

Causal Inference from Text: Unveiling Interactions between Variables

Yuxiang Zhou[♡] and Yulan He^{♡◇}

[♡]King’s College London, [◇]The Alan Turing Institute
{yuxiang.zhou, yulan.he}@kcl.ac.uk

Abstract

Adjusting for latent covariates is crucial for estimating causal effects from observational textual data. Most existing methods only account for *confounding* covariates that affect both *treatment* and *outcome*, potentially leading to biased causal effects. This bias arises from insufficient consideration of *non-confounding* covariates, which are relevant only to either the treatment or the outcome. In this work, we aim to mitigate the bias by unveiling interactions between different variables to disentangle the non-confounding covariates when estimating causal effects from text. The disentangling process ensures covariates only contribute to their respective objectives, enabling independence between variables. Additionally, we impose a constraint to balance representations from the treatment group and control group to alleviate selection bias. We conduct experiments on two different treatment factors under various scenarios, and the proposed model significantly outperforms recent strong baselines. Furthermore, our thorough analysis on earnings call transcripts demonstrates that our model can effectively disentangle the variables, and further investigations into real-world scenarios provide guidance for investors to make informed decisions¹.

1 Introduction

Causal Inference (Holland, 1985; Pearl, 2000; Morgan and Winship, 2007; Imbens and Rubin, 2015; Hernan and Robins, 2020) aims to identify how the *treatment* variable affects the *outcome* variable. For example, to estimate the effect of “*political risk*” (treatment) faced by a company on its “*stock movement*” (outcome). Early research efforts (Abadie and Imbens, 2004; Bardone-Cone and Cass, 2006; Kurth et al., 2006; Murnane and Willett, 2010; Keele, 2015) focusing on conducting randomized control trials (RCTs) to estimate

¹Our code and data are released at <https://github.com/zyxnlp/DIVA>.

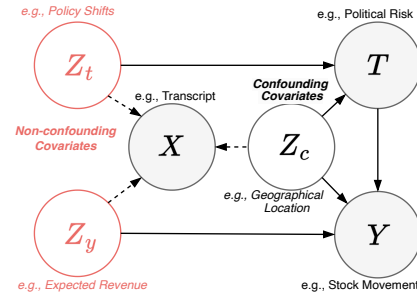


Figure 1: The causal diagram for our proposed model. Shaded nodes denote observed variables. Transparent nodes denote latent covariates derived from the transcripts, among which, nodes outlined in red represent non-confounding covariates that impact only either the treatment T or outcome O , whereas the node outlined in black denotes the confounding covariate that influences both T and O .

causal effects from structural numeric data have made significant progress. However, these methods requires extensive effort in treatment assignment mechanism (Halloran and Struchiner, 1995) and may suffer from ethical issues. Natural Language Processing (NLP) researchers are increasingly interested in estimating causal effects from observational unstructured text. Early literature (Choudhury et al., 2016; Olteanu et al., 2017; Pryzant et al., 2018) largely focuses on transforming texts into high-dimensional vectors using lexical features for confounding adjustment. Recent research primarily focuses on learning adequate representations through advanced NLP models. For example, Veitch et al. (2020) fine-tuned BERT (Devlin et al., 2019) to produce contextual text representations for efficient estimation of causal effects. Later, Pryzant et al. (2021) introduced strategies involving treatment enhancement and text adjustment to estimate the causal effects related to linguistic properties.

Despite their efficacy, such approaches operate under the assumption that text solely encompasses confounding covariates. This assumption raises

a potential issue due to the possible existence of unobserved non-confounding covariates that are pertinent exclusively to either the treatment or the outcome. The causal estimation may be biased if we fail to differentiate non-confounding covariates from confounding ones when learning an estimation function through effective modeling of variable interactions (Pearl, 2010; Wooldridge, 2016). As illustrated in Figure 1, if we aim to accurately estimate the causal effects of treatment T (e.g., Political Risk) on the outcome Y (e.g., Stock Movement), we intentionally omit the consideration of the impacts originating from Z_y (e.g., Expected Revenue). This mirrors our decision not to account for the influence of Z_c (e.g., Geographical Location) on Y , as such inclusion could obfuscate our ability to discern the true effects originating from T .

In this paper, we propose a framework named Disentangling Interaction of VArables (DIVA), specifically tailored for causal inference from text. We assume that the text carries sufficient information to identify the causal effects and consider the existence of non-confounding covariates. Drawing on the success of latent variable models for causal inference in literature (Louizos et al., 2017; Zhang et al., 2021), we use Variational Auto-Encoder (VAE) (Kingma and Welling, 2014) to infer confounding and non-confounding covariates. Additionally, we design a disentanglement module to ensure that covariates only contribute to their specific objectives, enabling independence between covariates. Furthermore, we propose to impose a constraint to balance representations from the treatment group and control group, which helps to mitigate selection bias.

Our contributions are summarized as follows:

- We propose the Disentangling Interaction of VArables (DIVA) approach, tailored to mitigate the bias issue in causal inference from text.
- Our model is able to effectively model interactions among diverse variables, ensuring that each variable primarily contributes to its specific objective and promotes maximal independence.
- Our experiments demonstrate state-of-the-art results in various scenarios. A detailed analysis shows that our model effectively disentangles different variables given inherently high-dimensional nature of text representation, pro-

viding valuable insights for estimating causal effects from text.

- To the best of our knowledge, we are pioneers in addressing biased issues arising from inadequate consideration of non-confounding covariates when estimating causal effects from text.

2 Related Work

Causal estimation with text data Early efforts in estimating causal effects from text focused on using lexical features for confounding adjustment (Choudhury et al., 2016; Choudhury and Kıcıman, 2017; Olteanu et al., 2017). Later studies investigating causal effects were devoted to effectively converting text into low-dimensional representations (Falavarjani et al., 2017; Pham and Shen, 2017; Pryzant et al., 2018; Weld et al., 2020; Cheng et al., 2021). Another line of work focused on using causal formalisms to make NLP methods more reliable (Wood-Doughty et al., 2018, 2021; Feder et al., 2021, 2022). Most recently, pre-trained language models such as BERT (Devlin et al., 2019) significantly benefited causal estimation. For example, Veitch et al. (2020) fine-tuned BERT using multi-task learning to produce contextual text representations for efficient estimation of causal effects. Later, Pryzant et al. (2021) introduced treatment-boosting and text-adjusting strategies to estimate the causal effects of linguistic properties. Our work differs from these works in three main aspects. First, we aim to mitigate the bias that arises from insufficient consideration of non-confounding covariates in causal inference. Second, we disentangle non-confounding covariates by encouraging independence among the variables, ensuring that each one contributes solely to its respective objective. Third, we introduce regularization to balance representations from the treatment group and control group, which helps to mitigate selection bias.

