On the Estimation of Treatment Effect with Text Covariates

Liuyi Yao[1] , **Sheng Li**[2] , **Yaliang Li**[3] , **Hongfei Xue**[1] , **Jing Gao**[1] and **Aidong Zhang**[4]
[1]University at Buffalo
[2]University of Georgia
[3]Alibaba Group
[4]University of Virginia
liuyiyao@buffalo.edu, sheng.li@uga.edu, yaliang.li@alibaba-inc.com, {hongfeix, jing}@buffalo.edu, aidong@virginia.edu

Abstract

Estimating the treatment effect benefits decision making in various domains as it can provide the potential outcomes of different choices. Existing work mainly focuses on covariates with numerical values, while how to handle covariates with textual information for treatment effect estimation is still an open question. One major challenge is how to filter out the nearly instrumental variables which are the variables more predictive to the treatment than the outcome. Conditioning on those variables to estimate the treatment effect would amplify the estimation bias. To address this challenge, we propose a conditional treatment-adversarial learning based matching method (CTAM). CTAM incorporates the treatment-adversarial learning to filter out the information related to nearly instrumental variables when learning the representations, and then it performs matching among the learned representations to estimate the treatment effects. The conditional treatment-adversarial learning helps reduce the bias of treatment effect estimation, which is demonstrated by our experimental results on both semi-synthetic and real-world datasets.

1 Introduction

Treatment effect, also known as causal effect, refers to the effect that one variable (i.e., the *treatment*) exerts on the other variable (i.e., the *outcome*). Accurate and reliable estimation of treatment effects would largely benefit decision making across various domains, due to its ability to reflect the outcomes of different choices. For example, in the medical domain, doctors can recommend the best therapy for a specific patient if the effect of the therapy on the recovery rate is known; in the education field, teachers can adopt the best teaching method according to the effect of teaching methods on the test score; and in the advertising area, business owners can choose the best platform to advertise based on the effect of the platform on the response rate. In the above examples, the therapy/teaching method/platform is the treatment, and the recovery rate/test score/response rate is the outcome.

The treatment effect is defined as the change of the outcome if the intervention is made on the treatment, supposing the covariates are unchanged (i.e., condition on those covariates), where covariates are the variables/features that are related to the treatment as well as the outcome. For example, in the aforementioned medical case, the patient's demographic information and physical examinations are the covariates. Many methods have been developed for treatment effect estimation [Imbens and Rubin, 2015].

Most of the existing work focuses on numerical covariates, while little attention has been paid to the textual covariates. However, in real world, text data are almost everywhere, such as clinical notes, movie reviews, news, social media posts, and etc. Different from the structured and well-defined numerical covariates, textual covariates contain richer information and can be summarized at different levels, such as word level, topic level, semantics level, and etc. This property of text data brings some new challenges into treatment effect estimation with textual covariates. In particular, some textual covariates that are very predictive to the treatment assignment might not be that predictive to the outcome. Such covariates are referred to as the *nearly instrumental variables*. In treatment effect estimation, existing work [Pearl, 2012; Wooldridge, 2016] has shown that conditioning on the nearly instrumental variables tends to amplify the bias in the analysis of causal effects. Therefore, the nearly instrumental variables should be excluded when estimating the treatment effect. Thus, the major challenge in estimating the treatment effect with textual covariates is: How to filter out the nearly instrumental variables?

In existing methods, filtering out the nearly instrumental variable is achieved by covariate re-weighting [Kuang *et al.*, 2017a; Chang and Dy, 2017; Diamond and Sekhon, 2013] or feature selection [Kuang *et al.*, 2017b; Tibshirani, 1996; Rassen *et al.*, 2011], when the covariates are numerical. However, when the covariate contains text data, the effectiveness of the re-weighting or feature selection based approaches would be limited, as those methods would be restricted to only one specific level of information contained in the textual variable, which leads to insufficient summarization of text covariates and further leads to insufficiency in filtering out nearly instrumental variables.

To handle the above challenges, we propose the **C**onditional **T**reatment-**A**dversarial learning based **M**atching method (**CTAM**), inspired by the conditional adversarial architecture in [Zhao *et al.*, 2017]. CTAM first learns the latent

representation of all covariates, in which the information contained in text variables can be fully summarized. Then in the learned representation space, we adopt the nearest neighbor matching (NNM), for its interpretability, to estimate the outcome if the treatment had been changed. The key characteristic of CTAM is the conditional treatment adversarial training procedure whose goal is to filter out the information related to nearly instrumental variables in the representation space. In this procedure, the treatment discriminator, along with the representation learner and the outcome predictor, play a minimax game: The treatment discriminator is trained to predict the treatment label correctly, while the representation learner, corporately working with the outcome predictor, aims to fool the treatment discriminator. Through the conditional treatment adversarial training procedure, the learned representation discards the extraneous information specific to treatment assignment, and meanwhile retains the information related to outcome prediction. Consequently, the proposed method benefits the treatment effect estimation with text covariates.

