The Journal of Supercomputing (2025) 81:386
https://doi.org/10.1007/s11227-024-06887-z

Estimating textual treatment effect via causal disentangled representation learning

Zhimi Yang[1,2] · **Bo Shen**[1,2]

Accepted: 23 December 2024

Abstract

Estimating causal effects from observational data reveals the potential outcomes of different treatments. However, the methods primarily focusing on numerical or categorical covariates leave causal inference with textual observational data as an unresolved issue. Specifically, the high-dimensional and unstructured nature of text complicates the learning of representation vectors of causal structure from textual covariates. This complexity is principally due to the interaction of different factors within the textual covariates, making separating these factors crucial for the accurate estimation of textual causal effects. To address this challenge, we propose a causal disentangled representation learning method based on variational inference. The method derives latent factors from observed textual covariates and decomposes them into instrumental, confounding, and adjustment factors. Additionally, a learning criterion that minimizes mutual information is employed to ensure the independence of disentangled factors, and targeted regularization based on nonparametric estimation is applied to reduce residual bias. Experimental results show that the proposed method performs well on textual causal effect datasets and has higher performance and competitiveness compared to strong baseline methods.

Keywords Disentangled representation · Causal inference · Textual observational data · Latent factors · Confounding bias

✉ Bo Shen
bshen@bjtu.edu.cn

Zhimi Yang
22120173@bjtu.edu.cn

[1] School of Electronic and Information Engineering, Beijing Jiaotong University, Beijing 100044, China

[2] Key Laboratory of Communication and Information Systems, Beijing Municipal Commission of Education, Beijing 100044, China

Published online: 11 January 2025

1 Introduction

Causal inference is the process of estimating the impact of treatments on outcomes [23, 36], which focuses on uncovering causal relationships, not just correlations [2]. The application of deep learning to causal inference has recently garnered significant interest from scholars [11, 54]. Most current deep causal effect estimation models focus on classification or numerical variables, while text data in natural language is crucial as it serves as the primary medium for human communication and knowledge documentation. Indeed, causal problems in computational social science often involve textual components, such as social media posts, essays, and reviews, making the estimation of causal effects from textual data essential. For instance, how do positive product reviews causally influence product sales [38], and how does the gender displayed in user profiles causally impact the number of likes received on posts [14]? Deep causal models focusing on non-textual data often struggle to effectively address causal inference in natural language. The limitation arises because high-dimensional textual data contains richer information spanning topics, semantics, and sentiments [9]. These characteristics present a new challenge in estimating the causal effect of textual data, as its high dimensionality and diversity make analysis complex.

When performing causal inference on textual data, it requires extracting information about treatments, confounders, and outcomes from the data. This involves transforming the text into simpler, lower-dimensional representation vectors [13, 42]. Due to the implicit information contained in textual data and the coupling between different factors, it is crucial to learn representation vectors with causal structures to effectively identify latent variables and their relationships [32]. Several current studies have attempted to combine the fields of natural language processing (NLP) and causal inference [14], providing a more comprehensive analysis. NLP offers various methods, such as topic modeling [24] and contextual embedding [32], to extract necessary information from text for causal effect estimation. Roberts et al. [39] proposed a matching approach to resolve the problem of confounding caused by textual data in observational studies and relied on topic modeling to reduce the dimension of the text. Recently, fine-tuned language models such as BERT [12] have performed well on semantic benchmarks. Some research has developed causally sufficient embeddings that retain enough information for causal identification by fine-tuning BERT models [38, 47]. By imposing multiple loss constraints, Zhou et al. [58] addressed the bias issues in estimating textual causal effects, which were caused by the insufficient consideration of non-confounding variables in previous work.

Most methods are built on the assumption that textual observational data contains sufficient information to identify causal relationships. In these methods, it is common practice to use methods of NLP for representing text and then refine these representations by using deep learning structures containing causal information [4]. These methods are effective to some extent. But high-dimensional textual covariates do not contain only confounders, as factors that affect treatment or outcome may also be contained. Similar to traditional causal inference challenges, estimating treatment effects requires proper identification and control of different types of variables. For instance, when examining the causal effect of

a medication on weight loss, the treatment variable is medication use, and the instrumental variable could be the doctor's prescription decision, which may influence medication use but has no direct effect on weight loss. Dietary habits function as a confounding variable, affecting both medication use and weight loss, while exercise serves as an adjustment variable that must be controlled to accurately evaluate the effect of the medication. Research indicates that incorporating unnecessary covariates in high-dimensional settings leads to suboptimal outcomes, which increase both the bias and variance of the treatment effect estimate [1, 17]. In textual data, the complex interactions among multiple features result in representations containing various latent information, which hinders accurate causal inference. Therefore, learning textual representations that separate coupled information and discard irrelevant factors is a critical issue.

To solve this issue, we introduce a method for estimating textual treatment effects named Text-CDRL (text-causal disentangled representation learning). The method decomposes textual data into three latent factors: instrumental factors, confounding factors, and adjustment factors. We construct a multitask learning framework to disentangle latent factors effectively and adjust neural network design and training to enhance the accuracy and robustness of textual treatment effect estimation. Our main contributions are:

- We address a crucial issue in estimating causal effects from textual observational data, where textual covariates may include instrumental factors, confounding factors, and adjustment factors.
- We propose a multitask learning framework, Text-CDRL, based on causal disentangled representation learning, to derive three latent factors from textual covariates. The framework employs a learning criterion of mutual information minimization to achieve optimal, independently disentangled representations.
- In estimating the effect of the treatment from the text using decomposed latent factors, a targeted regularization approach based on nonparametric estimation is introduced to incorporate targeted learning into the estimation process.
- We extensively experiment on a range of semisynthetic datasets for textual causal inference, empirically validating the effectiveness of the proposed algorithm.

2 Related work

There are several different perspectives on current research methods, including deep causal models for non-textual data, causal inference with NLP, and targeted learning for causal inference.

2.1 Deep causal models for non-textual data

The absence of counterfactuals and the presence of selection bias pose significant challenges to estimating treatment effects from observational data [18, 54]. Deep learning models have been extensively employed in recent years to mine data for

causation as opposed to merely correlation [25]. Many studies strive to integrate deep neural networks and causal modeling to enhance the accuracy and objectivity of treatment effect estimation [52, 56]. When dealing with observational data, inferring latent confounders to disentangle associations between covariates is considered an important method for learning counterfactual representations. TARNet [44] balanced the information of the treatment and control groups through a shared layer, enabling neural networks to learn similar representations. On this basis, the DragonNet [45] model utilized the characteristics of neural networks that are good at finding correlations, as well as the process of targeted regularization, to filter out meaningful covariates for the purpose of reducing confounding. There are also methods that focus on learning disentangled representations of confounders and non-confounders. Latent variable modeling was integrated by Louizos et al. [34] to fit the interaction between latent confounding factors and treatment effects. Zhang et al. [57] proposed a data-driven TEDVAE algorithm using variational inference to infer latent variables from observed data. In addition, there are several ways to estimate treatment effects in observational studies by disentangling latent variables and learning counterfactual regressions [5, 6, 19, 51].

However, these deep causal models frequently neglect to address causal inference in unstructured textual data, where the high dimensionality of such data presents significant challenges for covariate extraction. Therefore, our research focuses on how to effectively apply deep learning and causal inference to textual data. We leverage the structural extensibility and representational capabilities of deep learning models to uncover causal relationships embedded in textual data.

