Using Text Embeddings for Causal Inference

Victor Veitch *[1], Dhanya Sridhar *[2], and David M. Blei[1,2]

[1]*Department of Statistics, Columbia University*
[2]*Department of Computer Science, Columbia University*

Abstract

We address causal inference with text documents. For example, does adding a theorem to a paper affect its chance of acceptance? Does reporting the gender of a forum post author affect the popularity of the post? We estimate these effects from observational data, where they may be confounded by features of the text such as the subject or writing quality. Although the text suffices for causal adjustment, it is prohibitively high-dimensional. The challenge is to find a low-dimensional text representation that can be used in causal inference. A key insight is that causal adjustment requires only the aspects of text that are predictive of both the treatment and outcome. Our proposed method adapts deep language models to learn low-dimensional embeddings from text that predict these values well; these embeddings suffice for causal adjustment. We establish theoretical properties of this method. We study it empirically on semi-simulated and real data on paper acceptance and forum post popularity. Code is available at github.com/blei-lab/causal-text-embeddings.

1 Introduction

We develop a method for causal inference from observed text documents. We consider a binary treatment, an outcome of interest, and a document of text. We assume that the text carries sufficient information to identify the causal effect; it is either an observed confounder or an observed mediator.

Example 1.1. Consider a corpus of scientific papers submitted to a conference. Some have theorems; others do not. We want to infer the causal effect of including a theorem on paper acceptance. The effect is confounded by the subject of the paper—more technical topics demand theorems, but may have different rates of acceptance. The data does not explicitly list the subject, but it does include each paper's abstract. We want to use the text to adjust for the subject and estimate the causal effect.

Example 1.2. Consider comments from Reddit.com, an online forum. Each post has a popularity score and the author of the post may (optionally) list their gender. We want to know the direct effect of a 'male' label on the score of the post. However, the author's gender may affect the text of the post, e.g., through tone, style, or topic choices, which also affects its score. Again, we want to use the text to accurately estimate the causal effect.

In these two examples, we assume that the text carries sufficient information to identify the causal effect. In theory, we can use classical methods of causal inference to adjust for the text of the document. But in practice we have finite data and the text is high dimensional, prohibiting efficient and accurate causal inference. The challenge is to reduce the text to a low-dimensional representation that both suffices for causal identification and that allows effective estimation with finite data.

[*] Equal contribution.

Our strategy is to draw on text embedding methods to reduce the dimension of the text [e.g., Mik+13b; Mik+13a; Dev+18; Pet+18]. Informally, a text embedding method distills the text of each document to a real-valued vector, and these embeddings can be used as features for prediction problems. Black-box embedding methods are state-of-the-art for a range of natural language understanding tasks [Dev+18; Pet+18]. Here, we will adapt embedding methods in the service of causal inference.

The key insight is that to adjust for variables in causal inference, it suffices to use only the information relevant to the prediction of the treatment and outcome. Thus we harness modern embedding methods—BERT [Dev+18], in particular—to extract the information from the text required for this prediction problem . The learned embeddings capture information sufficient for causal identification and provide the necessary ingredients for various causal estimators.

Contribution. The main contribution of this paper is a method for adapting off-the-shelf text embedding methods to estimate treatment effects. We show that the method is theoretically sound, demonstrate its utility on semi-synthetic data, and apply it to real datasets for estimating causal effects of the properties of papers on acceptance and gender label on popularity.

2 Related work.

This paper connects to several areas of related work.

The first area is causal inference for text. Roberts et al. [Rob+18] also discuss how to estimate effects of treatments applied to text documents. They rely (in part) on topic modeling to reduce the dimension of the text. This strategy is reasonable if the learned topics reflect the confounding aspects of the text. In contrast, we replace the assumption that the topics capture confounding with the assumption that an embedding method can effectively extract predictive information. We compare to a topic-model based approach in section 5.

In other work, Egami et al. [Ega+18] reduce raw text to interpretable outcomes; Wood-Doughty et al. [WD+18] estimate treatment effects when confounders are observed, but missing or noisy treatments are inferred from text. In contrast, we are concerned with text as the confounder.

A second area of related work addresses causal inference with unobserved confounding when there is an observed proxy for the confounder [KM99; Pea12; KP14; Mia+18; Kal+18]. This work usually assumes that the observed proxy variables are noisy realizations of the unobserved confounder, and then derives conditions under which causal identification is possible. One view of our problem is that each unit has a latent attribute (e.g., topic) such that observing it would suffice for causal identification, and the text is a proxy for this attribute. Unlike the proxy variable approach, however, we assume the text fully captures confounding. Our interest is in methods for finite-sample estimation rather than infinite-data identification.