Causal inference with latent variable model

Latent variable models have demonstrated their effectiveness and gained significant popularity in causal inference (Fong and Grimmer, 2016; Sridhar and Getoor, 2019; Roberts et al., 2020). For example, Louizos et al. (2017) used Variational Auto-Encoder (VAE) (Kingma and Welling, 2014) to infer confounders from latent space to estimate the effect of job training on employment following the training. Rakesh et al. (2018) inferred the causation that leads to spillover effects between

pairs of units by incorporating VAE to learn the latent attributes as confounders. We follow the line of decomposing latent factors for causal inference (Hassanpour and Greiner, 2020; Wu et al., 2020; Vowels et al., 2020; Yang et al., 2021; Zhang et al., 2021). However, there are several key distinctions in our approach. Firstly, while previous studies attempted to disentangle variables for causal inference in structured numeric data, we specifically focus on estimating causal effects from textual data. The inherently high-dimensional nature of text features presents substantial challenges in disentangling various variables within the latent space, leading to biased causal estimations. Secondly, we tailor distinct constraints to effectively model interactions among diverse variables, ensuring that each variable primarily contributes to its specific objective and promotes maximal independence. Lastly, we optimize the maximum mean discrepancy loss to achieve a balanced representation of samples from both treatment and control groups.

NLP for earnings call transcripts Earnings call transcripts (Frankel et al., 1997; Bowen et al., 2001; Price et al., 2011) have gained much popularity in financial analysis using NLP tools. Early work by Wang and Hua (2014) formulated financial risk prediction as a text regression task and used handcrafted features to improve SVM performance. Later, researchers (Qin and Yang, 2019; Sawhney et al., 2020; Sang and Bao, 2022; Pataci et al., 2022; Shah et al., 2022; Yang et al., 2022) focused on stock prediction by employing sophisticated neural networks with financial pragmatic features. Another line of work focused on analyzing the content of earnings call transcripts (Sawhney et al., 2021; Alhamzeh et al., 2022). For example, Keith and Stent (2019) examined analysts’ decision-making behavior as it pertains to the language content of earnings calls. More in line with our work, Hassan et al. (2017) adapted linguistic tools to investigate the extent of political risk faced by firms over time and its correlation with stocks, hiring, and investment. In contrast with this prior work, our primary focus lies on estimating causal effects between financial interests, such as the impact of political risk on stocks, rather than measuring their correlations.

3 Preliminaries

Causal inference from text aims to estimate the causal effects based on observed textual data. Let

$\mathcal{D} = \{X_i, T_i, Y_i\}_{i=1}^N$ represent the N observational examples. Here, X_i is the observed textual data (e.g., earnings call transcript) for the i -th example (e.g., company), and $T_i \in \{0, 1\}$ is the binary treatment variable². $T_i = 1$ indicates that the i -th example belongs to the treatment group (e.g., a company faced high political risk). Conversely, $T_i = 0$ indicates that the i -th example belongs to the control group (e.g., a company faced low or no political risk). The causal effect τ_i for the i -th example is defined as the expected difference between its potential outcome Y_i (e.g., stock volatility) of the treatment and control groups, known as the Individual Treatment Effect (ITE):

$$\tau_i = Y_i(T_i = 1) - Y_i(T_i = 0) \quad (1)$$

One of the most challenging problems in estimating causal effects from observational data is the impossibility of simultaneously observing both potential outcomes $Y_i(T_i = 0)$ and $Y_i(T_i = 1)$ for a given example (Rubin, 1974; Holland, 1985). In other words, \mathcal{D} only includes the observed outcome Y_i for each example, but not the unobserved *counterfactual* outcome, which refers to the potential outcome for the i -th example in the alternative group. Nonetheless, it’s feasible to identify the Conditional Average Treatment Effect (CATE) and the Average Treatment Effect (ATE) from observational data under certain assumptions (Sława-Neyman et al., 1990; Rubin, 1974; Pearl, 2009):

Assumption 1 (Stable Unit Treatment Values Assumption (SUTVA)): The potential outcomes of one example are not influenced by the treatment assigned to other examples, and there are no varying forms or levels of the treatment that could result in different potential outcomes: $Y_i(t_1, \dots, t_i, \dots, t_n) = Y_i(t_i)$, and $Y(T = t_i) = Y_i(T_i)$.

Assumption 2 (Unconfoundedness): The potential outcomes are conditionally independent of the treatment given a set of observed covariates: $(Y(1), Y(0)) \perp\!\!\!\perp T$.

Assumption 3 (Positivity): Every individual has a non-zero probability of receiving treatment or control for all observed variables: $0 < P(T = 1|X = x) < 1$.

In line with the potential outcome framework outlined by Sława-Neyman et al. (1990) and Rubin (1974), and with the above assumptions, we

²We defer the scenarios involving multiple treatments for future exploration.