2.2 Causal inference with NLP

Several current studies have attempted to integrate causal inference into textual representation learning to better capture causal relationships. These studies involve methods such as text as treatment [15, 38, 46, 50], text as confounder [10, 27, 39, 49], and text as outcome [13, 16, 29]. We focus on the area of causally driven textual representation learning, where Egami et al. [13] contend that latent representations of text are necessary for nearly all text-based causal inference, providing a framework for learning latent representations that reduce raw text to interpretable results. Current methods for combining high-dimensional text with causal models for dimensionality reduction primarily include topic models and embedding methods. Roberts et al. [39] addressed the problem of text-conditioned confounding in observational studies and relied on topic modeling to reduce the dimensionality of the text. Liu et al. [31] proposed a graph-based causal inference framework for discovering causal information in texts. With the rapid development in the field of NLP, common text embedding methods such as GLoVe [37] and BERT [12] are widely used for causal inference tasks in textual data. Textual covariates were represented by GloVe in CTAM [53], and information pertaining to virtually instrumental variables in latent representations was filtered out via a conditional treatment-adversarial learning procedure. CausalBERT [47] developed causally sufficient embeddings to retain enough information for causal identification and

efficiently estimated textual causal effects by fine-tuning a pre-trained BERT model. Pryzant et al. [38] improved proxy labels and fine-tuned BERT to adjust confounding parts of text using remote supervision methods in order to estimate the causal effects of linguistic attributes. Recently, Zhou et al. [58] addressed the problem of bias in the estimation of textual causal effects due to insufficient consideration of non-confounding variables by imposing constraints and utilizing multitask learning.

Current approaches to causal inference in text data typically do not consider a deeper disentangled causal representation of textual covariates. Capturing the complex causal relationships inherent in textual data requires a more in-depth decompositional representation of the causal relationships because textual data involves frequent interdependencies and latent variables. Our method conducts dimensionality reduction on textual data and subsequently decomposes it into distinct latent factors. To achieve independent decomposition factors, we employ a multitask learning approach, ensuring their mutual independence. Through this process, the decomposed latent factors can be obtained, laying the groundwork for a more precise and efficient evaluation of textual causal effects.

2.3 Targeted learning for causal inference

In causal inference, the double robust approach is an effective tool that combines double modeling of model treatments and outcomes and provides unbiased estimates of causal effects when one of the models is unbiased [8, 26]. Targeted maximum likelihood estimation (TMLE) utilizes information from propensity score weights and conditional outcome models to update and optimize effect estimates by maximizing the likelihood function [30, 43]. These theories make the TMLE technique suited for application in neural networks for causal impact estimation since it yields doubly robust, asymptotically efficient estimates of causal or target parameters. Inspired by the idea of TMLE, Shi et al. [45] introduced additional parameters in the causal effect estimation process to generate estimates with finite sample behavior and strong asymptotic guarantees. Vowels et al. [48] combined structured inference and targeted learning, applying regularizers from influence curves to reduce residual bias. In this study, we incorporate targeted learning into the process of estimating textual causal effects. Through a method that introduces targeted regularization, our goal is to produce estimates of the causal effects that are asymptotically efficient, unbiased, and doubly robust.

3 Preliminaries

Based on the preceding discussion, causal inference in text is hindered by its high dimensionality and unstructured nature, leading to coupling between factors. Handling all textual observational data as confounded can be problematic. A relevant example involves studying the causal impact of sentiment attributes in review texts on sales. Previous approaches typically consider observed text covariates as confounding factors, neglecting the deeper causal disentanglement of these text covariates. Not all observed textual covariates generally act as confounders influencing

both the treatment and the outcome. For instance, the tone of textual reviews may serve as a confounder, while key positive vocabulary probably functions as an instrumental factor affecting only the treatment. In contrast, the price of the product may act as an adjusting factor influencing only the outcome.

We begin by defining the notations pertinent to our context. Assuming each observed unit in textual causal inference is represented by a tuple $O_i = \{X_i, T_i, Y_i\}$. X_i is the textual data we observe (e.g., each product review), $T_i \in \{0, 1\}$ is the binary treatment variable, and Y_i is the outcome. We focus on the binary treatment variable. At $T_i = 1$, the i-th unit is categorized within the treatment group, for instance, with a positive sentiment attribute. At $T_i = 0$, the i-th unit is categorized within the control group, for instance, with a negative sentiment attribute. The observed dataset contains n randomly independent and identical observations from the distribution $O_i \overset{\text{i.i.d.}}{\sim} P$. In causal inference, one of the challenges in estimating causal effects is that we cannot observe the potential outcomes $Y_i(T_i = 0)$ and $Y_i(T_i = 1)$ for each unit, i.e., the counterfactual outcomes are unobservable. Under certain assumptions, it is feasible to estimate the treatment effect from observational data [36, 41]. We assume that the following three fundamental assumptions for treatment effect estimation hold [40]:

Assumption 1 *(SUTVA)* The treatment assignment of one unit must not impact the possible outcomes of other units in order to comply with the stable unit treatment value assumption.

Assumption 2 *(Unconfoundedness)* The treatment assignment is independent of potential outcomes given observed covariates: $Y_0, Y_1 \perp t|x$.

Assumption 3 *(Positivity)* Every unit has a nonzero probability of receiving each treatment level: $0 < P(t = 1|x) < 1$.

Under the potential outcomes framework [41], if the causal assumptions are satisfied, then we can define the average treatment effect (ATE) as follows:

$$\psi = \mathbb{E}[Y|\text{do}(T = 1)] - \mathbb{E}[Y|\text{do}(T = 0)] \tag{1}$$

The do operator shows that the effect we are interested in is causal and is used to characterize the operation of the treatment on causality. We assume that the observed text x carries enough information to adjust the confounding between t and y, i.e., blocking all backdoor paths. We can therefore use the adjustment method in causal inference to estimate the causal effect. The individual treatment effect (ITE) is defined as follows:

$$\tau_i = Y_i(T_i = 1) - Y_i(T_i = 0) \tag{2}$$

The ATE can be determined from the observed data as follows:

$$\psi = \mathbb{E}[\mathbb{E}[Y|X, T = 1] - \mathbb{E}[Y|X, T = 0]] \tag{3}$$

To estimate the average treatment effect from finite observed textual data x, the estimator is defined as follows:

$$\hat{\tau}(\hat{Q};x) = \frac{1}{n}\sum_{i=1}^{n}(\hat{Q}(1,x_i) - \hat{Q}(0,x_i)) \tag{4}$$

where $Q(t,x) = \mathbb{E}[Y|t,x]$ represents the conditional expected outcome, $g(x) = P(T = 1|x)$ denotes the propensity score, and $\hat{g}$ is an estimate of g. An estimator function $\hat{Q}$ that accurately predicts potential outcomes is our objective to learn. We aim to model g and Q using neural networks to estimate treatment effects from textual observational data.

4 The Text-CDRL algorithm

In this section, we present the Text-CDRL (text-causal disentangled representation learning) algorithm, designed to estimate causal effects from text data by learning disentangled representations of underlying causal factors. The model follows the causal structure in Figure. 1, and its architecture is shown in Figure. 2. Specifically, z_t is the instrumental variable that influences only the treatment but not the outcome, z_y is the adjustment variable that influences only the outcome but not the treatment, and z_c is the confounding variable that influences both the treatment and the outcome. Our strategy involves generating low-dimensional embedding vectors from text sequences and subsequently performing disentangled representations of these vectors. The text embedding vector is decomposed into three independent latent factors, which are then used to accurately identify causal effects in textual causal inference. The following subsections describe the core components and steps involved in the algorithm.