To evaluate the effectiveness of the proposed method, we first conduct experiments on two semi-synthetic datasets. Experimental results show that the proposed method outperforms the state-of-the-art methods. Furthermore, in the real world dataset, we verify the matching quality, and demonstrate that by imposing the conditional treatment adversarial training, the dependency between the treatment assignment and the nearly instrumental variables are removed.

2 Preliminaries

We first introduce some important notations. Let W denote the treatment, and $W \in \mathcal{R}^N$, where N is the number of records in the dataset, and W_i is the treatment of the i-th record. When the treatment is binary, the records with $W_i = 1$ form the *treated group* and the others belong to the *control group*. Let X denote all covariates excluding the textual covariates, and $X \in \mathcal{R}^{N\times d}$, where d is the number of non-textual covariates. Let T denote the textual covariate, and $T = \{T_1, T_2, \ldots, T_N\}$, where T_i is the text that belongs to the i-th record, and $T_i = [t_{i,1}, t_{i,2}, \ldots, t_{i,N_{T_i}}]$, where $t_{i,j}$ is the j-th word in the i-th record, and N_{T_i} denotes the total number of words in T_i. The outcomes of different treatments are the potential outcomes, and let Y_i^ω denote the potential outcome of the i-th individual/record with the ω-th treatment. The observed outcome (i.e., factual outcome) is denoted as Y^F, and $Y^F \in \mathcal{R}^N$.

In this work, we follow the potential outcome framework [Splawa-Neyman *et al.*, 1990; Rubin, 1974] and the following assumptions ensure that the treatment effect can be identified.

Assumption 1: ***Stable Unit Treatment Value Assumption (SUTVA)***. *The potential outcomes for any unit do not vary with the treatment assigned to other units, and, for each unit, there are no different forms or versions of each treatment level, which lead to different potential outcomes.*

Assumption 2: ***Consistency***. *The potential outcome of treatment w equals to the observed outcome if the actual treatment received is w.*

Assumption 3: ***Ignorability.*** *Given pre-treatment covariates, i.e., the covariates affect the treatment, treatment assignment is independent of the potential outcomes.*

Assumption 4: ***Positivity.*** *For any set of values of pre-treatment covariates $\mathbf{X}$, treatment assignment is not deterministic [D'Amour* et al.*, 2017]: $\forall w$ and $\mathbf{x}$, $\exists\ \eta \in (0, 0.5)$, s.t. $\eta < P(W = w|\mathbf{X} = \mathbf{x}) < 1 - \eta$.*

The treatment effect can be measured at the individual, population, and treated group level, which is known as the individual treatment effect (ITE), the average treatment effect (ATE) and the average treatment effect on the treated group (ATT), respectively. With the above four assumptions, the ITE, ATE, and ATT can be identified as[1]:

$$\begin{aligned}
\text{ITE}_i &= \mathbf{E}[\mathbf{Y}^1|\mathbf{X} = \mathbf{x}_i] - \mathbf{E}[\mathbf{Y}^0|\mathbf{X} = \mathbf{x}_i] = Y_i^1 - Y_i^0; \\
\text{ATE} &= \mathbf{E}_U[\mathbf{Y}^1 - \mathbf{Y}^0] = \tfrac{1}{|U|}\textstyle\sum_{i\in U}(Y_i^1 - Y_i^0); \\
\text{ATT} &= \mathbf{E}_{U_1}[\mathbf{Y}^1 - \mathbf{Y}^0] = \tfrac{1}{|U_1|}\textstyle\sum_{i\in U_1}(Y_i^1 - Y_i^0);
\end{aligned} \tag{1}$$

where Y_i^1 and Y_i^0 are the potential treated and control outcomes; U is the whole population and U_1 is the treated group.

With the above knowledge, the problem is defined as

Input: The non-textual covariates X, textual covariates T, treatment assignment W and the observed outcome Y^F.

Output: ITE, ATE, and ATT.

3 Methodology

3.1 Motivation

The underlying causal graph of our proposed method is shown in Figure 1. In the figure, Z and Z' together are the latent representations of the observed textual covariates T and non-textual covariates X. Among the latent variables, Z' denotes the nearly instrumental variables, which is more predictive to the treatment assignment than the outcome Y. As mentioned previously, conditioning on the nearly instrumental variables would amplify the treatment effect estimation bias. Our objective is to learn the latent representations that filter out the information related to nearly instrumental variables.

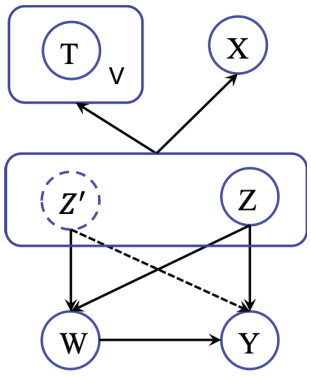

Figure 1: Causal Graph of the CTAM

[1]Here we define the treatment effect with binary treatment. For the case of treatment effect with multiple treatments, please refer to [Lopez *et al.*, 2017].