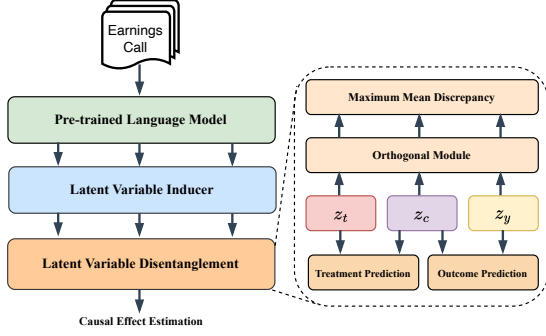


Figure 2: DIVA architecture.

can define the CATE as follows:

$$\mathbb{E}[\tau_i | X = x_i] = \mathbb{E}[Y_i(1) - Y_i(0) | X = x_i] \quad (2)$$

where $Y_i(1)$ and $Y_i(0)$ are the potential outcomes had the i -th individual received the treatment or control. X is the observed variable which is sufficient for causal estimation. The ATE can be written as:

$$\mathbb{E}[\tau_i] = \mathbb{E}_X[\tau_i | X = x_i] \quad (3)$$

Problem Definition Defining $Q(t, x) = \mathbb{E}[Y_i(t) | X = x]$ as the potential outcome of observing treatment $T = t$ for an example with $X = x$, the objective is to learn an estimation function $\hat{Q}(t, x)$ that can accurately predict both the observed outcome and counterfactual outcome from \mathcal{D} . Therefore, we can plug in \hat{Q} to estimate CATE:

$$\hat{\tau} = \frac{1}{n} \sum_{i=1}^n [\hat{Q}(1, x_i) - \hat{Q}(0, x_i)] \quad (4)$$

4 DIVA: Disentangling Interaction of Variables

In this section, we present the proposed Disentangling Interaction of Variables (DIVA) framework (Figure 2) for causal inference from textual earnings call transcripts. Although previous research (Veitch et al., 2020; Pryzant et al., 2021) has explored estimating causal effects from text, one of the core contributions of our work is that we disentangle various variables to effectively model the interactions among them. This in turn enables us to learn a more accurate estimation function \hat{Q} for predicting outcomes, thereby reducing the bias in the causal estimation.

Our proposed DIVA framework consists of a few steps. First, we extract the contextualized

text representation from the pre-trained language model. Following that, we employ a variational auto-encoder to determine the posterior distribution for various latent variables. Once this distribution is obtained, we use the variable disentanglement module to encourage independence among the variables, ensuring that each one contributes solely to its respective objective. Next, we utilize the disentangled variables to learn the \hat{Q} function via the outcome prediction task. Finally, we plug the trained \hat{Q} into a pre-determined statistic to estimate the ATE.

4.1 Text Encoder

Given a transcript $x = [w_1, \dots, w_n]$ that consists of n words, we adopt the pre-trained language (PLM) model FinBERT (Araci, 2019)³ to obtain the contextual representation \mathbf{h} for each transcript:

$$\mathbf{h} = \text{PLM}(x) \quad (5)$$

4.2 Latent Variable Inducer

Inspired by recent works (Louizos et al., 2017; Zhang et al., 2021), we use the VAE to induce latent variables. Given the contextualized representation \mathbf{h} . We compute the approximation variational posterior $q_\phi(\mathbf{z} | \mathbf{h})$ using the inference network $\Phi(\mathbf{h}; \phi)$:

$$\begin{aligned} \boldsymbol{\mu} &= \mathbf{W}_\mu \mathbf{h} + \mathbf{b}_\mu \\ \log \boldsymbol{\sigma}^2 &= \mathbf{W}_\sigma \mathbf{h} + \mathbf{b}_\sigma \\ \mathbf{z} &= \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{\epsilon} \end{aligned} \quad (6)$$

where \mathbf{W}_μ , \mathbf{W}_σ , \mathbf{b}_μ , and \mathbf{b}_σ are parameters for two MLPs. $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ define a multivariate Gaussian distribution with a diagonal covariance matrix, and $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I})$. Then, we sample from $q_\phi(\mathbf{z} | \mathbf{h}) \simeq \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2 \mathbf{I})$ to generate $\mathbf{z} \in \mathbb{R}^l$ as the latent representation, where l is the dimension of the representation. Under the assumption that a transcript contains not only the confounding covariates, which affects both treatment and outcome, but also the non-confounding covariates specific to either the treatment or the outcome, we use separate inference networks $\Phi_c(\mathbf{h}; \phi_c)$ for inferring confounding covariates z_c , and $\Phi_t(\mathbf{h}; \phi_t)$ and $\Phi_y(\mathbf{h}; \phi_y)$ for inferring non-confounding covariates z_t and z_y , respectively. We use a one-layer parameterized MLP $\Theta(\mathbf{h}; \theta) := p_\theta(\mathbf{h} | z_t, z_c, z_y)$

³We chose FinBERT due to its adaptability to text in finance domain. However, other PLMs could serve as suitable replacements.

as the decoder to reconstruct \mathbf{h} . The objective of the latent variable inducer is to maximize the evidence lower bound (ELBO):

$$\mathcal{L}_{vae} = \mathbb{E}_{\Phi_t, \Phi_c, \Phi_y} [\log \Theta(\mathbf{h}; \theta)] - \sum_k \text{KL}(\Phi_k || p(\mathbf{z}_k)) \quad (7)$$

where $k \in \{c, t, y\}$, and $p(\mathbf{z}_k)$ is the prior follows the Gaussian distribution $\mathcal{N}(0, \mathbf{I})$.

4.3 Latent Variable Disentanglement

Despite the successful application of decomposing variables in previous work (Zhang et al., 2021), unfortunately, the high-dimensional nature of text features presents significant obstacles in disentangling different variables in a latent space, leading to biased causal estimation. As will be shown in Section 5.2 (e.g., TEDVAE v.s. CEVAE), considering only non-confounding covariates, without the ability to effectively model interactions between different variables, fails to consistently achieve better performance in textual data.

To address this issue, we tailor various distinct constraints to effectively disentangle non-confounding covariates from confounding ones, ensuring that each variable primarily contributes to its specific objective and promotes maximal independence.

Specifically, we first minimize the Maximum Mean Discrepancy (MMD) (Gretton et al., 2012) loss to balance representations from the treatment group and the control group:

$$\mathcal{L}_{\text{mmd}} = \sum_{k \in \{c, t, y\}} \mathcal{M}(\mathbf{z}_k^{\text{treat}}, \mathbf{z}_k^{\text{contl}}) \quad (8)$$

where $\mathcal{M}(\cdot)$ denotes the maximum mean discrepancy metric. $\mathbf{z}_k^{\text{treat}}$ and $\mathbf{z}_k^{\text{contl}}$ are the representations in the treatment group and the control group, respectively. The nice property of this loss is that minimizing the loss essentially reduces the discrepancy between different groups, encouraging the satisfaction of the positivity assumption. Concurrently, it promotes the inference network to generalize from the factual to counterfactual domains, leading to better counterfactual inference (Johansson et al., 2016).

