words is that when multiple causally related words appear in the same 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 unexpectedly 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 is positive in each subgroup, we find that the absolute value of 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 spurious correlated words, thus decreasing the text classifier robustness.

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, 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 spurious correlated words and thus removed to achieve robust text classification. Extensive experiments are conducted on four public datasets, demonstrating the superiority of our proposal on both spurious correlated words identification and robust text classification.

Preliminaries

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. To classify the sentiment of the sentence, we 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 spurious correlated words. For instance, the word *and* is spurious correlated with the positive sentiment label in the IMDB movie reviews, but not in the Kindle book reviews. On the opposite, 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 $\mathcal{W}' = \{w'_1, w'_2, \ldots, w'_S\} \subseteq \mathcal{W}$ most likely to be spuriously correlated to the sentiment label and remove them from the sentences for training a robust text classifier $f(g(s \setminus \mathcal{W}'); \theta)$.

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, Shpitser, and Dredze 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.

Proposed Method

Unfortunately, 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 directly using causal effects, we propose to identify the causally related words via the probability of necessity (PN) and the probability of sufficiency (PS). Specifically, we first theoretically derive the identification results under different sentiment monotonicities, and further propose an 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 1 (Probability of Necessity (Pearl 2022)). *The*

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	×	?

Table 2: The sentences can be divided into eight strata according to the treatment T, observed outcome Y, and potential outcomes $Y(0)$ and $Y(1)$, with the unobserved one highlighted in red. For each stratum, counterfactual necessity and sufficiency either hold (✓), do not hold (×), or unknown (?).

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 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., $\mathbb{P}(Y(1) = 1 - y \mid T = 0, Y = y, X)$.*

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 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 1 (Identifiability Under Monotonicity)**.** *Under Assumption 1 that $Y(1) \geq Y(0)$, the probability of necessity and the probability of sufficiency are identifiable:*

$$\mathbb{P}(Y(0) = 0 \mid T = 1, Y = 1, X) = 1 + \frac{\mathbb{P}(Y = 0 \mid T = 0, X) - 1}{\mathbb{P}(Y = 1 \mid T = 1, X)},$$

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

Under Assumption 1 that $Y(1) \leq Y(0)$, the probability of necessity and the probability of sufficiency are identifiable:

$$\mathbb{P}(Y(0) = 1 \mid T = 1, Y = 0, X) = 1 + \frac{\mathbb{P}(Y = 1 \mid T = 0, X) - 1}{\mathbb{P}(Y = 0 \mid T = 1, X)},$$