Louizos et al. [Lou+17] also work with proxy variables, and consider the estimation problem. They fit a variational autoencoder using observed data and assume that it exactly recovers the true data generating distribution (including the latent confounder). We require weaker assumptions than the full recovery of the data generating distribution.

Work on causal inference with hidden confounding and many treatments is in the same vein [WB18; RP18; D'A19]. The idea is to use the treatments to infer the latent confounders. In contrast, we assume that the text suffices to adjust for confounding .

Finally, Veitch et al. [Vei+19] also use the reduction of causal estimation to prediction. In their case, to address unobserved confounding in the presence of network data.

3 Background

We begin by fixing notation and recalling some ideas from the estimation of causal effects. Each statistical unit is a document represented as a tuple $O_i = (Y_i, T_i, \mathbf{W}_i)$, where Y_i is the outcome, T_i is the treatment, and $\mathbf{W}_i$ is the sequence of words. The observed dataset consists of n observations drawn independently and identically at random from some distribution, $O_i \sim P$.

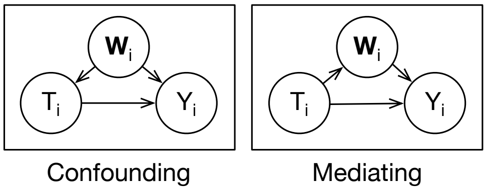

Figure 1: Models for the ATE (left) and NDE (right).

We review estimation of the average treatment effect and the natural direct effect. For both, we assume that the words are sufficient for adjustment.

Average treatment effect. The average treatment effect (ATE) is defined as

$$\psi = \mathbb{E}[Y \mid \mathrm{do}(T = 1)] - \mathbb{E}[Y \mid \mathrm{do}(T = 0)].$$

The use of Pearl's do notation indicates that the effect of interest is causal: what happens if we *intervene* by adding a theorem to a paper? We assume that the words $\mathbf{W}_i$ carry sufficient information to adjust for confounding (common causes) between T_i and Y_i. Figure 1 on the left depicts this assumption. We define $Z_i = f(\mathbf{W}_i)$ to be the part of $\mathbf{W}_i$ which blocks all 'backdoor paths' between Y_i and T_i. The causal effect is then identifiable from observational data as:

$$\psi = \mathbb{E}[\mathbb{E}[Y \mid Z, T = 1] - \mathbb{E}[Y \mid Z, T = 0]]. \tag{3.1}$$

Our task is to estimate the ATE ψ from a finite data sample. Define $Q(t,z) = \mathbb{E}[Y \mid t, z]$ to be the conditional expected outcome and $\hat{Q}$ to be an estimate for Q. Following 3.1, a natural estimator is:

$$\hat{\psi}^{Q} = \frac{1}{n} \sum_{i} \big[\hat{Q}(1, z_i) - \hat{Q}(0, z_i)\big]. \tag{3.2}$$

That is, ψ is estimated by a two-stage procedure: First produce an estimate for $\hat{Q}$ through a predictive model; then plug $\hat{Q}$ into a pre-determined statistic to compute the estimate of the ATE.

The estimator (3.2) is not the only possible choice. In principle, it is possible to do better by using estimators that also incorporate estimates $\hat{g}$ of the propensity scores $g(z) = \mathrm{P}(T = 1 \mid z)$ [e.g., Rob00; LR11; Rob+94; Che+17]. The general approach is a two-stage procedure. First fit a model for propensity scores and conditional outcomes; then plug the fitted model into a downstream estimator. What is important is that these estimators depend on z_i only through $\hat{g}(z_i)$ and $\hat{Q}(t, z_i)$.

Natural direct effect. The direct effect is the expected change in outcome if we apply the treatment while holding fixed any mediating variables that are affected by the treatment and that affect the outcome. Figure 1 on the right depicts the text as mediator of the treatment and outcome. For the estimation of the direct effect, we take $Z = f(\mathbf{W})$ to be the parts of $\mathbf{W}_i$ that mediate T and Y. The natural direct effect of treatment β is average difference in outcome induced by giving each unit the treatment, if the distribution of Z had been as though each unit received treatment. That is,

$$\beta = \mathbb{E}_{\mathrm{P}(Z \mid T=1)}[\mathbb{E}[Y \mid Z, \mathrm{do}(T = 1)] - \mathbb{E}[Y \mid Z, \mathrm{do}(T = 0)]].$$

In the gender example, this is the expected difference in score between a post labeled as written by a man versus labeled as written by a woman, where the expectation is taken over the distribution of posts written by men.