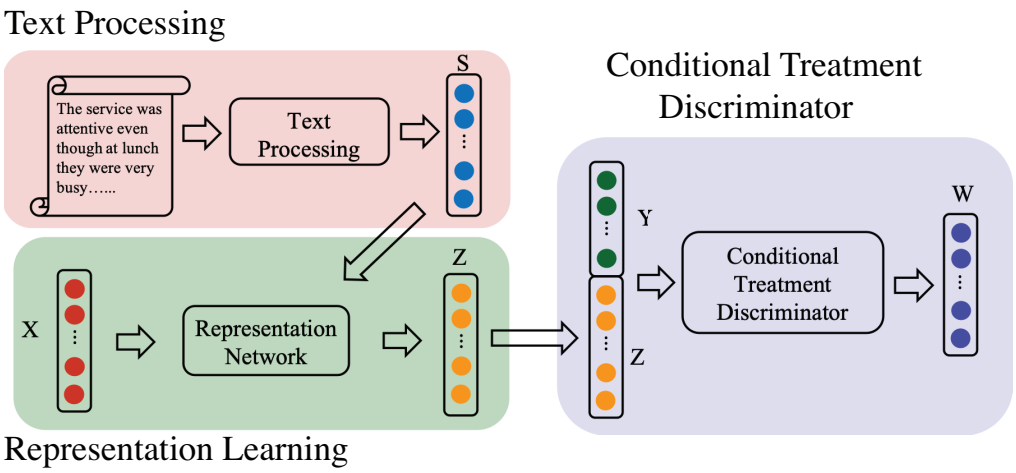

Figure 2: CTAM Framework

Our proposed method introduces the conditional treatment-adversarial learning to eliminate the information related to nearly instrumental variables $Z^{'}$ as much as possible in the latent representations.

3.2 Framework

Figure 2 shows the framework of the proposed CTAM method. CTAM contains three major components: text processing, representation learning, and conditional treatment discriminator. Through the text processing component, the original text is transformed into vectorized representation S. After that, S is concatenated with the non-textual covariates X to construct a unified feature vector, which is then fed into the representation neural network to get the latent representation Z. After learning the representation, Z, together with potential outcomes Y, are fed into the conditional treatment discriminator. During the training procedures, the representation learner plays a minimax game with the conditional treatment discriminator: By preventing the discriminator from assigning correct treatment, the representation learner can filter out the information related to nearly instrumental variables. The final matching procedure is performed in the representation space Z. The following sections introduce each component in detail.

3.3 Text Processing and Representation Learning

Text processing procedure converts the text data T to the numeric representation S. Various methods are developed to map the word from the vocabulary to the numerical vector, such as word embedding (GloVe [Pennington *et al.*, 2014], word2vec [Mikolov *et al.*, 2013], etc.), document-term matrix (bag of words [Harris, 1954; Manning *et al.*, 1999], inverse document frequency [Sparck Jones, 1972]). We adopt the word embedding learned by GloVe [Pennington *et al.*, 2014] in this procedure, and S is the average of all word-embeddings in one document.

Following the text processing, in the representation procedure, the learned numerical vector of textual covariate S is first concatenated with the numerical covariate X. The concatenated vector is denoted as C. After concatenation, a representation neural work is adopted to map the concatenated vector C to the latent representation Z: $Z = \Phi_{rep}(C; \Theta_\Phi)$, where Φ_{rep} is a feed-forward neural network with ReLU [Nair and Hinton, 2010] as the activation function and Θ_Φ denotes the set of its parameters.

The latent representation Z learned by Φ_{rep} contains the information related to nearly instrumental variables, which would amplify the treatment effect estimation bias. In order to eliminate such information, we design the following conditional treatment-adversarial learning procedure.

3.4 Conditional Treatment Discriminator and Conditional Treatment-Adversarial Learning

Conditional Treatment Discriminator

The input of the conditional treatment discriminator is the latent representation Z and the potential outcomes Y, and the output is the treatment assignment W. Our discriminator conditions on the potential outcomes, which allows the latent representation to correlate with the treatment only through the potential outcome distribution. In other words, by playing the minimax game, which is introduced in the following section, with the conditional treatment discriminator, the learned latent representation is capable of eliminating the conditional dependency with the treatment assignment.

The conditional treatment discriminator is a feed-forward neural network, denoted as $\mathcal{D}(Z, Y; \Theta_D)$, where Θ_D is a set of parameters. The goal of this discriminator is to correctly predict the treatment assignment. The loss of the conditional treatment discriminator is measured by the cross-entropy:

$$L_D(\Theta_D, \Theta_\Phi, \Theta_\Psi) = \mathbf{E}_{(c,w)\sim(C,W)}[-\log \mathcal{D}(\Phi_{rep}(c; \Theta_\Phi), \Psi_{pop}(\Phi_{rep}(c); \Theta_\Psi); \Theta_D)], \quad (2)$$

where $\Psi_{pop}(\cdot; \Theta_\Psi)$ is the pseudo outcome predictor with Θ_Ψ as its parameters, which is defined as follows.