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

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

Input: 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
        $\mathrm{PN}_m \leftarrow 1 + \frac{1}{n_{\mathrm{pos}}} \sum_{i: y_i = 1} \frac{\hat{\mathbb{P}}(Y=0|T=0,X)-1}{\hat{\mathbb{P}}(Y=1|T=1,X)}$;
        $\mathrm{PS}_m \leftarrow 1 + \frac{1}{n_{\mathrm{neg}}} \sum_{i: y_i = 0} \frac{\hat{\mathbb{P}}(Y=1|T=1,X)-1}{\hat{\mathbb{P}}(Y=0|T=0,X)}$;
    else
        $\mathrm{PN}_m \leftarrow 1 + \frac{1}{n_{\mathrm{neg}}} \sum_{i: y_i = 0} \frac{\hat{\mathbb{P}}(Y=1|T=0,X)-1}{\hat{\mathbb{P}}(Y=0|T=1,X)}$;
        $\mathrm{PS}_m \leftarrow 1 + \frac{1}{n_{\mathrm{pos}}} \sum_{i: y_i = 1} \frac{\hat{\mathbb{P}}(Y=0|T=1,X)-1}{\hat{\mathbb{P}}(Y=1|T=0,X)}$;
    end
end
Remove the words with the lowest $K_{\mathrm{PN}}\%$ PN and the
  lowest $K_{\mathrm{PS}}\%$ PS;
Train a robust $f(x; \theta)$ using the remaining words;
Output: robust transferable text classifier $f(x; \theta)$.
```

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 identifiability results:

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

	# docs	# top words	# causal	# spurious
IMDB	10,662	366	174	90
Kindle	20,232	270	100	119
Toxic	15,216	329	63	40
Toxic Tweets	6,774	341	45	72

Table 3: Datasets summary.

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

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

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

From Theorem 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 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 13). To obtain the robust text classifier, we finally remove the words with the lowest PN and PS by a certain percentage (line 14) and re-train the text classifier f using the remained sentence contexts (line 15). We summarize the overall algorithm in Algorithm 1.

Notice that the proposed algorithm does not require accurate estimations of PN, PS, or ATE. For PN (PS), we only need to make sure the bottom $K_{PN}\%$ ($K_{PS}\%$) words have smaller PN (PS) than the upper $1 - K_{PN}\%$ ($1 - K_{PS}\%$) words. While for ATE, the only requirement is that the sign of $\hat{\tau}_m$ is correct. This further enhance the robustness of our algorithm in addition to the advantages of the metrics PN and PS themselves over the widely adopted causal effects.

Experiments

Dataset and Preprocessing

Following Wang and Culotta (2020), we conduct experiments on four public datasets for binary classification tasks, specifically using the **IMDB** and **Kindle** datasets for sentiment classification and the **Toxic** and **Toxic Tweets** datasets for toxicity detection, with the detailed information as below:

- **IMDB:** This dataset includes sentences labeled with sentiment polarity (positive/negative) from movie reviews (Pang and Lee 2005).
- **Kindle:** This dataset collected product reviews from Amazon Kindle Store reviews (He and McAuley 2016). The original reviews are labeled as five-scale from 1 to 5. We follow Wang and Culotta (2020) to label the sentences with ratings 4 and 5 as positive, whereas 1 and 2 as negative, and remove the other sentences with rating 3.
- **Toxic:** This dataset collected comments from Wikipedia's talk page (Wulczyn, Thain, and Dixon 2017), in which the toxicity is labeled by human raters. Each comment was displayed to up to 10 raters with the original ratings from 0.0 to 1.0. For each comment, we take the average of the human-annotated ratings and label the comment as toxic (positive) if the average toxicity score is larger than 0.7 and as non-toxic (negative) if the average toxicity score is lower than 0.5. Similarly, we ignore the comments with the average toxicity score between 0.5 and 0.7.
- **Toxic Tweets:** This dataset collected comments from Twitter Streaming API and labeled toxic/non-toxic also by human raters (Radfar, Shivaram, and Culotta 2020).

In addition, Wang and Culotta (2020) manually labeled some of the words by identifying whether they are causally related or spurious correlated for each dataset. We summarize the summary statistics of the four datasets in Table 3.

Baselines

Despite there are many ATE-based methods for learning robust text classifiers, considering that the purpose of this paper is not to estimate more accurate ATEs, but to exploit PNs and PSs to determine whether words are causally related or spuriously correlated, we only compare our method with the matching-based ATE estimation methods (Falavarjani et al. 2017; Wang and Culotta 2020) for illustrative purposes.