Under minimal conditions, this quantity may be estimated from observational data [Pea14]. The natural estimator is [LR11, Ch. 8]

$$\hat{\beta}^{\text{plugin}} = \frac{1}{n}\sum_i \big[\hat{Q}(1,z_i) - \hat{Q}(0,z_i)\big]\,\hat{g}(z_i) / \left(\frac{1}{n}\sum\nolimits_i t_i\right).$$

As with the ATE, there are also more sophisticated estimators [e.g., LR11, Ch. 8]. Again, all such estimators rely on Z only through the estimated conditional outcomes and propensity scores.

4 Causal text embeddings

We first focus on estimation of the average treatment effect. Following the previous section, we want to produce estimates of the propensity score $g(z_i)$ and the conditional expected outcome $Q(t_i, z_i)$. We assume that some property $z_i = f(\mathbf{w}_i)$ of the text suffices for identification. The obstacle motivating this paper is that we do not directly observe the confounding features z_i. Instead, we must work with the raw text.

A simple approach is to abandon z_i altogether and learn models for the propensities and conditional outcomes directly from the words $\mathbf{w}_i$. Since $\mathbf{w}_i$ contains all information about z_i, the direct adjustment will also render the causal effect identifiable. Indeed, in an infinite-data setting this would be a sound approach. However, the dimensionality of the problem is prohibitive.

We require a reduction of the words $\mathbf{w}_i$ to a feature z_i that both contains sufficient information to render the causal effect identifiable, and that will allow us to effectively learn the propensity scores and conditional outcomes with a finite data sample. A key insight follows from [RR83, Thm. 3]. Recall $Q(t,z) = \mathbb{E}[Y \mid t, z]$ and $g(z) = \mathrm{P}(T = 1 \mid z)$.

Theorem 4.1. *Suppose $\lambda(\mathbf{w})$ is some function of the words such that at least one of the following is $\lambda(\mathbf{W})$-measurable:*

1. $(Q(1,\mathbf{W}), Q(1,\mathbf{W}))$,
2. $g(\mathbf{W})$,
3. $g((Q(1,\mathbf{W}), Q(1,\mathbf{W})))$ *or* $(Q(1, g(\mathbf{W})), Q(1, g(\mathbf{W})))$.

If adjusting for $\mathbf{W}$ suffices to render the average treatment effect identifiable then adjusting for only $\lambda(\mathbf{W})$ also suffices. That is, $\psi = \mathbb{E}[\mathbb{E}[Y \mid \lambda(\mathbf{W}), T = 1] - \mathbb{E}[Y \mid \lambda(\mathbf{W}), T = 0]]$.

In words: the random variable $\lambda(\mathbf{W})$ carries the information about $\mathbf{W}$ relevant to the prediction of both the propensity score and the conditional expected outcome. While $\lambda(\mathbf{W})$ will typically throw away much of the information in the words, Theorem 4.1 says that adjusting for it suffices to estimate causal effects. Item 3 says that this holds even if we throw away information relevant to Y, so long as this information is not also relevant to T (and vice versa). The utility of Theorem 4.1 is that if we can find features of $\mathbf{w}$ that suffice for the *prediction* problem, then adjusting for these features also suffices for the causal *estimation* problem.

Our strategy is to use the words of each document to produce an embedding vector $\lambda(\mathbf{w})$ that captures the confounding aspects of the text. These embeddings are satisfactory if we can use them to estimate the propensities and conditional outcomes required by the downstream effect estimator.

We will use embedding-based prediction models from the natural language processing literature. For our purposes, these models may viewed as black-boxes that take in words $\mathbf{w}_i$ and produce a tuple $(\lambda_i, \tilde{Q}(t_i, \lambda_i), \tilde{g}(\lambda_i))$, which contains an embedding λ_i and estimates of g and Q that use that embedding. The idea is that such models provide an effective black-box tool for both distilling the words into the information relevant to prediction problems, and for solving those prediction problems.

Finally, to estimate the average treatment effect, we follow the general strategy of section 3. First, we fit the embedding-based prediction model to produce estimated embeddings $\hat{\lambda}_i$, propensity scores $\tilde{g}(\hat{\lambda}_i)$ and conditional outcomes $\tilde{Q}(t_i, \hat{\lambda}_i)$. We then plug these values into a downstream estimator. We will see an explicit example below.

Validity. The next result gives conditions for this procedure to be valid.