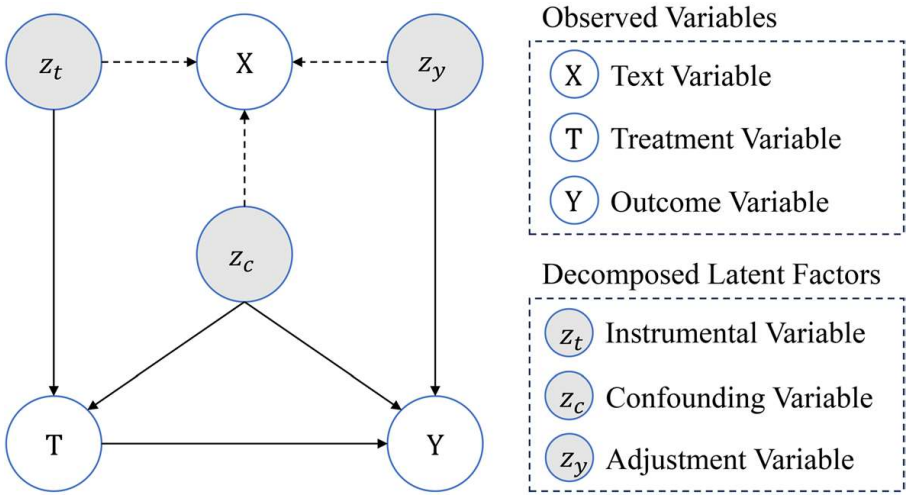

Fig. 1 Causal graph of the Text-CDRL model. Transparent nodes indicate variables that we can observe, and shaded nodes indicate latent factors decomposed in the text covariates

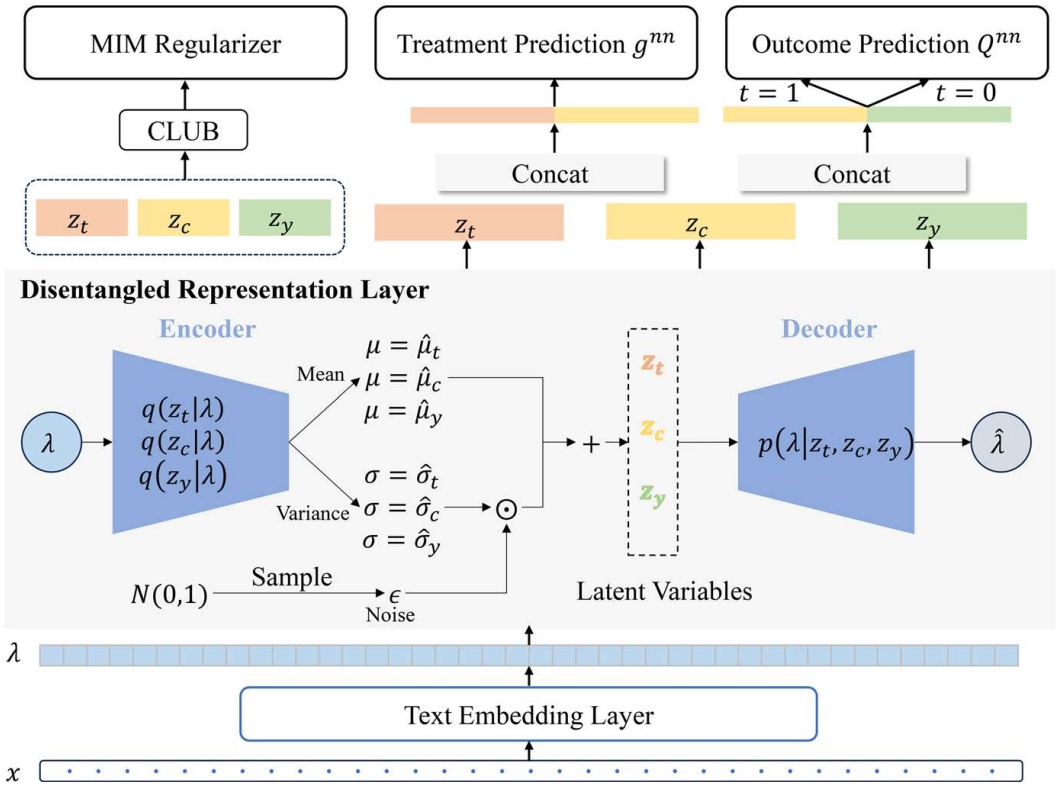

Fig. 2 The architecture of the proposed Text-CDRL algorithm. The text embedding is utilized to obtain the representation vector of the input observation text. The disentangled representation layer obtains three different representations of latent variables through variational inference, outputting z_t,z_c,z_y. Multitasking objectives are introduced at the top of the figure, including mutual information minimization regularizer, treatment prediction, and outcome prediction, ensuring that the treatment variables are predicted by z_t and z_c, and the outcome variables are predicted by z_y and z_c

4.1 Text embedding

We first transform the observed sentences into vectors through the text embedding layer. An effective strategy involves utilizing the pre-trained BERT model [12] to represent textual observational data. The input to BERT consists of the observed text x of each observation (e.g., product reviews), and the output is the hidden state of each token. We then extract the hidden state of the [CLS] token from the output of the BERT model to generate a representation of the text $\lambda = f(x)$.

Following [47], we introduce an objective function $\mathcal{L}_{mlm}$ based on a masked language model. The task involves randomly masking words from each text segment, and the model's objective is to predict the identity of these masked words.

4.2 Causal disentangled representation learning

According to the previous discussion, in textual causal inference, distinguishing between factors and removing unnecessary ones can result in a more precise and

efficient estimate of the causal effects. Therefore, building on the low-dimensional embedding $\lambda = f(x)$ of the text, we further disentangle the latent factors in the representation vector λ. To learn three independent latent factors, we propose a causal disentanglement representation layer. We use the latent variable model variational autoencoder (VAE) [21, 28] to infer latent instrumental, confounding, and adjustment factors.

Our objective is to learn the posterior distribution $p(z|\lambda)$ of the latent factor set $z = (z_t, z_c, z_y)$. Three distinct encoders, $q_{\phi_t}(z_t|\lambda)$, $q_{\phi_c}(z_c|\lambda)$, and $q_{\phi_y}(z_y|\lambda)$, are employed to predict the approximate variational posterior $q_\phi(z|\lambda)$. Encoder Φ_t infers latent instrumental factors, encoder Φ_c infers confounding factors, and encoder Φ_y infers adjustment factors. The variational posterior for approximating parameters $q_{\phi_t}(z_t|\lambda)$, $q_{\phi_c}(z_c|\lambda)$, and $q_{\phi_y}(z_y|\lambda)$ via the encoder is defined as follows:

$$q_{\phi_t}(z_t|\lambda) = \prod_{d=1}^{D_{z_t}} \mathcal{N}(\mu = \hat{\mu}_t, \sigma^2 = \hat{\sigma}_t^2) \tag{5}$$

$$q_{\phi_c}(z_c|\lambda) = \prod_{d=1}^{D_{z_c}} \mathcal{N}(\mu = \hat{\mu}_c, \sigma^2 = \hat{\sigma}_c^2) \tag{6}$$

$$q_{\phi_y}(z_y|\lambda) = \prod_{d=1}^{D_{z_y}} \mathcal{N}(\mu = \hat{\mu}_y, \sigma^2 = \hat{\sigma}_y^2) \tag{7}$$

where $\hat{\mu}_t$, $\hat{\mu}_c$, $\hat{\mu}_y$ and $\hat{\sigma}_t^2$, $\hat{\sigma}_c^2$, $\hat{\sigma}_y^2$ are the estimated mean and variance of latent variables z_t, z_c, z_y, respectively. We utilize a single decoder $p_\theta(\lambda|z_t, z_c, z_y)$ to reconstruct λ using latent variables.