Pseudo Outcome Prediction

The conditional treatment discriminator requires the potential outcomes of all treatments, which is formulated as:

$$Y = \Psi_{pop}(\Phi_{rep}(C); \Theta_\Phi) = \{\Psi_i(\Phi_{rep}(C); \Theta_\Phi^{(i)})\}_{i=1}^{n_w}, \quad (3)$$

where n_w is the number of total treatments, $\{\Psi_i\}_{i=1}^{n_w}$ is outcome prediction model for each treatment, Θ_Ψ is the set of parameters of Ψ_{pop}, and $\Theta_\Psi^{(i)}$ is the parameter set of i-th outcome prediction model. As the potential outcome here is only for the conditional treatment discriminator and is not the explicit result, we name it as the pseudo potential outcome.

Conditional Treatment-Adversarial Learning

The objective of conditional treatment-adversarial learning is to filter out the information related to the nearly instrumental variables. As the nearly instrumental variables refer to the variables that are more predictive to the treatment assignment instead of the outcome, this filtering strategy is equivalent to removing the conditional dependency between the latent representation and the treatment assignment. Therefore, we train an adversarial learning model to achieve this goal. The discriminator $\mathcal{D}$, along with Φ_{rep} and Ψ_{pop}, plays a minimax game. The discriminator $\mathcal{D}$ aims to minimize Eqn. (2) in order to assign correct treatment. Meanwhile, the representation learner Φ_{rep} and the outcome predictor Ψ_{pop} are trained to maximize the above loss to filter out the information that benefits the discriminator $\mathcal{D}$. When the conditional treatment discriminator can be successfully fooled, the information that enhances the treatment assignment is eliminated from the latent representation, i.e., the information related to nearly instrumental variables can be successfully filtered out.

3.5 Loss Function and Parameter Training

Loss Function

The final loss of the three-player game is:

$$\mathcal{L} = \mathcal{L}_p(\Theta_\Phi, \Theta_\Psi) - \lambda L_D(\Theta_D, \Theta_\Phi, \Theta_\Psi), \quad (4)$$

where L_D is defined in Eqn. (2) and λ is the hyper-parameter. $\mathcal{L}_p(\Theta_\Phi, \Theta_\Psi)$ is the sum of the group distance and the pseudo outcome prediction loss, which is defined as:

$$\begin{aligned} \mathcal{L}_p \quad = \quad & \textstyle\sum_{i=1}^{N_W} [\sum_{m=1}^{N_Y} \mathcal{L}_{pd}(Z^{\{i\}\{m\}}, Z^{\{i\}\{m\}}) \\ & -\alpha \textstyle\sum_{m \neq k} \mathcal{L}_{pd}(Z^{\{i\}\{m\}}, Z^{\{i\}\{k\}})] \\ & +\beta \mathcal{L}_{pseu}(\hat{Y}, Y^F), \end{aligned} \quad (5)$$

where N_Y is the number of label classes in the outcome, N_W is the number of treatments, $\mathcal{L}_{pd}(\cdot, \cdot)$ is the sum of pairwise distance between the two input matrices, $Z^{\{i\}\{m\}}$ denotes the representations of records that are assigned with the i-th treatment and their observed outcomes are the m-th label class, and $\hat{Y}$ is the corresponding prediction of the observed outcome, which can be obtained from Eqn. (3).

The first term in L_p measures the pairwise distance between the records sharing the observed outcome label under the same treatment, and the second term measures the pairwise distance between the records that have different observed outcomes. Minimizing the difference of two terms makes similar records close to each other, while dissimilar records far from each other in the representation space. The third term is the pseudo outcome prediction loss, and minimizing it allows better potential outcome predictions for conditional treatment discriminator.

Model Training

The training procedure involves optimizing the minimax game among the discriminator $\mathcal{D}$, representation learner Φ and the pseudo outcome predictor Ψ, which can be viewed as:

$$\text{mini}_{\Theta_\Phi, \Theta_\Psi} \text{max}_{\Theta_\mathcal{D}} \mathcal{L}_p(\Theta_\Phi, \Theta_\Psi) - \lambda L_D(\Theta_D, \Theta_\Phi, \Theta_\Psi). \quad (6)$$

The three players (D,Φ and Ψ) are alternatively updated as:

$$\begin{aligned} \Theta_D \quad & \leftarrow \Theta_D + \eta_D \tfrac{\partial \mathcal{L}}{\partial \Theta_D}, \\ \Theta_\Phi \quad & \leftarrow \Theta_\Phi - \eta_\Phi \tfrac{\partial \mathcal{L}}{\partial \Theta_\Phi}, \\ \Theta_\Psi \quad & \leftarrow \Theta_\Psi - \eta_\Psi \tfrac{\partial \mathcal{L}}{\partial \Theta_\Psi}, \end{aligned} \quad (7)$$

where η_Φ, η_Ψ, and η_D are the learning rates. To prevent the model collapsing, we can update the Θ_D several times before continuing to update the other parameters.

Nearest Neighbor Matching

After the model training, the estimated outcome of the i-th record/individual with the ω-th treatment, denoted as $\hat{Y}_i^\omega$, can be obtained by nearest neighbor matching: $\hat{Y}_i^\omega = Y_\nu^F$, with $\nu = \arg\min_{\nu \in U_\omega} \text{dist}(z_i, z_\nu)$, where Y_ν^F is the observed outcome of the ν-th record, U_ω is the group with the ω-th treatment, $dist(\cdot, \cdot)$ is the Euclidean distance, and z_i (z_ν) is the representation of the i-th (ν-th) record. Then the ITE, ATE and ATT can be obtained by Eqn. (1) accordingly.