Theorem 4.2. *Let $\eta(z) = (\mathbb{E}[Y \mid T=0, z], \mathbb{E}[Y \mid T=1, z], \mathrm{P}[T=1 \mid z])$ be the conditional outcomes and propensities given z. Suppose that $\hat{\psi}(\{(t_i, y_i, z_i)\}; \eta) = \frac{1}{n}\sum_i \phi(t_i, y_i, \eta(z_i)) + o_p(1)$ is some consistent estimator for the average treatment effect ψ. Further suppose that there is some function λ of the words such that*

1. *(identification) λ satisfies the condition of Theorem 4.1.*
2. *(consistency) $\|\eta(\lambda(\mathbf{W}_i)) - \tilde{\eta}(\hat{\lambda}_i)\|_{P,2} \to 0$ as $n \to \infty$, where $\tilde{\eta}$ is the estimated conditional outcome and propensity model.*
3. *(well-behaved estimator) $\|\nabla_\eta \phi(t, y, \eta)\|_2 \leq C$ for some constant $C \in \mathbb{R}_+$,*

then, $\tilde{\psi}(\{(t_i, y_i, \hat{\lambda}_i)\}; \tilde{\eta}) \xrightarrow{p} \psi$.

Remark 4.3. The requirement that the estimator $\hat{\psi}$ behaves asymptotically as a sample mean is not an important restriction; most commonly used estimators have this property [Ken16]. The third condition is a technical requirement on the estimator. In the cases we consider, it suffices that the range of Y and Q are bounded and that g is bounded away from 0 and 1. This later requirement is the common 'overlap' condition, and is anyway required for the estimation of the causal effects.

Proof. By Theorem 4.1 and assumption 1, $\hat{\psi}(\{(t_i, y_i, \lambda(\mathbf{w}_i))\}; \eta) \xrightarrow{p} \psi$.

For brevity, we write $\lambda_i = \lambda(\mathbf{w}_i)$. By Taylor's theorem,

$$\frac{1}{n}\sum_i \phi(t_i, y_i, \tilde{\eta}(\hat{\lambda}_i)) = \frac{1}{n}\sum_i \phi(t_i, y_i, \eta(\lambda_i)) + \frac{1}{n}\sum_i \nabla_\eta \phi(t_i, y_i, \eta_i^*)(\tilde{\eta}(\hat{\lambda}_i) - \eta(\lambda_i)),$$

for some $\{\eta_i^*\}$. By continuous mapping, it suffices to show that the second term goes to 0 in probability. By Cauchy-Schwarz and assumption 3,

$$\frac{1}{n}\sum_i \nabla_\eta \phi(t_i, y_i, \eta_i^*)(\tilde{\eta}(\hat{\lambda}_i) - \eta(\lambda_i)) \leq C\sqrt{\frac{1}{n}\sum_i \|\tilde{\eta}(\hat{\lambda}_i) - \eta(\lambda_i)\|_2^2}.$$

By Markov's inequality, $\mathrm{P}(\frac{1}{n}\sum_i \|\tilde{\eta}(\hat{\lambda}_i) - \eta(\lambda_i)\|_2^2 > \varepsilon) \leq \|\eta(\lambda_i) - \tilde{\eta}(\hat{\lambda}_i)\|_{P,2}^2 / \varepsilon$, for all $\varepsilon > 0$. The result follows by assumption 2. □

As with all causal inference, the validity of the procedure relies on uncheckable assumptions that the practitioner must assess on a case-by-case basis. Particularly, we require that:

1. (properties z of) the document text renders the effect identifiable,
2. the embedding method extracts text information relevant to the prediction of both t and y,
3. the conditional outcome and propensity score models are consistent.

Only the second assumption is non-standard. In practice, we use the best possible embedding method and take the strong performance on (predictive) natural language tasks in many contexts as evidence that the method effectively extracts information relevant to prediction tasks. Implicitly, we are assuming that features that are useful for language understanding tasks are also useful for eliminating confounding. This is reasonable in settings where we expect the confounding to be aspects such as topic, writing quality, or sentiment. Informally, assumption 2 is satisfied if we use a good natural-language model, so we satisfy it by using the best available model.

Causal BERT. We modify BERT, a state-of-the-art language model [Dev+18]. Each input to BERT is the document text, a sequence of word-piece tokens $\mathbf{w}_i = (w_{i1}, \dots, w_{il})$. The model is tasked with producing three kinds of outputs: 1) document-level embeddings, 2) a map from the embeddings to treatment probability, 3) a map from the embeddings to expected outcomes for the treated and untreated.

The model assigns an embedding ξ_w to each word-piece w. It then produces a document-level embedding for document text $\mathbf{w}_i$ as $\lambda_i = f((\xi_{w_{i1}}, \dots, \xi_{w_{il}}), \gamma^{\mathrm{U}})$ for a particular function f. The embeddings and global parameter γ^{U} are trained by minimizing an unsupervised objective, denoted as $L_{\mathrm{U}}(\mathbf{w}_i; \xi, \gamma^{\mathrm{U}})$. Informally, random word-piece tokens are censored from each document and the model is tasked with predicting their identities.[1]

Following Devlin et al. [Dev+18], we use a fine-tuning approach to solve the prediction problem. We add a logit-linear layer mapping $\lambda_i \to \tilde{g}(\lambda_i; \gamma^g)$ and a 2-hidden layer neural net for each of $\lambda_i \to \tilde{Q}(0, \lambda_i; \gamma^{Q_0})$ and $\lambda_i \to \tilde{Q}(1, \lambda_i; \gamma^{Q_1})$. We learn the parameters for the embedding model and the prediction model jointly. Intuitively, this adapts the embeddings to be useful for the downstream prediction task, i.e., for causal inference.