The prior distributions for $p(z_t)$, $p(z_c)$, and $p(z_y)$ are chosen as Gaussian distributions. By modeling the latent variables with Gaussian distributions, we can effectively capture the mean and variance of each factor, ensuring that the model accounts for variability in the data. The approach facilitates the disentangling of instrumental, confounding, and adjustment factors while allowing for a flexible representation of the underlying factors in the data. We use (8) to infer latent variables:

$$\begin{aligned} \mu &= W_\mu \lambda + b_\mu \\ \log \sigma^2 &= W_\sigma \lambda + b_\sigma \\ z &= \mu + \sigma \odot \epsilon \end{aligned} \tag{8}$$

where W_μ and W_σ are weight matrices, b_μ and b_σ are bias terms for the respective linear transformations, and $\epsilon \sim \mathcal{N}(0, \mathbf{I})$. μ and σ^2 represent the mean and variance of a Gaussian distribution parameterized by a neural network. The reparameterization trick is expressed as $z = \mu + \sigma \odot \epsilon$, where ϵ is sampled from a standard normal distribution, and $\odot$ denotes element-wise multiplication. This reparameterization enables efficient backpropagation of gradients through the random sampling

process, allowing for gradient-based optimization during training. We sample from $q_\phi(z|\lambda) \simeq \mathcal{N}(\mu, \sigma^2 I)$ to generate $z \in \mathbb{R}^{D_z}$ as the latent vector, where D_z denotes the dimensionality of the latent space. Given a set of training samples, the objective of inferring latent variables from text representation is to maximize the evidence lower bound. Thus, the objective function of latent variables inferred through VAE is:

$$\begin{aligned}\mathcal{L}_{\text{vae}} = &- \mathbb{E}_{q_{\phi_t} q_{\phi_c} q_{\phi_y}} [\log p_\theta(\lambda|z_t, z_c, z_y)] \\ &+ D_{\text{KL}}(q_{\phi_t}(z_t|\lambda) \parallel p(z_t)) \\ &+ D_{\text{KL}}(q_{\phi_c}(z_c|\lambda) \parallel p(z_c)) \\ &+ D_{\text{KL}}(q_{\phi_y}(z_y|\lambda) \parallel p(z_y))\end{aligned} \tag{9}$$

To promote the disentanglement of latent factors and ensure that the inferred latent variables z_t, z_c, and z_y are as independent as possible, a learning criterion is proposed with the objective of minimizing mutual information. We aim to minimize the mutual information between latent variables to ensure their independence. Mutual information quantifies the degree of interdependence between two random variables. Given two random variables x and y, the mutual information I(x;y) is defined as:

$$\text{I}(x;y) = \mathbb{E}_{p(x,y)}\left[\log \frac{p(x,y)}{p(x)p(y)}\right] \tag{10}$$

We employ CLUB [7] as an upper bound to minimize mutual information between disentangled representations due to its computational efficiency in estimating mutual information. While direct minimization is often intractable, CLUB provides a practical upper bound that can be optimized efficiently, facilitating the separation of latent variables while ensuring their independence. When the conditional distribution between variables is known, CLUB is defined as:

$$\begin{aligned}\text{I}_{\text{club}}(x, y) =&\mathbb{E}_{p(x,y)}\left[\log p(y|x)\right] \\ &- \mathbb{E}_{p(x)}\mathbb{E}_{p(y)}\left[\log p(y|x)\right]\end{aligned} \tag{11}$$

Since the conditional distribution $p(y|x)$ is unknown in the actual scenario, we use the variational distribution $q_\theta(y|x)$ to approximate it:

$$\begin{aligned}\text{I}_{\text{vclub}}(x, y) =&\mathbb{E}_{p(x,y)}\left[\log q_\theta(y|x)\right] \\ &- \mathbb{E}_{p(x)}\mathbb{E}_{p(y)}\left[\log q_\theta(y|x)\right]\end{aligned} \tag{12}$$

Thus, we can learn independent latent variables in a way that minimizes mutual information:

$$\mathcal{L}_{\text{club}} = \text{I}_{\text{vclub}}(z_t, z_c) + \text{I}_{\text{vclub}}(z_c, z_y) + \text{I}_{\text{vclub}}(z_t, z_y) \tag{13}$$

To generate robust causal embeddings, we leverage the latent factors obtained in the previous step to predict both the treatment and the outcome. We include two auxiliary classification models, Q^{nn} and g^{nn}, to ensure that treatments t are predictable

from factors z_t and z_c, and outcomes y are predictable from factors z_y and z_c. For each conditional outcome model, we utilize neural networks $Q^{nn}(t_i, z_c, z_y; \theta^Q)$ to realize mappings: $z_i \rightarrow \hat{Q}(0, z_i^c, z_i^y; \theta^{Q_0})$ and $z_i \rightarrow \hat{Q}(1, z_i^c, z_i^y; \theta^{Q_1})$. Similarly, for the propensity score model, we employ neural networks $g^{nn}(z_t, z_c; \theta^g)$ to realize the mapping $z_i \rightarrow \hat{g}(z_i^t, z_i^c; \theta^g)$. Our training objective is to minimize $\mathcal{L}_s$:

$$\begin{aligned} \mathcal{L}_s = & \frac{1}{n} \sum_i [(Q^{nn}(t_i, z_c, z_y; \theta) - y_i)^2 \\ & + \text{CrossEntropy}(g^{nn}(z_t, z_y; \theta), t_i)] \end{aligned} \tag{14}$$

Through optimizing multitask objectives, we aim to disentangle causal representations in text. We expect factors influencing treatment variables to be included in z_t, factors affecting outcomes in z_y, and confounding information in z_c. The objective function for predicting text treatment effects using disentangled latent variables is:

$$\mathcal{L}_d = \mathcal{L}_s + \alpha \mathcal{L}_{vae} + \beta \mathcal{L}_{club} + \gamma \mathcal{L}_{mlm} \tag{15}$$

where the hyperparameters are α, β, and γ.

Algorithm 1 Text-CDRL: Text-Causal Disentangled Representation Learning

Algorithm 1 Text-CDRL: Text-Causal Disentangled Representation Learning

Require: Observed data $\{X_i, T_i, Y_i\}_{i=1}^{N}$, pre-trained BERT $f(\cdot)$, VAE encoder q_ϕ, VAE decoder p_θ, outcome predictor Q^{nn}, treatment predictor g^{nn}, regularization parameter ε, hyperparameters $\{\alpha, \beta, \gamma, \eta\}$, batch size m, and limit on the total number of iterations I