Next, we introduce an orthogonal loss to maximize the independence between \mathbf{z}_t , \mathbf{z}_c , and \mathbf{z}_y as much as possible:

$$\mathcal{L}_{\text{ort}} = \sum_{k, v} \text{Orth}(\mathbf{z}_k; \mathbf{z}_v) \quad (9)$$

where $k, v \in \{t, c, y; k \neq v\}$. $\text{Orth}(\mathbf{z}_k; \mathbf{z}_v) = \|\mathbf{z}_k \cdot \mathbf{z}_v^T - \mathbb{I}\|$, and \mathbb{I} is the identity matrix.

Intuitively, we expect that the prediction of the treatment label should primarily rely on \mathbf{z}_t and \mathbf{z}_c , rather than \mathbf{z}_y . To ensure this holds, we introduce the treatment loss:

$$\mathcal{L}_t = \log P(t|\mathbf{z}_y) - \log P(t|\mathbf{z}_t, \mathbf{z}_c) \quad (10)$$

where $t \in \{0, 1\}$ indicates whether the transcript belongs to the treatment group.

Similarly, we expect the prediction of outcome should primarily rely on \mathbf{z}_y and \mathbf{z}_c , and define the outcome loss:

$$\mathcal{L}_o = \mathcal{O}(y, \hat{Q}(t, \mathbf{z}_y, \mathbf{z}_c)) \quad (11)$$

where $y \in Y$ is the potential outcome. \mathcal{O} is an MSE loss for real-valued outcomes and a cross-entropy loss for the binary outcomes.

The overall objective function of the latent variable disentanglement module is formulated as:

$$\mathcal{L}_d = \mathcal{L}_{vae} + \alpha \mathcal{L}_t + \beta \mathcal{L}_o + \gamma \mathcal{L}_{\text{ort}} + \eta \mathcal{L}_{\text{mmd}} \quad (12)$$

where α, β, γ , and η are hyper-parameters.

4.4 Final Training Objective

Following Veitch et al. (2020) and Pryzant et al. (2021), we introduce a Masked Language Model (MLM) objective that predicts words that are randomly⁴ masked, in order to adapt text representation, making it more efficient for treatment and outcome prediction. Our final objective function is a multi-task learning objective:

$$\mathcal{L} = \mathcal{L}_d + \lambda \mathcal{L}_{\text{mlm}} \quad (13)$$

where λ is the coefficient that balances the contribution of each component in the training process.

5 Experiments

We conduct experiments on both semi-synthetic data and real-world application scenarios with two objectives: 1) to empirically evaluate the effectiveness of our proposed model, and 2) to investigate practical questions in the field of finance and gain insights from the application of our model to these real-world scenarios.

⁴Following Devlin et al. (2019), we masked 15% of the words in each transcript.

5.1 Experimental Setup

Baselines The baseline models selected for comparison can be broadly categorized into three groups: deep outcome regression models, latent variable models, and representation learning models. Deep outcome regression models include:

- **TARNet** Shalit et al. (2017) uses separate feed-forward networks to predict outcomes and counterfactuals.
- **CFRNet** Shalit et al. (2017) adds an integral probability metric (IPM) regularization term to TARNet to balance representation from different groups.
- **DragonNet** Shi et al. (2019) extends TARNet with an additional head adapts representation by modeling the propensity score.

The latent variable based models are:

- **CEVAE** Louizos et al. (2017) uses VAE to infer confounders from an unknown latent space to estimate causal effects.
- **TEDVAE** Zhang et al. (2021) extends CEVAE by decomposing latent factors into three sets: instrumental, confounding, and risk factors.

The representation learning models are:

- **CausalBert** Veitch et al. (2020) develops an approach to adjust for confounding features of text to estimate causal effects from observational data.
- **TextCause** Pryzant et al. (2021) introduces treatment-boosting and text-adjusting strategies to estimate causal effects of linguistic properties.

Whenever possible, we generate results for baselines using the officially released source code. In cases where the code of models is not available at the time of writing, we independently implement those models using the optimal hyper-parameter settings reported in the respective papers. For a fair comparison, we use FinBERT (Araci, 2019) to encode text for generating contextualized feature representations for all models.

Evaluation Metric We evaluate the results using the precision in estimation of heterogeneous effect (PEHE) (Hill, 2011), which reflects model’s individual-level estimation performance: $\sqrt{\text{PEHE}} = \sqrt{\frac{1}{N} \sum_{i=1}^N (\tau_i - \hat{\tau}_i)^2}$. We also report the error of ATE estimation $\delta\text{ATE} = |\tau - \frac{1}{N} \sum_{i=1}^N \hat{\tau}_i|$, which measure the model’s population-level estimation performance.

Setup Details In our experimental evaluations, each model is trained for 30 epochs with a linear warmup for the first 10% of the training steps. We employ AdamW (Loshchilov and Hutter, 2019) as the optimizer. We set the maximum learning rate at 5e-5 and use a batch size of 86. We select the optimal model weights based on either accuracy or the MSE loss of the \hat{Q} function on the development set⁵. We report the average results along with the mean absolute deviations across five runs with randomly initialized parameters.