Specifically, given the estimated ATEs for words that are most strongly associated with the class labels, a straightforward approach is to rank the absolute values of the estimated ATEs of these words and predict words with ATEs less than a certain threshold as spuriously correlated words, named **model-free** method. To further exploit the human-annotated labels of whether words are causally related or spurious correlated, we train a new word classification model using logistic regression for predicting the probability of a word being causally related, by augmenting the estimated ATEs to have more diverse input features, e.g., the ATE restricted to the top-5 most similar matches for sentences, the word's coefficient from the initial sentence classifier using a logistic regression with the bag-of-words as features to predict the sensitment/toxicity label (Wang and Culotta 2020). We rank the probability of being spurious derived from the word classification model and predict words with predicted probability of being spurious less than a given threshold as spuriously correlated words, named **model-based** method.

We also apply the **Oracle** method that identifies the spurious words with the ground truth label as baseline.

Hyperparameters and Implementation Details

First, following Wang and Culotta (2020), we use a logistic regression model with bag-of-words as the features to predict the sentiment/toxicity of the sentence and extract words that are most strongly associated with the class, named top words. For IMDB and Kindle datasets, words with coefficients larger than 1.0/lower than -1.0 are chosen as positive/negative top words, and for Toxic and Toxic Tweets, words with coefficients larger than 1.0/0.7 are selected as top words. Note that we do not consider the word with large negative coefficients

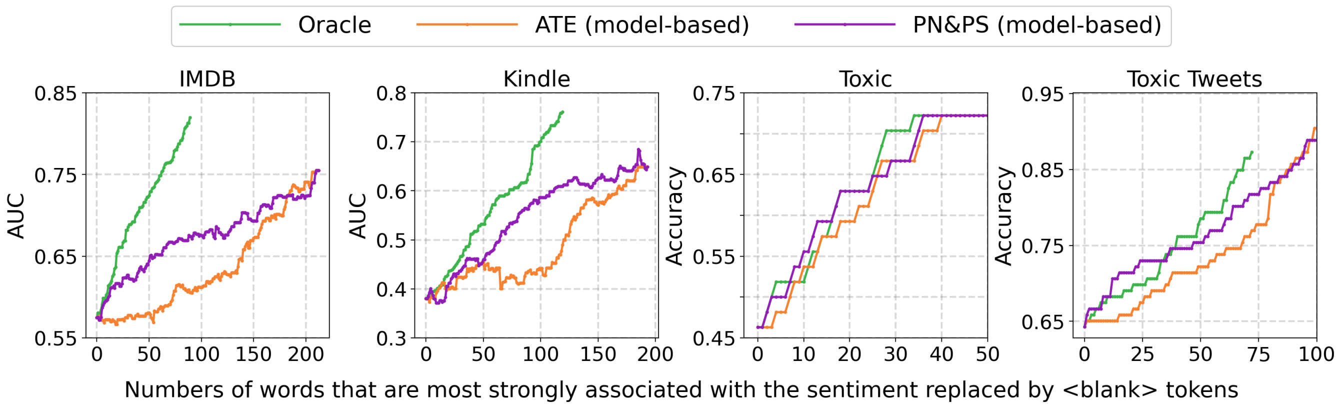

Figure 1: Performance of sentiment/toxicity classifications on test sets for four datasets.

because the goal of these two datasets are only to identify the toxic words and regard the other words as non-toxic.

Second, for each top word that appears in more than 10 sentences in each dataset, we estimate the PN and PS for our method using the Algorithm 1. To compute $\hat{\mathbb{P}}(Y \mid T = 1, X)$ and $\mathbb{P}(Y \mid T = 0, X)$ within the estimators of PN and PS, we first use principle component analysis (PCA) on the Bert embeddings for each sentence to obtain the covariates X, and then fit two logistic regression models separately on sentences with and without the top word of interest. We tune the PCA dimension in $\{3, 5, 10, 20, 30, 40, 50\}$.

Finally, for both model-free and model based method identified by estimated ATEs and estimated PNs and PSs, we replace the spurious correlated words by the <blank> token in the sentence, and then train a final sentiment/toxicity classification model. Motivated by (Wang and Culotta 2020), we select the sentence where spurious word is negatively related to the sentiment/toxicity label, that is, the spurious positive/negative word in the negative/positive sentiment or toxic/non-toxic sentence as the test set.

Performance Comparison

Sentiment Classification and Toxicity Detection We sequentially replace the words with the highest predicted probabilities of being spurious to <blank> token and examine the classification performance after replacing words identified as spuriously correlated words with model-based methods. Figure 1 shows the classification performance on the test set of four datasets with the varying numbers of replaced words. The results show that both the ATE-based and PN&PS-based method can increase the classification accuracy in test sets when as the number of replaced spurious words increasing. This highlights the importance of discovering spurious words and replacing spuriously correlated words in classification. The PN&PS-based method outperforms the ATE-based method, suggesting that PN and PS can better facilitate the classification task compared with the ATE.

Word Classification To investigate the reasons of the sentiment/toxicity classification performance improvement, an important question is whether our method can distinguish the spuriously correlated words more accurately than the ATE-based baselines. Table 4 shows the accuracy for each classifier on four datasets. To ensure the fair comparison, we fix the number of the replaced words as the overall number of manually labeled spurious words as shown in Table 3. The accuracy are measured by the proportion of spurious words among all replaced words. The results indicate that our PN&PS-based approaches perform better than the ATE-based approaches in distinguishing spurious words for sentiment datasets (IMDB and Kindle). On toxic datasets, though the model-free models perform poorly, the model-based methods can greatly enhance the word classification performance, which still can demonstrate the effectiveness of identifying causal words of our PN&PS-based methods.

	IMDB	Kindle	Toxic	Toxic Tweets
ATE (model-free)	0.348	0.445	**0.421**	**0.625**
PN&PS (model-free)	**0.393**	**0.454**	0.289	0.611
ATE (model-based)	0.315	0.370	0.553	0.736
PN&PS (model-based)	**0.461**	**0.412**	**0.632**	**0.750**

Table 4: Causal/spurious word classification performance measured by accuracy with the outperforming results bolded.

Conclusion