We write γ for the full collection of global parameters. The final model is trained as:

$$\hat{\lambda}_i = f((\hat{\xi}_{n,w_{i1}}, \dots, \hat{\xi}_{n,w_{il}}), \hat{\gamma}^{\mathrm{U}})$$
$$\hat{\xi}, \hat{\gamma} = \operatorname*{argmin}_{\xi,\gamma} \frac{1}{n} \sum_i L(\mathbf{w}_i; \xi, \gamma),$$

where the objective is designed to predict both the treatment and outcome. It is

$$L(\mathbf{w}_i; \xi, \gamma) = \big(y_i - \tilde{Q}(t_i, \lambda_i; \gamma)\big)^2 + \mathsf{CrossEnt}\big(t_i, \tilde{g}(\lambda_i; \gamma)\big) + L_{\mathrm{U}}(\mathbf{w}_i; \xi, \gamma).$$

Effect estimation. Computing causal effect estimates simply requires plugging in the propensity scores and expected outcomes that the trained model predicts on the held-out units. For example, using the plug-in estimator (3.2),

$$\hat{\psi}^Q := \frac{1}{n} \sum_i \tilde{Q}(1, \hat{\lambda}_{n,i}; \hat{\gamma}_n^Q) - \tilde{Q}(0, \hat{\lambda}_{n,i}; \hat{\gamma}_n^Q). \tag{4.1}$$

The same procedure applies to other estimators as well.

Natural direct effect. We now discuss the analogous development for the natural direct effect. In this setting, the text serve as mediators between the treatment and the outcome. We are interested in understanding the causal effect of the treatment that does *not* go through the text.

The key result is the analogue of Theorem 4.1. Namely, suppose λ is some function of the words such that $\lambda(\mathbf{W})$ carries all information relevant to both the prediction of the treatment and outcome. Then the natural direct effect is equal to

$$\beta = \mathbb{E}_{\mathrm{P}(\lambda(\mathbf{W})|T=1)}[\mathbb{E}[Y \mid \lambda(\mathbf{W}), \mathrm{do}(T=1)] - \mathbb{E}[Y \mid \lambda(\mathbf{W}), \mathrm{do}(T=0)]]. \tag{4.2}$$

[1] BERT also considers a ‘next sentence’ prediction task, which we do not use.

That is, adjusting for $\lambda(\mathbf{W})$ suffices to adjust for any mediating effect in the words. This result is essentially by definition: any mediator must be predictive of both the treatment and outcome, so it suffices to adjust only for the parts of $\mathbf{w}$ that are predictive of both treatment and outcome.

The remaining development is identical to the average treatment effect case. We estimate embeddings, propensities, and conditional expected outcomes using Causal BERT, and then plug these estimates into a downstream direct effect estimator. For example,

$$\hat{\beta}^{\text{plugin}} = \frac{1}{n}\sum_i \left[\tilde{Q}(1,\hat{\lambda}_i) - \tilde{Q}(0,\hat{\lambda}_i)\right]\tilde{g}(\hat{\lambda}_i)/(\frac{1}{n}\sum_i t_i). \tag{4.3}$$

The proof of validity is the same as Theorem 4.2.

5 Experiments

We now empirically study the quality of Causal BERT embeddings for causal estimation. The questions of interest are: 1) do the learned embeddings identify causal effects in realistic simulations? 2) what happens in the presence of unobserved confounding exogenous to the text? Additionally, we apply the proposed method to the two motivating examples in the introduction. We estimate causal effects on paper acceptance and post popularity on Reddit.com.[2]

We find: 1) The method is able to effectively adjust for confounding. And, 2) it is robust to exogenous confounding. Our application suggests that much of the apparent effect of the treatments we study is attributable to confounding in the text.

5.1 Setup

PeerRead. PeerRead is a corpus of computer-science papers [Kan+18]. We consider a subset of the corpus consisting of papers posted to the arXiv under `cs.cl`, `cs.lg`, or `cs.ai` between 2007 and 2017 inclusive. The data only includes papers which are not cross listed with any non-cs categories and are within a month of the submission deadline for a target conferences. The conferences are: ACL, EMNLP, NAACL, EACL, TACL, NeurIPS, ICML, ICLR and AAAI. A paper is marked as accepted if it appeared in one of the target venues. Otherwise, the paper is marked as rejected. The dataset includes 11,778 papers, of which 2,891 are accepted.