5.2 Experiments on Synthetic Data

Dataset

Since ground truth causal effects ITE τ_i and ATE τ , are typically inaccessible in real-world scenarios, directly training a model for causal inference is impractical. Therefore, we follow Veitch et al. (2020) and Pryzant et al. (2021), using real text and metadata to generate semi-synthetic data to empirically evaluate our proposed model. We collect 115,880 transcripts from 1,438 companies across twelve different sectors, for earnings calls held between May 2001 and October 2019. Then, we construct different datasets for two distinct treatment variables - political risk (T_{pr}) and sentiment (T_s) - under two separate scenarios: stock volatility (Y_{vol}) and stock movement (Y_{mov}). To derive T_{pr} , we follow Hasan et al. (2017) to calculate the political risk score⁶ for each transcript. We then select the top 15,000 transcripts with the highest scores as the treatment group ($T_{pr} = 1$), indicating that the company faces high political risk. Conversely, we designate the bottom 15,000 transcripts with the lowest scores as the control group ($T_{pr} = 0$), suggesting these companies face lower or no political risk. To derive T_s , we follow Maia et al. (2018) and Araci (2019) to calculate the sentiment score⁷ for each transcripts. We select the top 15,000 transcripts with the highest scores as the treatment group ($T_s = 1$) and select the bottom 15,000 transcripts with the lowest scores as the control group ($T_s = 0$). Finally, we simulate the outcomes by using the treatment variable $T \in \{T_{pr}, T_s\}$ along with observed covariates, C_{size} and C_{sect} , which represent the size of the company in terms of the number of full-time employees and the industrial sector that the company operates. The real-valued stock volatility Y_{vol} can

⁵Please refer to Appendix B for detailed hyper-parameters.

⁶<https://github.com/mschwedeler/firmlevelrisk>

⁷<https://github.com/ProsusAI/finBERT>

Model	Political risk		Sentiment	
	$\sqrt{\text{PEHE}}$	δATE	$\sqrt{\text{PEHE}}$	δATE
<i>Stock Volatility</i>				
TARNet	1.196±0.019	0.480±0.049	1.213±0.019	0.491±0.049
DragonNet	1.173±0.022	0.450±0.048	1.190±0.021	0.463±0.046
CFRNet	1.169±0.020	0.445±0.045	1.185±0.020	0.455±0.044
CEVAE	1.197±0.025	0.477±0.050	1.211±0.024	0.491±0.044
TEDVAE	1.212±0.056	0.447±0.101	1.228±0.056	0.459±0.099
CausalBert	1.097±0.032	0.313±0.079	1.121±0.034	0.336±0.080
TextCause	<u>1.096±0.019</u>	<u>0.114±0.042</u>	<u>1.100±0.019</u>	<u>0.114±0.028</u>
DIVA	1.003±0.003 [†]	0.033±0.012 [†]	1.010±0.007 [†]	0.027±0.008 [†]
<i>Stock Movement</i>				
TARNet	0.497±0.001	0.086±0.009	0.497±0.001	0.089±0.010
DragonNet	0.497±0.003	0.084±0.025	0.497±0.004	0.088±0.026
CFRNet	0.497±0.003	0.083±0.025	0.497±0.004	0.086±0.025
CEVAE	0.499±0.004	0.076±0.022	0.499±0.004	0.079±0.020
TEDVAE	0.498±0.007	0.095±0.024	0.497±0.007	0.098±0.023
CausalBert	<u>0.496±0.002</u>	0.083±0.020	<u>0.496±0.001</u>	0.088±0.017
TextCause	<u>0.526±0.008</u>	<u>0.038±0.028</u>	<u>0.522±0.009</u>	<u>0.030±0.028</u>
DIVA	0.483±0.001 [†]	0.009±0.004 [†]	0.481±0.001 [†]	0.015±0.003 [†]

Table 1: The causal estimation results of different treatment factors on *stock volatility* and *stock movement*. Lower is better. The best results on each dataset are in bold. The second-best ones are underlined. The † marker indicates that the p -value is less than 0.05 compared to the second-best results. The parameter setting used is ($\alpha=1, \beta=1, \gamma=0.5, \epsilon=1$) for Equation (14) and (15).

be simulated as follows:

$$Y_{vol} = \alpha_v T + \beta_{v1}(\pi(C_{sect}) - \gamma_{v0}) + \beta_{v2}(\pi(C_{size}) - \gamma_{v1}) + \epsilon_v \quad (14)$$

The binary stock movement (*Up* or *Down*), Y_{mov} can be simulated as:

$$Y_{mov} \sim \text{Bernoulli}(\sigma(\alpha_m T + \beta_{m1}(\pi(C_{sect}) - \gamma_{m0}) + \beta_{m2}(\pi(C_{size}) - \gamma_{m1}) + \epsilon_m)) \quad (15)$$

where $\pi(C_{size})$ and $\pi(C_{sect})$ are propensity scores estimated from meta data. α_v and α_m control treatment strength. $\beta_{v1}, \beta_{v2}, \beta_{m1}$, and β_{m2} control confound strength. $\gamma_{v1}, \gamma_{v2}, \gamma_{m1}$, and γ_{m2} are offset. σ is the sigmoid function.

We split the dataset into the training, validation, and test sets in an 8:1:6 ratio and conduct experiments in a cross-validated manner, following Egami et al. (2018) and Pryzant et al. (2021). We conduct experiments for the two different treatment variable T_{pr} and T_s under the scenarios of stock volatility and stock movement, respectively. Detailed statistics of each scenario can be found in the Appendix.

Main Results

As shown in Table 1, DragonNet and CFRNet generally achieve better results than TARNet, suggest-

ing that additional constraints indeed benefit the outcome regression model in causal estimation. For example, DragonNet improves upon the TARNet by 0.03 in terms of δATE based on political risk in the stock volatility scenario. We also observe that Causalbert and TextCause generally achieve better results than the deep outcome regression models such as TARNet, DragonNet, and CFRNet, as well as latent variable models such as CEVAE and TEDVAE. This suggests that the inclusion of the masked language modeling task has a positive impact on causal inference from text. Our model consistently outperforms all compared baseline models across both evaluation metrics and under both scenarios. For instance, DIVA demonstrates a significant improvement (with $p < 0.05$) over the best-performing baseline TextCause and the CausalBert model.