4 Experiment

In this section, we conduct experiments on both semi-synthetic and real world datasets to evaluate the following: 1) our proposed method can work well on both the textual covariates and the non-textual covariates, and 2) the conditional treatment discriminator in our proposed method improves the performance on treatment effect estimation.

4.1 Experiment Settings

Baselines

We compare our proposed CTAM method with the following widely-adopted nearest neighbor matching (NNM) based methods: Mahalanobis distance matching (**MDM**) [Rubin, 1979], propensity score matching with logistic regression (**PSM**) [Rosenbaum and Rubin, 1983], dimensionality reduction by random linear projections (**DR-RLP**) [Li *et al.*, 2016], Hilbert-Schmidt independence criterion based nearest neighbor matching (**HSIC-NNM**) [Chang and Dy, 2017], and structural topic model based matching (**STM**) [Mozer *et al.*, 2018]. Besides the NNM based methods, we also compare the proposed method with the following representative baselines: linear regression with ℓ_1 regularization (**LASSO**) [Tibshirani, 1996]; Bayesian additive regression trees (**BART**) [Chipman *et al.*, 2010]; and causal forest regression (**CF**) [Wager and Athey, 2017]. Among the above baselines, LASSO and HSIC-NNM take the nearly

	PEHE	ϵ_{ATE}	ϵ_{ATT}
LASSO	$3.47 \pm 1.26^*$	$0.88 \pm 0.33^*$	$1.75 \pm 0.73^*$
BART	$4.10 \pm 1.27^*$	$1.98 \pm 1.36^*$	$2.87 \pm 1.44^*$
CF	2.69 ± 0.98	$1.88 \pm 0.53^*$	$2.20 \pm 0.80^*$
MDM	$3.29 \pm 0.80^*$	$0.64 \pm 0.61^*$	$0.74 \pm 0.57^*$
PSM	$2.69 \pm 0.33*$	$0.21 \pm 0.14^*$	$0.15 \pm 0.11^*$
DR-RLP	$4.03 \pm 1.44^*$	$0.85 \pm 0.57^*$	0.10 ± 0.05
STM	2.29 ± 0.41	$0.20 \pm 0.15^*$	$\mathbf{0.07 \pm 0.04}$
NNM-HSIC	$4.25 \pm 1.21^*$	$0.83 \pm 0.71^*$	0.12 ± 0.11
CTAM (Ours)	$\mathbf{2.06 \pm 0.03}$	$\mathbf{0.08 \pm 0.01}$	0.09 ± 0.01

Table 1: Results on News Dataset. Each entry is the mean and standard deviation of evaluation metric over 50 repeated realizations. The star marker (*) indicates that the results of that baseline and CTAM have statistically significant difference.

instrumental variables into consideration by covariates reweighting/selection. For comparison fairness, the baselines share the same text processing procedure with CTAM.

Parameter Setting

The parameters of baselines are set as suggested by the original papers, and the hyper-parameter search of CTAM follows the scheme in [Shalit *et al.*, 2017].

Evaluation Metrics

The following evaluation metrics are adopted to compare the proposed methods with the baselines:
(1) PEHE: precision in estimation of heterogeneous effect [Hill, 2011], which is defined as: $\text{PEHE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\text{ITE}_i - \hat{\text{ITE}}_i\right)^2}$. (2) $\mathcal{E}_{\text{ATE}}$: error of ATE estimation. $\mathcal{E}_{\text{ATE}}$ is defined as: $\mathcal{E}_{\text{ATE}} = |\text{ATE} - \hat{\text{ATE}}|$. (3) $\mathcal{E}_{\text{ATT}}$: error of ATT estimation. $\mathcal{E}_{\text{ATT}}$ is defined as: $\mathcal{E}_{\text{ATT}} = |\text{ATT} - \hat{\text{ATT}}|$.

4.2 Experiment on News Dataset

Dataset

The News dataset is first introduced in [Johansson *et al.*, 2016], which studies the effect of viewing devices to the user experience. The text covariate T is represented by the term-document matrix, and the vocabulary size is 3,477. The treatments are different devices: $W_i = 1$ denotes the news in the i-th record is viewed in mobile and $W_i = 0$ denotes the desktop. The generations of treatment assignment and the outcome (reading experience) are the same as [Johansson *et al.*, 2016]. We generate $1,000$ samples for each realization and repeat the sampling procedure 50 times.

Results and Analysis

The results of our proposed method and the baseline approaches are shown in Table 1. Overall, our method has the best performance under PEHE and $\mathcal{E}_{\text{ATE}}$ measurements, and has competing performance compared with the best baseline STM under $\mathcal{E}_{\text{ATT}}$ measurement. This observation demonstrates that the conditional treatment discriminator can effectively filter out information related to the nearly instrumental variables and therefore reduce the bias of treatment effect estimation.