Ensure: Predicted potential outcomes $\hat{Y}(t)$ for $t \in \{0, 1\}$

Initialize Model parameters θ, regularization parameter ε, and weights $\{\alpha, \beta, \gamma, \eta\}$
for $iter = 1$ to I **do**
  Sample mini-batch $\{i_1, i_2, \ldots, i_m\} \subset \{1, 2, \ldots, N\}$
  Extract text embeddings: $\lambda_i \leftarrow f(X_i)$
  Calculate $L_{\text{mlm}} \leftarrow \text{MaskedLanguageModelLoss}(X_i, \lambda_i)$
  Disentangle latent causal factors using VAE:
  $(z_t, z_c, z_y), (\mu_t, \mu_c, \mu_y), (\log \sigma_t^2, \log \sigma_c^2, \log \sigma_y^2) \leftarrow q_\phi(\lambda_i)$
  Minimize mutual information using CLUB: $L_{\text{club}} \leftarrow \text{CLUBLoss}(z_t, z_c, z_y)$
  Calculate prediction loss L_s:
  $L_q \leftarrow \text{MSE}(Q^{nn}(t, z_c, z_y), Y), \quad L_g \leftarrow \text{CrossEntropy}(g^{nn}(z_c, z_t), T)$
  $L_s \leftarrow L_q + L_g$
  Calculate targeted regularization:
  $\xi_i \leftarrow (y_i - Q_{\text{reg}})^2, \quad L_{\text{targeted}} \leftarrow \eta \frac{1}{n} \sum \xi_i$
  Reconstruct embedding from latent factors:
  $\hat{\lambda}_i \leftarrow p_\theta(\text{concat}(z_t, z_c, z_y))$
  Calculate VAE optimization objectives:
  $L_{\text{recon}} \leftarrow \text{ReconstructionLoss}(\lambda_i, \hat{\lambda}_i)$
  $L_{\text{kl}} \leftarrow \text{KLDivergence}((\mu, \log \sigma^2), \mathcal{N}(0, I))$
  $L_{\text{vae}} \leftarrow L_{\text{recon}} + L_{\text{kl}}$
  Update model parameters: $L_{\text{total}} \leftarrow L_s + \alpha L_{\text{vae}} + \beta L_{\text{club}} + \gamma L_{\text{mlm}} + L_{\text{targeted}}$
  Optimize $[\hat{\theta}, \hat{\varepsilon}] \leftarrow \arg\min_{\theta, \varepsilon} L_{\text{total}}$
end for
for $t \in \{0, 1\}$ **do**
  $\hat{Y}(t) \leftarrow Q^{nn}(t, z_c, z_y)$
end for
return Predicted potential outcomes $\hat{Y}(0), \hat{Y}(1)$

4.3 Targeted regularization

Building upon propensity score models and expected conditional treatment effect models, inspired by targeted learning [30, 45, 48], we additionally propose a method of targeted regularization to enhance the estimation of text treatment effects. This method combines the advantages of nonparametric estimation theory, ensuring both consistency and robustness of the estimates. We introduce an additional model parameter ε, and a regularization term ξ, where I is an indicator function.

$$\tilde{Q}(\hat{g}, t_i, z_i^c, z_i^y, \hat{\varepsilon}) = \hat{Q}(t_i, z_i^c, z_i^y) + \hat{\varepsilon}\left(\frac{I(t_i = 1)}{\hat{g}(t_i = 1; z_i^t, z_i^c)} - \frac{I(t_i = 0)}{\hat{g}(t_i = 0; z_i^t, z_i^c)}\right) \tag{16}$$

$$\xi(y_i, t_i, z_i; \theta, \hat{\varepsilon}) = (y_i - \tilde{Q}(\hat{g}, t_i, z_i^c, z_i^y, \hat{\varepsilon}))^2 \tag{17}$$

The final objective function of our model after introducing targeted regularization is:

$$\mathcal{L} = \mathcal{L}_d + \eta \frac{1}{n} \sum_i \xi(y_i, t_i, z_i; \theta, \varepsilon) \tag{18}$$

$$\hat{\theta}, \hat{\varepsilon} = \underset{\theta, \varepsilon}{\operatorname{argmin}} \left[\mathcal{L}_d + \eta \frac{1}{n} \sum_i \xi(y_i, t_i, z_i; \theta, \varepsilon)\right] \tag{19}$$

where η represents the hyperparameter of the targeted regularization weights, controlling the regularization weight for the target. At convergence, $\hat{Q}$ and $\hat{g}$ have good nonparametric asymptotic properties, thereby satisfying the conditions of the efficient influence curve [20, 30]. By implementing targeted regularization, the estimation effect of the propensity score model $\hat{g}$ and the expected outcome model $\hat{Q}$ can be effectively improved to ensure the consistency and robustness of the estimation. Algorithm 1 provides a detailed description of the procedure.

5 Experiments

In this section, we empirically evaluate the performance of Text-CDRL for estimating textual causal effects via causal disentangled representation learning. Experimentally evaluating treatment effect estimates is challenging because known causal effects in the text are not available. Existing work in the field addresses this problem by generating semisynthetic datasets. In the following experiment, we empirically evaluate the validity of our proposed model.

5.1 Datasets

Estimating textual treatment effects requires a benchmark true causal effect. However, in practice, counterfactual outcomes are unobservable, making it challenging to obtain the actual causal effect. To address this challenge, following [47], we utilize real data, including text and metadata, in combination with data generated by a specific mechanism. This approach maintains data diversity and complexity while enabling model evaluation. In our experimental evaluation, we primarily use the following two datasets.

Amazon reviews dataset We utilize the same simulation methodology as in [38] to generate a semisynthetic dataset of Amazon reviews. Our objective is to investigate

the causal impact of sentiment in product reviews on the decision to click on the product. The treatment T represents the inferred sentiment from the review text, where $T = 1$ indicates positive sentiment, classifying the review into the treatment group, and $T = 0$ indicates negative sentiment, classifying the review into the control group. We employ the autocoder method from [35, 55] to infer the sentiment score for each review. Let C denote the observed additional covariates, representing the product types discussed in the reviews. We simulate the outcome Y.

$$Y \sim \text{Bernoulli}(\sigma(\beta_t T + \beta_c(\pi(C) - \beta_o) + \epsilon)) \tag{20}$$

where β_t controls the strength of the treatment, β_c controls the strength of confounding, β_o is the bias, $\pi(C) = P(T = 1|C)$ is the propensity score estimated from the metadata, and ϵ represents noise. The dataset ultimately contains 21,289 records. We partition the dataset into training, validation, and test sets in a 6:1:2 ratio, respectively, and conduct experiments using cross-validation.

Earnings call transcripts dataset This dataset is a balanced dataset constructed by [58] for causal inference from textual data. The datasets are constructed for two different treatment variables political risk and sentiment and are designed to study issues in finance. For example, consider how sentiment causally affects stock movement and stock volatility. Here, sentiment T serves as a binary treatment variable. Stock movement Y_{mov} is represented as a binary outcome variable, and stock volatility Y_{vol} is represented as a real-valued outcome variable. The specific information of this dataset is shown in Table 1. The dataset has 30,000 data points, split into training, validation, and test sets in an 8:1:6 ratio. Cross-validation is used to run the experiments.

5.2 Experiment setting

5.2.1 Evaluation criteria

To evaluate the performance of conditional average treatment effect (CATE) estimation, we use the precision in estimating heterogeneous effects (PEHE) metric [22], which assesses the model's accuracy in estimating causal effects at the individual level: $\sqrt{\epsilon_{\text{PEHE}}} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\tau_i - \hat{\tau}_i\right)^2}$. To evaluate the performance of average treatment effect (ATE) estimation, we use $\epsilon_{\text{ATE}} = |\tau - \frac{1}{N}\sum_{i=1}^{N}\hat{\tau}_i|$ as the metric. We report $\sqrt{\epsilon_{\text{PEHE}}}$ for the Amazon reviews dataset, as PEHE is employed as a more robust metric for individual-level causal effect estimation, whereas ATE may be less

Table 1 Specific information of the earnings call transcripts dataset. The treatment variable is sentiment

	Train		Dev		Test	
Outcome	Treatment	Control	Treatment	Control	Treatment	Control
Stock Movement	8,000	8,000	1,000	1,000	6,000	6,000
Stock Volatility	8,000	8,000	1,000	1,000	6,000	6,000

reliable under the given conditions. The results are detailed for both within-sample and out-of-sample performance. Please note that, as the neural network is not supervised for treatment effects, the estimation of treatment effects remains equally valid for both within-sample and out-of-sample results [45]. For the earnings call transcript dataset, we report the metrics $\sqrt{\epsilon_{\text{PEHE}}}$ and ϵ_{ATE}, consistent with previous work.