For each paper, we consider the text of abstract, the accept/reject decision, and two attributes:

1. `buzzy`: the title contains any of 'deep', 'neural', 'embed', or 'adversarial net'.
2. `theorem`: the word 'Theorem' appears in the paper.

These attributes can be predicted from the abstract text.

Reddit. Reddit is an online forum divided into topic-specific subforums called 'subreddits'. We consider three subreddits: `keto`, `okcupid`, and `childfree`. In these subreddits, we identify users whose username flair includes a gender label (usually 'M' or 'F'). We collect all top-level comments from these users in 2018. We use each comment's text and score, the number of likes minus dislikes from other users. The dataset includes 90k comments in the selected subreddits. We consider the direct effect of the labeled gender on posts' scores.

[2] Software and data at github.com/blei-lab/causal-text-embeddings.

Table 1: Embedding adjustment recovers the NDE on Reddit. This persists even with high confounding and high noise. Table entries are estimated NDE. Columns are labeled by confounding level. Low, Med., and High correspond to $\beta_1 = 1.0, 10.0$ and 100.0.

Noise:	$\sigma = 1.0$			$\sigma = 4.0$		
Confounding:	Low	Med.	High	Low	Med.	High
Ground truth	1.00	1.00	1.00	1.00	1.00	1.00
Unadjusted	1.03	1.24	3.48	0.99	1.22	3.51
Words $\hat{\beta}^{\text{plugin}}$	1.01	1.17	2.69	1.04	1.16	2.63
Words $\hat{\beta}^{\text{TMLE}}$	1.02	1.18	2.71	1.04	1.17	2.65
LDA $\hat{\beta}^{\text{plugin}}$	1.01	1.20	2.95	1.02	1.19	2.91
LDA $\hat{\beta}^{\text{TMLE}}$	1.01	1.20	2.96	1.02	1.19	2.91
$\hat{\beta}^{\text{plugin}}$	0.96	1.05	1.24	0.83	0.63	1.31
$\hat{\beta}^{\text{TMLE}}$	0.98	1.05	1.58	0.95	1.00	1.51

Table 2: Embedding adjustment recovers the ATE on PeerRead. This persists even with high confounding. Table entries are estimated ATE. Columns are labeled by confounding level. Low, Med., and High correspond to $\beta_1 = 1.0, 5.0$ and 25.0.

Confounding:	Low	Med.	High
Ground truth	0.06	0.05	0.03
Unadjusted	0.08	0.15	0.16
Words $\hat{\psi}^{Q}$	0.07	0.13	0.15
Words $\hat{\psi}^{\text{TMLE}}$	0.07	0.13	0.15
LDA $\hat{\psi}^{Q}$	0.06	0.06	0.06
LDA $\hat{\psi}^{\text{TMLE}}$	0.06	0.06	0.06
$\hat{\psi}^{Q}$	0.07	0.06	−0.01
$\hat{\psi}^{\text{TMLE}}$	0.06	0.07	0.04

Estimator. We use Causal BERT, explained in section 4. We truncate PeerRead abstracts to 250 word-piece tokens, and Reddit posts to 128 word-piece tokens. We begin with a BERT model pre-trained on a general English language corpus. We further pre-train a BERT model on each dataset, running training on the unsupervised objective until convergence. In all cases, we use a logit-linear layer to predict treatment from embeddings, and a 2 hidden layer neural network for the expected outcome predictor.

For each experiment, we consider two downstream estimators: The simple estimators eqs. (4.1) and (4.3), and 'one-step' TMLE estimators [vG16]. The latter are more sophisticated estimators that combine estimated conditional outcomes and propensities to achieve asymptotic robustness and efficiency properties. For all estimators, we exclude units that have a predicted propensity score greater than 0.97 or less than 0.03.

5.2 Results

Estimator Evaluation Empirical evaluation of causal estimation procedures requires semi-synthetic data because ground truth causal effects are usually not available for real-world data. For such evaluations to be compelling, the semi-synthetic model must be reflective of real-world data. This is challenging for text data: there are no realistic generative models of text, so it is not possible to generate a confounder and then generate the text, treatment, and outcome on the basis of this confounder.