	PEHE	ϵ_{ATE}	ϵ_{ATT}
LASSO	$6.59 \pm 4.36^*$	$1.58 \pm 0.58^*$	$2.33 \pm 0.61*$
BART	$5.55 \pm 4.07^*$	$2.12 \pm 1.25^*$	0.22 ± 0.38
CF	$5.10 \pm 4.24^*$	$2.95 \pm 1.74*$	$2.75 \pm 1.62*$
MDM	$2.38 \pm 1.05^*$	0.12 ± 0.08	0.31 ± 0.31
PSM	$3.52 \pm 1.34^*$	$0.78 \pm 0.59*$	$4.08 \pm 1.17*$
DR-RLP	$3.07 \pm 1.65^*$	$0.16 \pm 0.13*$	$0.42 \pm 0.35*$
NNM-HSIC	1.71 ± 0.43	$0.14 \pm 0.11*$	0.17 ± 0.17
CTAM (Ours)	$\mathbf{1.64 \pm 0.04}$	$\mathbf{0.09 \pm 0.01}$	$\mathbf{0.14 \pm 0.02}$

Table 2: Results on IHDP Dataset. Each entry is the mean and standard deviation of evaluation metrics over 100 repeated realizations. The star marker (*) indicates the baselines over which CTAM has statistically significant improvement.

4.3 Experiment on IHDP Dataset

The proposed method is motivated by the existence of nearly instrumental variables which are commonly observed in text covariates, but it can also work well on numerical cases to filter out information that may amplify the bias in the treatment effect estimation and improve the result. To validate this claim, we evaluate the proposed method on the widely used benchmark dataset IHDP.

Dataset

This dataset is from the Infant Health and Development Program [Brooks-Gunn *et al.*, 1992] targeting low-birth-weight, premature infants. The treated group is provided with both intensive high-quality child care and specialist home visits [Hill, 2011]. The dataset also provides 25 covariates related to the children and their mothers. The outcome is the infants' cognitive test score, which is simulated by the setting "A" of NPCI package[2]. Additionally, a biased subset of the treated group is removed to create the correlation between the treatment assignment and the covariates. In all, there are 747 records in the dataset, with 139 in the treated group and 608 in the control group.

Results and Analysis

Table 2 shows the results of our proposed method as well as the baselines. Similar to the performance on News dataset, our proposed method consistently performs best under all three evaluation metrics, which demonstrates the effectiveness of imposing the conditional adversarial-treatment learning.

4.4 Experiment on CFPB Dataset

Dataset

The dataset comes from Consumer Financial Protection Bureau (CFPB)[3]. The CFPB solicits complaints from consumers across a variety of financial products and then addresses those complaints [Egami *et al.*, 2018]. In this dataset, our goal is to analyze the effect of the financial company's public response to the consumers' disputation. In this dataset, we select the following responses as three treatments:

[2]https://github.com/vdorie/npci

[3]https://www.consumerfinance.gov/data-research/consumer-complaints/

Type I	*Type II*	*Type III*
this account is old, and should have been removed. it appears to have been transferred several times, they mention continue to update as un paid on my credit file, what for the rest of my life. another threat involving my way of living. amazing how I 'm following the law and getting robbed and bullied, controlled.	Mortgage company failed to maintain proper accounting and balance for escrow and attempted to increase mortgage by XXXX $400.00 per month on two occasions. After multiple requests for an explanation of increase, company is yet to provide consistent accounting.	Debt was paid multiple times due to collection agency stating they never received payment and refuse to remove it from credit bureaus. National credit systems owes me money!
This collection agency, FOCUS RM, has placed a debt on my credit report that is not mine. I researched the company. Apparently they are scam artists that place false debts that do not belong, and they have done this to XXXX of people. The debt on my credit report, which has ruined my credit is not mine. I never even used XXXX. Please shut this scam company down!	Credit Collection Services has attempted to collect a debt from me for XXXX and I never established a contract with this company, in addition this company has defamed my name and character and reported false information to both my XXXX and XXXX credit report.	I am receiving letters from credit collection services for XXXX bill that I do not owe, I have never had any accounts with XXXX.
All information matches except for the last four of SSN (XXXX). The SSN they have on file (XXXX) is not mine but the name, address, phone # is correct. I have requested that they remove me from there dialer and cease contact with me until they can prove this debt belongs to me.	This debt is in legal dispute at Superior Court of XXXX XXXX. Also, it was subsequently assigned to another Collection Agency And both are showing on my credit report. I have made repeated requests to have this removed since it 's been assigned to another agency but nothing has been done as of XXXX XXXX, 2015.	This company was found on my credit report and I did not receive any 30 day notice, any disclosure to my physical address. They do n't have a court order, verbal express consent or written consent with my authorized signature to report on my credit report.

Table 3: Matched complaints. Each row contains three matched complaints.

Type I. The company believes it acted appropriately as authorized by contract or law;
Type II. The company believes the complaint is the result of a misunderstanding;
Type III. The company disputes the facts presented in the complaint.
The treatment represents the company's overall attitude in dealing with the complaint. Different attitudes would affect customers' degree of satisfaction. The outcome is whether the consumer disputed. In total, there are $15,187$ consumers who receive the *Type I* response, $1,512$ consumers who receive the *Type II* response, $1,468$ consumers who receive the *Type III* response. The covariates are consumers' textual complaints, and each complaint is represented as the average of the word embeddings generated by Glove [Pennington *et al.*, 2014].