5.2.2 Implementation details

It is worth noting that the text embedding layer in our model is highly flexible and can accommodate any pre-trained model or its variants as the core architecture. To ensure a fair comparison with previous methods, for the Amazon review dataset, we employ the BERT-base-uncased [12] model for text embedding. For the earnings call transcripts dataset, we utilize FinBERT [3] for textual representation, as it is fine-tuned for financial texts. To reduce overfitting, we apply a dropout rate of 0.2 to the hidden representations. We optimize the model using the AdamW algorithm [33]. In the experiments, the model is trained for 3 epochs, the first 10% of the training steps are linear warm-ups, the learning rate is set to 1e-5, and the batch size is set to 32. In the disentangled representation, the dimensions of the latent variable spaces D_{z_t}, D_{z_c}, and D_{z_y} are 200. The weights for the loss functions are assigned as follows: $\alpha = 0.001$, $\beta = 1.0$, $\gamma = 0.01$, and $\eta = 1.0$. By employing cross-validation, the model is trained, with the best model being chosen according to its estimated performance on the validation set. The hyperparameter values are determined through a systematic search, where different combinations are tested to identify the configuration that optimizes model performance while ensuring robustness. To reduce the impact of randomness on individual experiment results, every experiment is carried out five times using distinct random seeds. We report the mean and standard deviation of the results from the random initialization parameters.

5.3 Baseline methods

To compare the performance of Text-CDRL for estimating textual causal effects, we evaluate it against seven baseline models. These models are commonly employed for estimating treatment effects from observational data and can be categorized into two groups: causal models based on deep neural networks and text representation models driven by causality.

Causal models based on deep neural networks include:

- *TARNet* [44] learns shared information representations through shared layers and estimates individual treatment effects by predicting potential outcomes through its network structure.
- *DragonNet* [45] leverages the sufficiency of propensity scores and the neural network's ability to identify correlations to select meaningful covariates.

- *CEVAE* [34] addresses the important issue of dealing with confounders as a means of estimating causal effects from observed data, incorporating the idea of VAE to uncover hidden confounding variables.
- *TEDVAE* [57] uses the observable variables to infer latent components, which are then used to untangle the variables and estimate the causal impact.

Text representation models driven by causality include:

- *CausalBERT* [47] creates causally adequate document embeddings. This allows for the estimation of causal effects from observational text data while maintaining pertinent causal information and controlling for confounding variables.
- *TextCause* [38] builds upon foundational assumptions, elevates the quality of proxy labels through distant supervision, and integrates causal adjustment methods to estimate the causal effects of latent linguistic features.
- *DIVA* [58] addresses the oversight of non-confounded covariates in causal inference from text by imposing various constraints to unveil interactions among different variables and alleviate bias in the estimation process.

For each baseline model, the same dataset is used for training as per the official source code or independent implementations. Ensure that all models are evaluated under the same experimental conditions.

5.4 Main results

The results of the experiments on the semisynthetic dataset of Amazon reviews are shown in Table 2. We observe that causally driven text representation learning models, such as CausalBERT, TextCause, and DIVA, consistently outperform deep causal models like TARNet, DragonNet, CEVAE, and TEDVAE. This suggests that traditional deep causal models fail to fully utilize the implicit information in text when dealing with complex textual data. And causally driven text representation learning models perform well in textual causal inference tasks by introducing language modeling and additional supervision.

Table 2 Mean and variance of the PEHE metric on the semisynthetic dataset of Amazon reviews. Lower is better, and bolded values indicate the best performers. 'ws' indicates within-sample and 'oos' indicates out-of-sample. The parameters are set to $\beta_t = 1, \beta_c = 1, \beta_o = 0.5, \epsilon = 1$

Method	$\sqrt{\epsilon_{\text{PEHE}}}$ ws	$\sqrt{\epsilon_{\text{PEHE}}}$ oos
TARNet	0.694±0.004	0.693±0.002
DragonNet	0.692±0.003	0.689±0.002
CEVAE	0.692±0.003	0.691±0.003
TEDVAE	0.689±0.003	0.689±0.002
CausalBERT	0.686±0.001	0.685±0.001
TextCause	0.684±0.001	0.683±0.001
DIVA	0.556±0.008	0.555±0.009
Text-CDRL	**0.504±0.003**	**0.504±0.004**

The evaluation metrics of our proposed model consistently outperform those of the leading baseline models, DIVA and TextCause, with the lowest $\sqrt{\epsilon_{\text{PEHE}}}$, demonstrating significant improvements. This provides empirical evidence that Text-CDRL is effective for estimating treatment effects from text. Furthermore, we observe that the causal effect estimates produced by Text-CDRL and DIVA, which utilize disentangled representations, significantly surpass those generated by CausalBERT and TextCause, which do not employ disentangled representations. This demonstrates that by decoupling covariates in textual data through representation learning, we can disentangle different factors within the text, thereby reducing confounding effects. This approach significantly enhances the accuracy of causal effect estimation and improves the interpretability and robustness of the model.

The difficulty in evaluating causal effect estimation in realistic scenarios lies in the lack of real causal validation data. However, a suitable technique for estimating the causal impact should not require unnecessary parameter adjustments and should be able to perform consistently across various datasets. Our experimental results on the earnings call transcripts dataset used by DIVA verify this. We analyze the causal effects of treatment sentiment on stock movements and stock volatility. From Table 3, it is evident that our approach achieves the lowest $\sqrt{\epsilon_{\text{PEHE}}}$ and ϵ_{ATE} among the compared methods, demonstrating outstanding performance. The experimental results illustrate that our model achieves accurate estimation of causal effects from textual observations by using a causally disentangled representation of the text and combining it with multitask learning. These results also further demonstrate the applicability of our model across various real-world scenarios and underscore its robustness to parameter selection, further validating its versatility and accuracy.

5.5 Simulation robustness evaluation

To assess our model's robustness across varied simulation settings, we systematically adjust simulation parameters and contrast their performance with baseline text representation models guided by causality, which exhibit superior performance. As

Table 3 Experimental results on a dataset of telephone earnings meeting transcripts. Causal effects of sentiment on stock movements as well as stock volatility. Lower is better and best results are shown in bold

Method	Stock Movement		Stock Volatility	
	$\sqrt{\epsilon_{\text{PEHE}}}$	ϵ_{ATE}	$\sqrt{\epsilon_{\text{PEHE}}}$	ϵ_{ATE}
TARNet	0.497±0.001	0.089±0.010	1.213±0.019	0.491±0.049
DragonNet	0.497±0.004	0.088±0.026	1.190±0.021	0.463±0.046
CEVAE	0.499±0.004	0.079±0.020	1.211±0.024	0.491±0.044
TEDVAE	0.497±0.007	0.098±0.023	1.228±0.056	0.459±0.099
CausalBERT	0.496±0.001	0.088±0.017	1.121±0.034	0.336±0.080
TextCause	0.522±0.009	0.030±0.028	1.100±0.019	0.114±0.028
DIVA	0.481±0.001	0.015±0.003	1.010±0.007	0.027±0.008
Text-CDRL	**0.480±0.001**	**0.007±0.006**	**1.005±0.003**	**0.014±0.005**

Table 4 Results of changing simulation parameter settings for the Amazon reviews semisynthetic dataset. Columns are labeled by the level of confounding: Low, medium, and high correspond to $\beta_c = 1$, 5, and 10, respectively. Lower is better, and bolded values indicate the best performers. 'ws' indicates within-sample, and 'oos' indicates out-of-sample

Confounding	Low		Medium		High	
Method	$\sqrt{\epsilon_{\text{PEHE}}}$ ws	$\sqrt{\epsilon_{\text{PEHE}}}$ oos	$\sqrt{\epsilon_{\text{PEHE}}}$ ws	$\sqrt{\epsilon_{\text{PEHE}}}$ oos	$\sqrt{\epsilon_{\text{PEHE}}}$ ws	$\sqrt{\epsilon_{\text{PEHE}}}$ oos
CausalBERT	0.6959	0.6715	0.7226	0.7299	0.8259	0.8136
TextCause	0.6945	0.6958	0.7334	0.7264	0.8221	0.8101
DIVA	0.5674	0.5649	0.6065	0.6067	0.6423	0.6384
Text-CDRL	**0.5116**	**0.5102**	**0.5604**	**0.5609**	**0.5659**	**0.5647**