Interestingly, we observe that TEDVAE struggles to consistently outperform CEVAE. In particular, TEDVAE achieves better results in terms of δATE but performs worse in terms of $\sqrt{\text{PEHE}}$ compared to CEVAE in the stock volatility scenario. We have contrary observations for TEDVAE and CEVAE under the stock movement setting. These results demonstrate that only considering non-confounding covariates, without the ability to

Latent Covariates			Political Risk		Sentiment	
z_t	z_c	z_y	$\sqrt{\text{PEHE}}$	δATE	$\sqrt{\text{PEHE}}$	δATE
<i>Stock Volatility</i>						
	✓		1.0107	0.0684	1.0179	0.0860
✓	✓		1.0062	0.0502	1.0110	0.0519
		✓	1.0054	0.0696	1.0140	0.0516
✓	✓	✓	1.0034	0.0332	1.0102	0.0273
<i>Stock Movement</i>						
	✓		0.4891	0.0407	0.4857	0.0358
✓	✓		0.4900	0.0440	0.4881	0.0579
		✓	0.4845	0.0390	0.4841	0.0409
✓	✓	✓	0.4831	0.0095	0.4814	0.0145

Table 2: Ablation study of our proposed model considering various latent covariates. Lower values are better.

effectively modeling interactions among various variables, falls short of consistently delivering satisfactory performance in textual data. However, our DIVA model consistently surpasses both CEVAE and TEDVAE by a substantial margin across all scenarios, which clearly demonstrates the importance of the constraints we introduced and underscores the effectiveness of our proposed model to estimate causal effects more accurately from text data.

Latent Covariates Analysis

To further investigate the influence of various covariates on model performance, we conduct an in-depth analysis of DIVA, focusing on the disentanglement of different covariates. As shown in Table 2, merely disentangling non-confounding covariates z_t or z_y from the confounding covariate z_c fails to consistently achieve better results compared to considering only z_c . Our model yields the best performance with the simultaneous disentanglement of z_t , z_c , and z_y . This results underscore the necessity of comprehensive covariate disentanglement, specifically, disentangling both non-confounding covariates z_t and z_y from the confounding covariate z_c , as opposed to a partial or singular focus.

Simulation Sensitivity Analysis

To evaluate the robustness of our proposed DIVA model, we have chosen to compare it with the two strongest baseline CausalBert and TextCause, under different simulation settings ($\alpha=1$, $\beta=10$, $\gamma=0.5$, $\epsilon=4$) in Equation (14) and (15). As shown in Table 3, our DIVA model consistently outperforms both CausalBert and TextCause across both evaluation metrics and under both scenarios. These results suggest that the superior performance of our model is not sensitive to changes in the simulation

Model	Political Risk		Sentiment	
	$\sqrt{\text{PEHE}}$	δATE	$\sqrt{\text{PEHE}}$	δATE
<i>Stock Volatility</i>				
CausalBert	4.0810	0.3858	4.1610	0.3904
TextCause	4.4103	0.2968	4.3927	0.2926
DIVA	4.0589	0.0534	4.1378	0.0592
<i>Stock Movement</i>				
CausalBert	0.4992	0.0400	0.4999	0.0531
TextCause	0.5306	0.0148	0.5337	0.0286
DIVA	0.4973	0.0072	0.4966	0.0103

Table 3: Causal estimation results (lower is better) under parameter settings ($\alpha=1$, $\beta=10$, $\gamma=0.5$, $\epsilon=4$) in Equation (14) and (15).

Model	Political risk		Sentiment	
	$\sqrt{\text{PEHE}}$	δATE	$\sqrt{\text{PEHE}}$	δATE
<i>Stock Volatility</i>				
DIVA	1.003	0.033	1.010	0.027
w/o mlm	1.004	0.062	1.011	0.068
w/o mmd	1.003	0.040	1.010	0.032
w/o ort	1.003	0.034	1.010	0.036
<i>Stock Movement</i>				
DIVA	0.483	0.009	0.481	0.015
w/o mlm	0.487	0.057	0.485	0.044
w/o mmd	0.486	0.036	0.485	0.035
w/o ort	0.486	0.036	0.485	0.030

Table 4: Ablation study of our proposed model on various scenarios. Lower values are better. ‘w/o mlm’ – without masked language modeling objective; ‘w/o mmd’ – without the Maximum Mean Discrepancy (MMD) objective; ‘w/o ort’ – without the orthogonal loss.

parameter setting, demonstrating the robustness of our DIVA model.

Ablation Study

We conducted experiments to examine the effectiveness of the major components of our proposed model. Table 4 shows the ablation results on stock volatility and stock movement scenarios. We observe that each component, namely \mathcal{L}_{mlm} , \mathcal{L}_{mmd} , and \mathcal{L}_{ort} contributes to the overall performance of the model. Specifically, with the removal of the \mathcal{L}_{mmd} , the performance of the full model drops considerably in terms of $\sqrt{\text{PEHE}}$. Similarly, removing \mathcal{L}_{mlm} results in a considerable drop in performance, measured by δATE . These observations demonstrate the vital role played by the \mathcal{L}_{mmd} regularization term, which encourages closer representations of individuals from different groups in the latent space. Incorporating the \mathcal{L}_{mlm} term benefits the es-

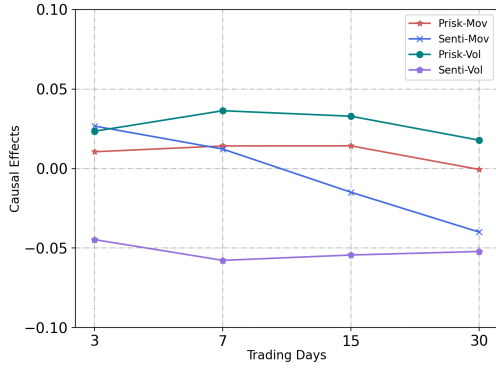


Figure 3: Causal effect of political risk and sentiment on the actual stock over trading days.

timination of CATE from text data. This phenomenon aligns with previous studies such as (Veitch et al., 2020; Pryzant et al., 2018).

5.3 Real World Scenario Application

To answer the questions of "How does political risk faced by a company affect its stock?" and "How does the sentiment conveyed in the earning call transcription of a company affect its stock?", we apply our proposed model to estimate the treatment effect of political risk and sentiment on actual stock volatility and stock movement.