This paper proposes a robust text classification approach using PN and PS to distinguish causally related words from spuriously correlated words. Theoretically, we derive the identifiability results of PN and PS under different sentiment monotonicities. Empirically, we conduct extensive experiments to validate the superiority of our approach in causally related words identification and downsteam tasks such as sentiment classification and toxicity detection. One possible limitation of this paper is that the monotonicity assumption may be violated for a few sentences with the presence of words like *not*, *never* and *hardly*. Another limitation of this paper, which also served as our future research direction, is to explore a more effective way to utilize the estimated PN and PS for robust text classification rather than simply removing the words with PN and PS lower than a given threshold.

References

De Choudhury, M.; Kiciman, E.; Dredze, M.; Coppersmith, G.; and Kumar, M. 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*, 2098–2110.

Duan, Y.; Zhao, Z.; Qi, L.; Zhou, L.; Wang, L.; and Shi, Y. 2023. Towards semi-supervised learning with non-random missing labels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 16121–16131.

Falavarjani, S. M.; Hosseini, H.; Noorian, Z.; and Bagheri, E. 2017. Estimating the effect of exercising on users' online behavior. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11, 734–738.

He, R.; and McAuley, J. 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*, 507–517.

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

Hu, X.; Niu, Y.; Miao, C.; Hua, X.-S.; and Zhang, H. 2022. On non-random missing labels in semi-supervised learning. In *International Conference on Learning Representations*.

Imbens, G. W.; and Rubin, D. B. 2015. *Causal inference in statistics, social, and biomedical sciences*. Cambridge university press.

Johansson, F.; Shalit, U.; and Sontag, D. 2016. Learning representations for counterfactual inference. In *International Conference on Machine Learning*, 3020–3029. PMLR.

Li, H.; Lyu, Y.; Zheng, C.; and Wu, P. 2023a. TDR-CL: Targeted doubly robust collaborative learning for debiased recommendations. In *International Conference on Learning Representations*.

Li, H.; Xiao, Y.; Zheng, C.; Wu, P.; and Cui, P. 2023b. Propensity matters: Measuring and enhancing balancing for recommendation. In *International Conference on Machine Learning*, 20182–20194. PMLR.

Maas, A.; Daly, R. E.; Pham, P. T.; Huang, D.; Ng, A. Y.; and Potts, C. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human Language Technologies*, 142–150.

Morgan, S. 2015. *Counterfactuals and causal inference*. Cambridge University Press.

Olteanu, A.; Varol, O.; and Kiciman, E. 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*, 370–386.

Pang, B.; and Lee, L. 2005. Seeing Stars: Exploiting Class Relationships for Sentiment Categorization with Respect to Rating Scales. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, 115–124. Association for Computational Linguistics.

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

Pearl, J. 2009. *Causality*. Cambridge university press.

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

Pham, T. T.; and Shen, Y. 2017. A Deep Causal Inference Approach to Measuring the Effects of Forming Group Loans in Online Non-profit Microfinance Platform. arXiv:1706.02795.

Pryzant, R.; Card, D.; Jurafsky, D.; Veitch, V.; and Sridhar, D. 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*, 4095–4109.

Radfar, B.; Shivaram, K.; and Culotta, A. 2020. Characterizing Variation in Toxic Language by Social Context. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, 959–963.

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

Saha, K.; Sugar, B.; Torous, J.; Abrahao, B.; Kıcıman, E.; and De Choudhury, M. 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, 440–451.

Schnabel, T.; Swaminathan, A.; Singh, A.; Chandak, N.; and Joachims, T. 2016. Recommendations as treatments: Debiasing learning and evaluation. In *International Conference on Machine Learning*, 1670–1679. PMLR.

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

Sridhar, D.; Springer, A.; Hollis, V.; Whittaker, S.; and Getoor, L. 2018. Estimating causal effects of exercise from mood logging data. In *IJCAI/ICML Workshop on CausalML*.

Veitch, V.; Sridhar, D.; and Blei, D. 2020. Adapting text embeddings for causal inference. In *Conference on Uncertainty in Artificial Intelligence*, 919–928. PMLR.

Wang, H.; Fan, J.; Chen, Z.; Li, H.; Liu, W.; Liu, T.; Dai, Q.; Wang, Y.; Dong, Z.; and Tang, R. 2024. Optimal transport for treatment effect estimation. *Advances in Neural Information Processing Systems*, 36.

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

Wood-Doughty, Z.; Shpitser, I.; and Dredze, M. 2018. Challenges of using text classifiers for causal inference. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, volume 2018, 4586.

Wulczyn, E.; Thain, N.; and Dixon, L. 2017. Ex Machina: Personal Attacks Seen at Scale. In *Proceedings of the 26th International Conference on World Wide Web*, 1391–1399.