Table 5 Ablation studies on latent representations. Results on semisynthetic dataset of Amazon reviews. Lower is better, and bolded values indicate the best performers. 'ws' indicates within-sample, and 'oos' indicates out-of-sample. 'w/o z_t': without z_t; 'w/o z_c': without z_c; 'w/o z_y': without z_y

Method	$\sqrt{\epsilon_{\text{PEHE}}}$ ws	$\sqrt{\epsilon_{\text{PEHE}}}$ oos
w/o z_t	0.5049	0.5049
w/o z_c	0.5118	0.5110
w/o z_y	0.5102	0.5093
Text-CDRL	**0.5041**	**0.5046**

shown in Table 4, after increasing the noise level to $\epsilon = 2$ and varying the confounding strength across different levels, including low ($\beta_c = 1$), medium ($\beta_c = 5$), and high ($\beta_c = 10$), our model consistently outperforms others under all simulation conditions. It also demonstrates insensitivity to changes in simulation parameters. This further highlights its robustness and adaptability to varying settings, providing stable results across diverse simulation environments and confirming its suitability for handling complex, real-world scenarios.

5.6 A study on latent representations

The impact of three latent representations, z_t, z_c, and z_y, on the Text-CDRL model's performance is then examined. When setting the latent representation dimensions D_{z_t}, D_{z_c}, and D_{z_y} to 200, we fix one of the latent variables, for instance, D_{z_t}, at 0 to evaluate the performance of Text-CDRL while disregarding the influence of z_t. Table 5 illustrates the ability of Text-CDRL to decompose latent representations. The data indicates that excluding any of the factors z_t, z_c, or z_y prevents the model from achieving optimal performance. However, when we set the dimensions of all latent factors to nonzero and decompose these three latent representations simultaneously, the model achieves optimal performance. The results indicate that decomposing the text covariates into instrumental, confounding, and adjustment factors is crucial for improving the model's performance. This highlights the significant role of each latent variable within the model. Considering all three latent variables

Table 6 Ablation studies in our proposed model. Lower is better, and bolded values indicate best performance. 'ws' indicates in-sample, and 'oos' indicates out-of-sample. 'w/o MIM': without mutual information minimization; 'w/o target-reg': without targeted regularization

Method	$\sqrt{\epsilon_{\text{PEHE}}}$ ws	$\sqrt{\epsilon_{\text{PEHE}}}$ oos
w/o MIM	0.5084	0.5095
w/o target-reg	0.5059	0.5053
Text-CDRL	**0.5041**	**0.5046**

simultaneously allows the model to more effectively extract useful information from the text data, thereby enhancing the accuracy of causal effect estimation.

5.7 Ablation study

To validate each part of our model's effectiveness, we do ablation experiments. Results of the ablation investigation are summarized in Table 6. The results on the Amazon review semisynthetic dataset indicate that the removal of mutual information minimization and targeted regularization leads to a performance degradation of Text-CDRL compared to the full model in terms of the $\sqrt{\epsilon_{\text{PEHE}}}$ metric. This demonstrates the contribution of each designed component to the model.

Removing the mutual information minimization component significantly decreases performance, indicating that this learning criterion effectively promotes the independence of latent variables, which is crucial for downstream text-causal effect estimation. Similarly, the removal of targeted regularization results in performance degradation, showing that the introduction of targeted regularization improves the performance of causal effect estimation and exhibits good asymptotic properties. The model achieves optimal performance only when all components are integrated. This holistic approach, combining mutual information minimization and targeted regularization, ensures that the latent factors remain independent and that the textual causal effect estimations are accurate and reliable.

6 Conclusion

In this paper, we present Text-CDRL, a novel method for textual causal inference. This method employs multitask learning to infer and disentangle three distinct latent representations from observed textual variables. During the text embedding stage, we represent and vectorize the input textual observational data using the pre-trained BERT model. To infer latent variables with causal structures from textual data, variational inference is employed to decouple different latent factors. Through the application of multiple constraints, our model effectively decomposes latent factors that contain causal information from the text. Subsequently, we utilize these disentangled factors for downstream estimation of textual causal effects. The effectiveness of our method is validated across a range of semisynthetic datasets, including the Amazon reviews dataset and the earnings call transcripts dataset.

Although our approach demonstrates promising results on the current dataset, several limitations remain. One challenge is scaling to larger datasets, where memory constraints and computational efficiency issues may arise, particularly during the large-scale parameter optimization involved in the disentangling of representations and causal effect estimation. Furthermore, while the model has primarily been evaluated on monolingual data, its performance in multilingual environments requires further investigation and refinement to ensure broader applicability. Moreover, real-world challenges, such as selection bias, skewed distributions of response variables, and violations of causal assumptions, could compromise the validity of causal inference, and these factors have yet to be fully addressed in our study. In future work, we intend to explore more robust disentangled representation methods for textual effect estimation, expand our approach to multilingual scenarios, and investigate ways to empirically validate methods in this domain to improve their applicability in complex real-world settings.

Funding This work was supported by National Natural Science Foundation of China (Grant No. 62376018).

References

1. Abadie Alberto, Imbens Guido W. (2006) Large sample properties of matching estimators for average treatment effects. Econometrica 74(1):235–267. https://doi.org/10.1111/j.1468-0262.2006.00655.x
2. Altman N, Krzywinski M (2015) Points of significance: association, correlation and causation. Nat Methods 12(10):899–900. https://doi.org/10.1038/nmeth.3587
3. Araci D (2019) Finbert: Financial sentiment analysis with pre-trained language models. arXiv preprint arXiv:1908.10063
4. Chalupka K, Eberhardt F, Perona P (2017) Causal feature learning: an overview. Behaviormetrika 44:137–164. https://doi.org/10.1007/s41237-016-0008-2
5. Cheng D, Xie Y, Xu Z, et al (2023) Disentangled latent representation learning for tackling the confounding m-bias problem in causal inference. In: 2023 IEEE International Conference on Data Mining (ICDM), IEEE, pp 51–60, https://doi.org/10.1109/ICDM58522.2023.00014
6. Cheng M, Liao X, Liu Q, et al (2022) Learning disentangled representations for counterfactual regression via mutual information minimization. In: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp 1802–1806, https://doi.org/10.1145/3477495.3532011
7. Cheng P, Hao W, Dai S, et al (2020) Club: A contrastive log-ratio upper bound of mutual information. In: International Conference on Machine Learning, PMLR, pp 1779–1788
8. Chernozhukov V, Chetverikov D, Demirer M et al (2018) Double/debiased machine learning for treatment and structural parameters. Economet J 21(1):C1–C68. https://doi.org/10.1111/ectj.12097
9. Chowdhary K, Chowdhary K (2020) Natural language processing. Fundamentals of artificial intelligence pp 603–649. https://doi.org/10.1007/978-81-322-3972-7_19
10. Daoud A, Jerzak CT, Johansson R (2022) Conceptualizing treatment leakage in text-based causal inference. arXiv preprint arXiv:2205.00465
11. Deng Z, Zheng X, Tian H, et al (2022) Deep causal learning: representation, discovery and inference. arXiv preprint arXiv:2211.03374
12. Devlin J, Chang MW, Lee K, et al (2018) Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805
13. Egami N, Fong CJ, Grimmer J, Roberts ME, Stewart BM (2022) How to make causal inferences using texts. Sci Adv 8(42):eabg2652. https://doi.org/10.1126/sciadv.abg2652