Stock Volatility Following Qin and Yang (2019) and Kogan et al. (2009), we obtain the stock prices from Yahoo Finance⁸ by stock-market-scraper⁹ and calculate stock volatility as:

$$v_{[t-\mu,t]} = \ln \left(\sqrt{\frac{\sum_{i=0}^{\mu} (r_{t-i} - \bar{r})^2}{\mu}} \right) \quad (16)$$

where $r_t = \frac{P_t}{P_{t-1}} - 1$ is the stock return between the close of trading day $t - 1$ and day t , P_t is the dividend-adjusted closing stock price at t . \bar{r} is the mean of r_t over the period of day $t - \mu$ to day t . We choose different $\mu \in \{3, 7, 15, 30\}$ to evaluate the short-term and long-term causal effects.

Stock Movement Following (Medya et al., 2022), we define stock movement as:

$$m_t = \begin{cases} 1, & \text{if } r_t \geq \bar{v}_{[t-\mu,t]} \\ 0, & \text{else} \end{cases} \quad (17)$$

where $\bar{v}_{[t-\mu,t]}$ is the mean stock volatility over the period of day $t - \mu$ to day t .

⁸<https://finance.yahoo.com/>

⁹<https://github.com/gunjannandy/stock-market-scraper>

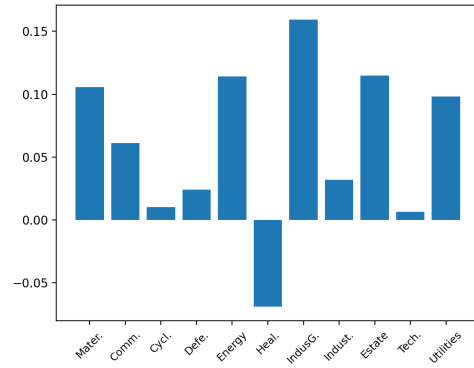


Figure 4: Causal effect of political risk on stock volatility over companies in different sectors.

Result As shown in Figure 3, we observe that the causal effects of political risk on stock increases in the short term (3 days) and begin to decline over time. Conversely, the causal effect of sentiment on stock movement decreases over time.

Analysis To further investigate the effect of political risks on the stock market for different types of companies, we examine the causal effect of political risk faced by companies in different sectors on their stock prices. Figure 4 shows that the stock volatility of companies in Industrials Goods, Real Estate, and Energy are most significantly affected by the political risk they faced, while companies in Consumer Cyclical and Technology are affected to the smallest extent. The political risks faced by the Healthcare companies have no effect on their stock volatility.

6 Conclusion

In this paper, we propose DIVA, a novel framework designed specifically for causal inference from text. We verify its effectiveness by estimating the causal effects of treatment factor (e.g., political risk or sentiment) on a company's stock (e.g., stock volatility or movement) from the earnings conference call transcripts. The experimental results demonstrate that our model can effectively disentangle representations with different functionalities from text features by imposing constraints and utilizing multi-task learning. Furthermore, our analysis of real-world applications highlights the causal relationship between political risks faced by a company and its stock prices, providing valuable insights for the finance and investment industry.

Limitations

Our work has a number of limitations. First, we constructed a balanced dataset in which the number of transcripts in the treatment group is equal to that in the control group. While this facilitated relatively easier causal estimation, it does not account for the selection bias that commonly exists in real-world scenarios. Consequently, causal estimation in such scenarios becomes more challenging. Second, we modeled the relation between treatment factors and stocks as a linear relation. However, in reality, this relationship is likely to be much more complex and nonlinear. A more precise modeling of this relationship would enhance the accuracy of our causal estimation.

Acknowledgements

We would like to thank the anonymous reviewers, our meta-reviewer, and senior area chairs for their constructive comments and support with our work. We would also like to thank Rui Qiao for helpful discussions and suggestions, and Zhanming Jie for feedback on the manuscript. This work was funded by the the UK Engineering and Physical Sciences Research Council (grant no. EP/T017112/1, EP/T017112/2, EP/V048597/1). YH is supported by a Turing AI Fellowship funded by the UK Research and Innovation (grant no. EP/V020579/1, EP/V020579/2).

References

- Alberto Abadie and Guido Imbens. 2004. Large sample properties of matching estimators for average treatment effects. *Econometrica*.
- Alaa Alhamzeh, Romain Fonck, Erwan Versmée, Elöd Egyed-Zsigmond, Harald Kosch, and Lionel Brunie. 2022. [It’s time to reason: Annotating argumentation structures in financial earnings calls: The FinArg dataset](#). In *Proceedings of Workshop on FinNLP*.
- Dogu Araci. 2019. [Finbert: Financial sentiment analysis with pre-trained language models](#). *ArXiv preprint*.
- Anna M. Bardone-Cone and Kamila M Cass. 2006. Investigating the impact of pro-anorexia websites: a pilot study. *European Eating Disorders Review*.
- Robert M. Bowen, Angela K. Davis, and Dawn Matsumoto. 2001. Do conference calls affect analysts’ forecasts? *Financial Accounting eJournal*.
- Lu Cheng, Ruocheng Guo, and Huan Liu. 2021. [Estimating causal effects of multi-aspect online reviews with multi-modal proxies](#). *Proceedings of WSDM*.
- Munmun De Choudhury, Emre Kiciman, Mark Dredze, Glen Coppersmith, and Mrinal Kumar. 2016. [Discovering shifts to suicidal ideation from mental health content in social media](#). In *Proceedings of the 2016 CHI CHFCS*.
- Munmun De Choudhury and Emre Kiciman. 2017. The language of social support in social media and its effect on suicidal ideation risk. *Proceedings of the AAAI*.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of NAACL*.
- Naoki Egami, Christian Fong, Justin Grimmer, Margaret E. Roberts, and Brandon M Stewart. 2018. How to make causal inferences using texts. *Science Advances*.
- Seyed Amin Mirlohi Falavarjani, Hawre Hosseini, Zeinab Noorian, and Ebrahim Bagheri. 2017. Estimating the effect of exercising on users’ online behavior. *Proceedings of AAAI*.
- Amir Feder, Katherine A. Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, Justin Grimmer, Roi Reichart, Margaret E. Roberts, Brandon M. Stewart, Victor Veitch, and Diyi Yang. 2022. [Causal inference in natural language processing: Estimation, prediction, interpretation and beyond](#). *Transactions of the Association for Computational Linguistics*.
- Amir Feder, Nadav Oved, Uri Shalit, and Roi Reichart. 2021. [CausaLM: Causal model explanation through counterfactual language models](#). *Computational Linguistics*, (2).
- Christian Fong and Justin Grimmer. 2016. [Discovery of treatments from text corpora](#). In *Proceedings of ACL*.
- Richard Frankel, Marilyn F. Johnson, and Douglas J. Skinner. 1997. An empirical examination of conference calls as a voluntary disclosure medium. *Financial Accounting*.
- Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alex Smola. 2012. A kernel two-sample test. *J. Mach. Learn. Res.*
- M. Elizabeth Halloran and Cláudio José Struchiner. 1995. Causal inference in infectious diseases. *Epidemiology*.
- Tarek Alexander Hassan, Stephan Hollander, Laurence van Lent, and Ahmed Tahoun. 2017. Firm-level political risk: Measurement and effects. *Political Economy: Government Expenditures & Related Policies eJournal*.
- Negar Hassanpour and Russell Greiner. 2020. [Learning disentangled representations for counterfactual regression](#). In *Proceedings of ICLR*.