To circumvent this, we use real metadata—subreddit and title buzziness—as the confounders $\tilde{z}$ for the simulation. We simulate only the outcomes, using the treatment and the confounder. We compute the true propensity score $\pi(\tilde{z})$ as the proportion of units with $t_i = 1$ in each strata of $\tilde{z}$. Then, Y_i is simulated from the model:

$$Y_i = t_i + \beta_1(\pi(\tilde{z}_i) - 0.5) + \varepsilon_i \qquad \varepsilon_i \sim N(0, \sigma).$$

Or, for binary outcomes,

$$Y_i \sim \text{Bernoulli}(\text{sigmoid}(0.25 t_i + b_1(\pi(\tilde{z}_i) - 0.2)))$$

The parameter b_1 controls the level of confounding; e.g., the bias of the unadjusted difference $\mathbb{E}[Y|T=1] - \mathbb{E}[Y|T=0]$ increases with b_1. For PeerRead, we report estimates of the ATE for binary simulated outcomes. For Reddit, we compute the NDE for simulated real-valued outcomes.

Additionally, we compare against two baselines. The first is a two-stage procedure that uses LDA to estimate document-topic proportions $\hat{z}$ and linear/logistic regression for $\hat{Q}(\hat{z})$ and $\hat{g}(\hat{z})$. The second fits linear/logistic regression for the expected outcomes and treatments using word counts directly without dimensionality reduction.

Results are summarized in tables 1 and 2. Compared to the unadjusted estimate, all methods for adjustment reduce confounding. However, causal BERT does substantially better for moderate to high confounding. This is even in a simulation setting favorable to LDA (the true confounding is topic, and has a simple relation to outcome). The benefits of dimensionality reduction on text are clear in PeerRead, where adjustment based on LDA is much better than using the words alone.

The effect of exogeneity. We assume that the text carries all information about the confounding (or mediation) necessary to identify the causal effect. In many situations, this assumption may not be fully realistic. For example, in the simulations just discussed, it may not be possible to exactly recover the confounding from the text. We study the effect of violating this assumption by simulating both treatment and outcome from a confounder that consists of a part that can be fully inferred from the text and part that is wholly exogenous.

The challenge is finding a realistic confounder that can be exactly inferred from the text. Our approach is to (i) train BERT to predict the actual treatment of interest, producing propensity scores $\hat{g}_i$ for each i, and (ii) use $\hat{g}_i$ as the inferrable part of the confounding. Precisely, we simulate propensity scores as $\operatorname{logit} g_{\text{sim}} = (1-p) \operatorname{logit} \hat{g}_i + p\xi_i$, with $\xi_i \overset{\text{iid}}{\sim} \mathrm{N}(0,1)$. The outcome is simulated as above. When $p = 0$, the simulation is fully-inferrable and closely matches real data. Increasing p allows us to study the effect of exogeneity; see Figure 2. As expected, the adjustment quality decays. Remarkably, the adjustment improves the naive estimate at all levels of exogeneity—the method is robust to violations of the theoretical assumptions.

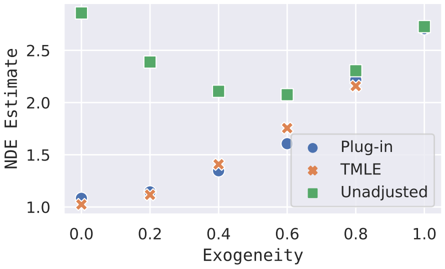

Figure 2: The method improves the unadjusted estimator even with exogeneous mediations. Plot shows estimates of NDE from simulated data based on Reddit. Ground truth is 1.

Application We apply causal BERT to estimate the treatment effect of `buzzy` and `theorem`, and the effect of gender on log-score in each subreddit; see tables 3 and 4. Although unadjusted estimates suggest strong effects, our results show this is in large part explainable

Table 3: Embedding adjustment reduces estimated treatment effects in PeerRead. Entries are estimated treatment effect and 10-fold bootstrap standard deviation.

	buzzy	theorem
Unadjusted	0.08 ± 0.01	0.21 ± 0.01
$\hat{\psi}^{Q}$	0.01 ± 0.03	0.03 ± 0.03
$\hat{\psi}^{\mathrm{TMLE}}$	0.06 ± 0.04	0.10 ± 0.03

Table 4: Embedding adjustment reduces estimated direct effects in Reddit. Entries are estimated treatment effect and 10-fold bootstrap standard deviation.

	okcupid	childfree	keto
Unadjusted	-0.18 ± 0.01	-0.19 ± 0.01	-0.00 ± 0.00
$\hat{\beta}^{\mathrm{plugin}}$	-0.10 ± 0.04	-0.10 ± 0.04	-0.03 ± 0.02
$\hat{\beta}^{\mathrm{TMLE}}$	-0.15 ± 0.05	-0.16 ± 0.05	-0.01 ± 0.00

by confounding or mediating. On PeerRead, the TMLE estimate $\hat{\psi}^{\mathrm{TMLE}}$ suggests a positive effect from including a theorem on paper acceptance, but the Q-only estimator does not. On Reddit, both estimates suggest a positive effect from labeling a post as female on its score in okcupid and childfree.