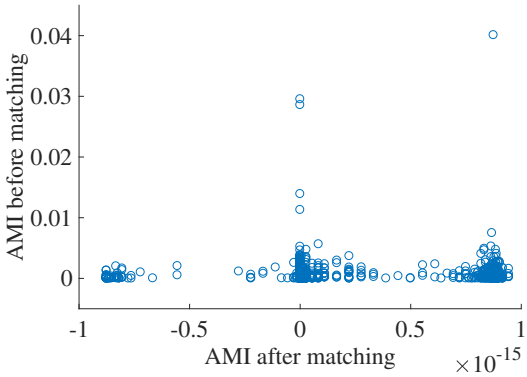

Figure 3: Adjusted Mutual Information

Results and Analysis
Since the ground truth information on treatment effect is not available in the CFPB Dataset, we examine the matching quality of the paired complaints. Figure 3 shows the adjust mutual information (AMI) [Vinh *et al.*, 2010] between the word occurrence and treatment assignment. Each point in the figure represents one word, the value of y-axis represents the AMI before matching, and the value of x-axis represents the AMI after matching. It can be observed from the figure that after matching, almost every word's AMI is close to zero, which indicates that our proposed method can effectively remove the information predictive to the treatment assignment.

Case Study
To better measure the matching quality, we show the matched complaints in Table 3 to validate whether the matched pairs are semantically similar or not. The length of compliants varies, and due to the space limit, we only list three short complaint pairs in the table. It is observed from Table 3 that the matched pairs are similar in our human sense: in the first row of Table 3, the three matched complaints are all about unexpected debt charging or account balancing; in the second row, the claims are all about wrongly placed debt or bill, and in the third row, the complaints are all about an unauthorized account.

5 Related Work

Due to its ability of estimating the changes in the outcome after making the intervention on the treatment selection, treatment effect estimation is prevalent across various domains [Guo *et al.*, 2018; Imbens and Rubin, 2015].

Among the existing treatment effect estimation methods, few of them focus on the case where the dataset contains

textual covariates. In [Egami *et al.*, 2018], complimentary to our problem setting, the authors consider the case when the treatment or outcome is text, and present a framework to estimate the effect of text or the effect on text. In [Wood-Doughty *et al.*, 2018], the text classifiers are integrated into the causal graph to recover the underlying distribution of other covariates, but the textual covariates are not involved in treatment effects. In [Mozer *et al.*, 2018; Roberts *et al.*, 2018], the authors adopt the topic model to represent the textual covariates and apply matching approaches to estimate the treatment effect. However, the topic model based matching approaches highly rely on the accuracy of the topic model and also ignore the nearly instrumental variables contained in the topic representation. Compared with existing approaches, our proposed CTAM method is flexible to any text representation and is capable of filtering out the information related to nearly instrumental variables, which effectively decreases the estimation bias.

In terms of the nearly instrumental variable related methods, various methods have been developed to handle the numerical covariates, such as covariate re-weighting [Kuang *et al.*, 2017a; Chang and Dy, 2017; Diamond and Sekhon, 2013] and feature selection [Kuang *et al.*, 2017b; Tibshirani, 1996; Rassen *et al.*, 2011]. In our work, instead of restricting the filtering process on certain covariates, CTAM learns the latent representation of all covariates with the information related to the nearly instrumental variables removed.

6 Conclusions

Text data are almost everywhere in real life. However, few of the existing treatment effect estimation methods deal with the textual covariates. The major challenge of treatment effect estimation with textual covariates is how to effectively filter out the nearly instrumental variables which degrade the performance of the treatment effect estimation. Existing covariate selection or covariate re-weighting technologies cannot handle the challenge due to the abundant information contained in the textual covariates. Therefore, we propose a novel conditional treatment-adversarial training based matching method, named as CTAM. By imposing the conditional adversarial training procedure, CTAM can learn the latent representation of all covariates with information related to near instrumental variables discarded. Through the experiments on three datasets, it is demonstrated that our proposed CTAM method can improve the treatment effect estimation with textual covariates.

Acknowledgements

This work was supported in part by the US National Science Foundation under grants NSF-IIS 1747614 and IIS-1514204. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

[Brooks-Gunn *et al.*, 1992] Jeanne Brooks-Gunn, Fong-ruey Liaw, and Pamela Kato Klebanov. Effects of early intervention on cognitive function of low birth weight preterm infants. *The Journal of pediatrics*, 120(3):350–359, 1992.

[Chang and Dy, 2017] Yale Chang and Jennifer G. Dy. Informative subspace learning for counterfactual inference. In *Proc. of AAAI'17*, pages 1770–1776, 2017.

[Chipman *et al.*, 2010] Hugh A Chipman, Edward I George, Robert E McCulloch, et al. Bart: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1):266–298, 2010.