14. Feder A, Keith KA, Manzoor E et al (2022) Causal inference in natural language processing: estimation, prediction, interpretation and beyond. Trans Assoc Comput Linguist 10:1138–1158. https://doi.org/10.1162/tacl_a_00511
15. Fong C, Grimmer J (2016) Discovery of treatments from text corpora. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp 1600–1609
16. Gill M, Hall A (2015) How judicial identity changes the text of legal rulings. Available at SSRN 262078. https://doi.org/10.2139/ssrn.2620781
17. Häggström J (2018) Data-driven confounder selection via Markov and Bayesian networks. Biometrics 74(2):389–398. https://doi.org/10.1111/biom.12788
18. Hammerton G, Munafò MR (2021) Causal inference with observational data: the need for triangulation of evidence. Psychological Med 51(4):563–578. https://doi.org/10.1017/S0033291720005127
19. Hassanpour N, Greiner R (2019) Learning disentangled representations for counterfactual regression. In: International Conference on Learning Representations
20. He H, Wu P, Chen DG et al (2016) Statistical causal inferences and their applications in public health research. Springer
21. Higgins I, Matthey L, Pal A, et al (2017) beta-vae: Learning basic visual concepts with a constrained variational framework. ICLR (Poster) 3
22. Hill JL (2011) Bayesian nonparametric modeling for causal inference. J Comput Graph Stat 20(1):217–240. https://doi.org/10.1198/jcgs.2010.08162
23. Imbens GW, Rubin DB (2015) Causal inference in statistics, social, and biomedical sciences. Cambridge University Press
24. Jelodar H, Wang Y, Yuan C et al (2019) Latent dirichlet allocation (lda) and topic modeling: models, applications, a survey. Multimed Tools Appl 78:15169–15211. https://doi.org/10.1007/s11042-018-6894-4
25. Johansson F, Shalit U, Sontag D (2016) Learning representations for counterfactual inference. In: International Conference on Machine Learning, PMLR, pp 3020–3029
26. Kang JDY, Schafer JL (2007) Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data. Stat Sci 22(4):523–539. https://doi.org/10.1214/07-sts227
27. Keith K, Jensen D, O'Connor B (2020) Text and causal inference: A review of using text to remove confounding from causal estimates. In: Jurafsky D, Chai J, Schluter N, et al (eds) Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Online, pp 5332–5344, https://doi.org/10.18653/v1/2020.acl-main.474
28. Kingma DP, Welling M (2013) Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114
29. Koroleva A, Kamath S, Paroubek P (2019) Measuring semantic similarity of clinical trial outcomes using deep pre-trained language representations. J Biomed Inf 100:100058. https://doi.org/10.1016/j.yjbinx.2019.100058
30. van der Laan M, Rose S (2011) Targeted Learning: Causal Inference for Observational and Experimental Data. Springer Series in Statistics, Springer New York, https://books.google.com.tw/books?id=RGnSX5aCAgQC
31. Liu X, Yin D, Feng Y, et al (2021) Everything has a cause: Leveraging causal inference in legal text analysis. In: North American Chapter of the Association for Computational Linguistics
32. Liu Y, Lapata M (2018) Learning structured text representations. Trans Assoc Comput Linguist 6:63–75. https://doi.org/10.1162/tacl_a_00005
33. Loshchilov I, Hutter F (2017) Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101
34. Louizos C, Shalit U, Mooij JM, et al (2017) Causal effect inference with deep latent-variable models. Adv Neural Inf Process Syst 30
35. Maiya AS (2021) Causalnlp: A practical toolkit for causal inference with text. arXiv preprint arXiv:2106.08043
36. Pearl J (2009) Causality. Cambridge University Press
37. Pennington J, Socher R, Manning CD (2014) Glove: Global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp 1532–1543
38. Pryzant R, Card D, Jurafsky D, et al (2021) Causal effects of linguistic properties. In: Toutanova K, Rumshisky A, Zettlemoyer L, et al (eds) Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.

Association for Computational Linguistics, Online, pp 4095–4109https://doi.org/10.18653/v1/2021.naacl-main.323
39. Roberts ME, Stewart BM, Nielsen RA (2020) Adjusting for confounding with text matching. Am J Political Sci 64(4):887–903. https://doi.org/10.1111/ajps.12526
40. Rosenbaum PR, Rubin DB (1983) The central role of the propensity score in observational studies for causal effects. Biometrika 70(1):41–55. https://doi.org/10.1093/biomet/70.1.41
41. Rubin DB (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. J Edu Psychol 66(5):688. https://doi.org/10.1037/h0037350
42. Schölkopf B, Locatello F, Bauer S et al (2021) Toward causal representation learning. Proc IEEE 109(5):612–634. https://doi.org/10.1109/JPROC.2021.3058954
43. Schuler MS, Rose S (2017) Targeted maximum likelihood estimation for causal inference in observational studies. Am J Epidemiol 185(1):65–73. https://doi.org/10.1093/aje/kww165
44. Shalit U, Johansson FD, Sontag D (2017) Estimating individual treatment effect: generalization bounds and algorithms. In: International Conference on Machine Learning, PMLR, pp 3076–3085
45. Shi C, Blei D, Veitch V (2019) Adapting neural networks for the estimation of treatment effects. Adv Neural Inf Process Syst 32
46. Tan C, Lee L, Pang B (2014) The effect of wording on message propagation: Topic-and author-controlled natural experiments on twitter. arXiv preprint arXiv:1405.1438
47. Veitch V, Sridhar D, Blei D (2020) Adapting text embeddings for causal inference. In: Conference on Uncertainty in Artificial Intelligence, PMLR, pp 919–928
48. Vowels MJ, Camgoz NC, Bowden R (2021) Targeted vae: Variational and targeted learning for causal inference. In: 2021 IEEE International Conference on Smart Data Services (SMDS), IEEE, pp 132–141, https://doi.org/10.1109/SMDS53860.2021.00027
49. Weld G, West P, Glenski M, et al (2022) Adjusting for confounders with text: Challenges and an empirical evaluation framework for causal inference. In: Proceedings of the International AAAI Conference on Web And Social Media, pp 1109–1120, https://doi.org/10.1609/icwsm.v16i1.19362
50. Wood-Doughty Z, Shpitser I, Dredze M (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, NIH Public Access, p 4586
51. Wu A, Yuan J, Kuang K et al (2022) Learning decomposed representations for treatment effect estimation. IEEE Trans Knowl Data Eng 35(5):4989–5001. https://doi.org/10.1109/TKDE.2022.3150807
52. Yao L, Li S, Li Y, et al (2018) Representation learning for treatment effect estimation from observational data. Advances in neural information processing systems 31
53. Yao L, Li S, Li Y, et al (2019) On the estimation of treatment effect with text covariates. In: International Joint Conference on Artificial Intelligence, https://doi.org/10.24963/ijcai.2019/570
54. Yao L, Chu Z, Li S et al (2021) A survey on causal inference. ACM Trans Knowl Discov from Data (TKDD) 15(5):1–46. https://doi.org/10.1145/3444944
55. Yin W, Hay J, Roth D (2019) Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. arXiv preprint arXiv:1909.00161
56. Yoon J, Jordon J, Van Der Schaar M (2018) Ganite: Estimation of individualized treatment effects using generative adversarial nets. In: International conference on learning representations
57. Zhang W, Liu L, Li J (2021) Treatment effect estimation with disentangled latent factors. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp 10923–10930, https://doi.org/10.1609/aaai.v35i12.17304
58. Zhou Y, He Y (2023) Causal inference from text: Unveiling interactions between variables. arXiv preprint arXiv:2311.05286