References

[Che+17] V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins. "Double/debiased machine learning for treatment and structural parameters". In: *The Econometrics Journal* (2017).

[D'A19] A. D'Amour. "On multi-cause approaches to causal inference with unobserved counfounding: two cautionary failure cases and a promising alternative". In: *International conference on artificial intelligence and statistics*. 2019.

[Dev+18] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. "BERT: pre-training of deep bidirectional transformers for language understanding". In: *arXiv e-prints*, arXiv:1810.04805 (2018).

[Ega+18] N. Egami, C. J. Fong, J. Grimmer, M. E. Roberts, and B. M. Stewart. "How to make causal inferences using texts". In: *arXiv preprint arXiv:1802.02163* (2018).

[Kal+18] N. Kallus, X. Mao, and M. Udell. "Causal inference with noisy and missing covariates via matrix factorization". In: *Advances in neural information processing systems*. 2018.

[Kan+18] D. Kang, W. Ammar, B. Dalvi, M. van Zuylen, S. Kohlmeier, E. Hovy, and R. Schwartz. "A dataset of peer reviews (peerread): collection, insights and nlp applications". In: *arXiv e-prints*, arXiv:1804.09635 (2018).

[Ken16] E. H. Kennedy. "Semiparametric theory and empirical processes in causal inference". In: *Statistical Causal Inferences and their Applications in Public Health Research*. 2016.

[KM99] M. Kuroki and M. Miyakawa. "Identifiability criteria for causal effects of joint interventions". In: *Journal of the Japan Statistical Society* 2 (1999).

[KP14] M. Kuroki and J. Pearl. "Measurement bias and effect restoration in causal inference". In: *Biometrika* 2 (2014).

[LR11] M. van der Laan and S. Rose. *Targeted Learning: Causal Inference for Observational and Experimental Data*. 2011.

[Lou+17] C. Louizos, U. Shalit, J. M. Mooij, D. Sontag, R. Zemel, and M. Welling. "Causal effect inference with deep latent-variable models". In: *Advances in neural information processing systems*. 2017.

[Mia+18] W. Miao, Z. Geng, and E. J. Tchetgen Tchetgen. "Identifying causal effects with proxy variables of an unmeasured confounder". In: *Biometrika* 4 (2018).

[Mik+13a] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. "Distributed representations of words and phrases and their compositionality". In: *Advances in neural information processing systems*. 2013.

[Mik+13b] T. Mikolov, K. Chen, G. Corrado, and J. Dean. "Efficient estimation of word representations in vector space". In: *arXiv preprint arXiv:1301.3781* (2013).

[Pea12] J. Pearl. "On measurement bias in causal inference". In: *arXiv e-prints,* arXiv:1203.3504 (2012).

[Pea14] J. Pearl. "Interpretation and identification of causal mediation". In: *Psychological methods* (2014).

[Pet+18] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. "Deep contextualized word representations". In: *arXiv e-prints*, arXiv:1802.05365 (2018).

[RP18] R. Ranganath and A. Perotte. "Multiple causal inference with latent confounding". In: *arXiv preprint arXiv:1805.08273* (2018).

[Rob+18] M. E. Roberts, B. M. Stewart, and R. A. Nielsen. *Adjusting for confounding with text matching*. 2018.

[Rob00] J. M. Robins. "Robust estimation in sequentially ignorable missing data and causal inference models". In: *Bayesian Statistical Science* (2000).

[Rob+94] J. M. Robins, A. Rotnitzky, and L. P. Zhao. "Estimation of regression coefficients when some regressors are not always observed". In: *Journal of the American Statistical Association* 427 (1994).

[RR83] P. R. Rosenbaum and D. B. Rubin. "The central role of the propensity score in observational studies for causal effects". In: *Biometrika* 1 (1983).

[vG16] M. van der Laan and S. Gruber. "One-step targeted minimum loss-based estimation based on universal least favorable one-dimensional submodels". In: *Int. J. Biostat.* 1 (2016).

[Vei+19] V. Veitch, Y. Wang, and D. M. Blei. "Using embeddings to correct for unobserved confounding in networks". In: *arXiv e-prints*, arXiv:1902.04114 (2019).

[WB18] Y. Wang and D. M. Blei. "The Blessings of Multiple Causes". In: *arXiv e-prints*, arXiv:1805.06826 (2018).

[WD+18] Z. Wood-Doughty, I. Shpitser, and M. Dredze. "Challenges of using text classifiers for causal inference". In: *Empirical methods in natural language processing*. 2018.