[D'Amour *et al.*, 2017] Alexander D'Amour, Peng Ding, Avi Feller, Lihua Lei, and Jasjeet Sekhon. Overlap in observational studies with high-dimensional covariates. *arXiv preprint arXiv:1711.02582*, 2017.

[Diamond and Sekhon, 2013] Alexis Diamond and Jasjeet S Sekhon. Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies. *Review of Economics and Statistics*, 95(3):932–945, 2013.

[Egami *et al.*, 2018] Naoki Egami, Christian J Fong, Justin Grimmer, Margaret E Roberts, and Brandon M Stewart. How to make causal inferences using texts. *arXiv preprint arXiv:1802.02163*, 2018.

[Guo *et al.*, 2018] Ruocheng Guo, Lu Cheng, Jundong Li, P Richard Hahn, and Huan Liu. A survey of learning causality with data: Problems and methods. *arXiv preprint arXiv:1809.09337*, 2018.

[Harris, 1954] Zellig S Harris. Distributional structure. *Word*, 10(2-3):146–162, 1954.

[Hill, 2011] Jennifer L Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011.

[Imbens and Rubin, 2015] Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

[Johansson *et al.*, 2016] Fredrik D. Johansson, Uri Shalit, and David Sontag. Learning representations for counterfactual inference. In *Proc. of ICML'16*, pages 3020–3029, 2016.

[Kuang *et al.*, 2017a] Kun Kuang, Peng Cui, Bo Li, Meng Jiang, and Shiqiang Yang. Estimating treatment effect in the wild via differentiated confounder balancing. In *Proc. of KDD'17*, pages 265–274, 2017.

[Kuang *et al.*, 2017b] Kun Kuang, Peng Cui, Bo Li, Meng Jiang, Shiqiang Yang, and Fei Wang. Treatment effect estimation with data-driven variable decomposition. In *Proc. of AAAI'17*, 2017.

[Li *et al.*, 2016] Sheng Li, Nikos Vlassis, Jaya Kawale, and Yun Fu. Matching via dimensionality reduction for estimation of treatment effects in digital marketing campaigns. In *Proc. of IJCAI'16*, pages 3768–3774, 2016.

[Lopez *et al.*, 2017] Michael J Lopez, Roee Gutman, et al. Estimation of causal effects with multiple treatments: a review and new ideas. *Statistical Science*, 32(3):432–454, 2017.

[Manning *et al.*, 1999] Christopher D Manning, Christopher D Manning, and Hinrich Schütze. *Foundations of statistical natural language processing*. MIT press, 1999.

[Mikolov *et al.*, 2013] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[Mozer *et al.*, 2018] Reagan Mozer, Luke Miratrix, Aaron Russell Kaufman, and L Jason Anastasopoulos. Matching with text data: An experimental evaluation of methods for matching documents and of measuring match quality. *arXiv preprint arXiv:1801.00644*, 2018.

[Nair and Hinton, 2010] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proc. of ICML'10*, 2010.

[Pearl, 2012] Judea Pearl. On a class of bias-amplifying variables that endanger effect estimates. *arXiv preprint arXiv:1203.3503*, 2012.

[Pennington *et al.*, 2014] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Proc. of EMNLP'14*, 2014.

[Rassen *et al.*, 2011] Jeremy A Rassen, Robert J Glynn, M Alan Brookhart, and Sebastian Schneeweiss. Covariate selection in high-dimensional propensity score analyses of treatment effects in small samples. *American journal of epidemiology*, 173(12):1404–1413, 2011.

[Roberts *et al.*, 2018] Margaret E Roberts, Brandon M Stewart, and Richard A Nielsen. Adjusting for confounding with text matching. 2018.

[Rosenbaum and Rubin, 1983] Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

[Rubin, 1974] Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

[Rubin, 1979] Donald B Rubin. Using multivariate matched sampling and regression adjustment to control bias in observational studies. *Journal of the American Statistical Association*, 74(366a):318–328, 1979.

[Shalit *et al.*, 2017] Uri Shalit, Fredrik D. Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *Proc. of ICML'17*, 2017.

[Sparck Jones, 1972] Karen Sparck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21, 1972.

[Splawa-Neyman *et al.*, 1990] Jerzy Splawa-Neyman, Dorota M Dabrowska, and TP Speed. On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, pages 465–472, 1990.

[Tibshirani, 1996] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[Vinh *et al.*, 2010] Nguyen Xuan Vinh, Julien Epps, and James Bailey. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research*, 11(Oct):2837–2854, 2010.

[Wager and Athey, 2017] Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, (just-accepted), 2017.

[Wood-Doughty *et al.*, 2018] Zach Wood-Doughty, Ilya Shpitser, and Mark Dredze. Challenges of using text classifiers for causal inference. In *Proc. of ENMLP'18*, 2018.

[Wooldridge, 2016] Jeffrey M Wooldridge. Should instrumental variables be used as matching variables? *Research in Economics*, 70(2):232–237, 2016.

[Zhao *et al.*, 2017] Mingmin Zhao, Shichao Yue, Dina Katabi, Tommi S Jaakkola, and Matt T Bianchi. Learning sleep stages from radio signals: A conditional adversarial architecture. In *Proc. of the ICML'17*, 2017.