- MA Hernan and J Robins. 2020. Causal inference: What if. boca raton: Chapman & hill/crc.
- Jennifer L. Hill. 2011. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*.
- Paul Holland. 1985. Statistics and causal inference. *Journal of the American Statistical Association*.
- Guido Imbens and Donald B. Rubin. 2015. Causal inference for statistics, social, and biomedical sciences: An introduction.
- Fredrik D. Johansson, Uri Shalit, and David A. Sontag. 2016. [Learning representations for counterfactual inference](#). In *Proceedings of ICML*.
- Luke J. Keele. 2015. The statistics of causal inference: A view from political methodology. *Political Analysis*.
- Katherine Keith and Amanda Stent. 2019. [Modeling financial analysts' decision making via the pragmatics and semantics of earnings calls](#). In *Proceedings of ACL*.
- Diederik P. Kingma and Max Welling. 2014. [Auto-encoding variational bayes](#). In *Proceedings of ICLR*.
- Shimon Kogan, Dimitry Levin, Bryan R. Routledge, Jacob S. Sagi, and Noah A. Smith. 2009. [Predicting risk from financial reports with regression](#). In *Proceedings of NAACL*.
- Tobias Kurth, Alexander M. Walker, Robert J. Glynn, K. Arnold Chan, J. Michael Gaziano, Klaus Berger, and James M. Robins. 2006. Results of multivariable logistic regression, propensity matching, propensity adjustment, and propensity-based weighting under conditions of nonuniform effect. *American journal of epidemiology*.
- Ilya Loshchilov and Frank Hutter. 2019. [Decoupled weight decay regularization](#). In *Proceedings of ICLR*.
- Christos Louizos, Uri Shalit, Joris M. Mooij, David A. Sontag, Richard S. Zemel, and Max Welling. 2017. [Causal effect inference with deep latent-variable models](#). In *Proceedings of NeurIPS*.
- Macedo Maia, Siegfried Handschuh, André Freitas, Brian Davis, Ross McDermott, Manel Zarrouk, and Alexandra Balahur. 2018. Www'18 open challenge: Financial opinion mining and question answering. *Companion Proceedings of The Web Conference*.
- Sourav Medya, Mohammad Rasoolinejad, Yang Yang, and Brian Uzzi. 2022. An exploratory study of stock price movements from earnings calls. *Companion Proceedings of the Web Conference 2022*.
- Stephen L. Morgan and Christopher Winship. 2007. Counterfactuals and causal inference: Methods and principles for social research.
- Richard J. Murnane and John B. Willett. 2010. Methods matter: Improving causal inference in educational and social science research.
- Alexandra Olteanu, Onur Varol, and Emre Kıcıman. 2017. Distilling the outcomes of personal experiences: A propensity-scored analysis of social media. *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*.
- Hilal Pataci, Yunyao Li, Yannis Katsis, Yada Zhu, and Lucian Popa. 2022. [Stock price volatility prediction: A case study with AutoML](#). In *Proceedings of Workshop on FinNLP*.
- Judea Pearl. 2000. Causality: Models, reasoning and inference.
- Judea Pearl. 2009. Causal inference in statistics: An overview. *Statistics Surveys*.
- Judea Pearl. 2010. [On a class of bias-amplifying variables that endanger effect estimates](#). In *Proceedings of UAI*.
- Thai Thuy Pham Pham and Yuanyuan Shen. 2017. A deep causal inference approach to measuring the effects of forming group loans in online non-profit microfinance platform. *arXiv: Machine Learning*.
- S. McKay Price, James S. Doran, David R. Peterson, and Barbara A. Bliss. 2011. Earnings conference calls and stock returns: The incremental informativeness of textual tone. *ERN: Uncertainty & Risk Modeling (Topic)*.
- Reid Pryzant, Dallas Card, Dan Jurafsky, Victor Veitch, and Dhanya Sridhar. 2021. [Causal effects of linguistic properties](#). In *Proceedings of NAACL*.
- Reid Pryzant, Kelly Shen, Dan Jurafsky, and Stefan Wagner. 2018. [Deconfounded lexicon induction for interpretable social science](#). In *Proceedings of NAACL*.
- Yu Qin and Yi Yang. 2019. [What you say and how you say it matters: Predicting stock volatility using verbal and vocal cues](#). In *Proceedings of ACL*.
- Vineeth Rakesh, Ruocheng Guo, Raha Moraffah, Nitin Agarwal, and Huan Liu. 2018. [Linked causal variational autoencoder for inferring paired spillover effects](#). In *Proceedings of CIKM*.
- Margaret E. Roberts, Brandon M Stewart, and Richard A. Nielsen. 2020. Adjusting for confounding with text matching. *American Journal of Political Science*.
- Donald B. Rubin. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*.
- Yunxin Sang and Yang Bao. 2022. [DialogueGAT: A graph attention network for financial risk prediction by modeling the dialogues in earnings conference calls](#). In *Findings of EMNLP*.

- Ramit Sawhney, Arshiya Aggarwal, and Rajiv Ratn Shah. 2021. [An empirical investigation of bias in the multimodal analysis of financial earnings calls](#). In *Proceedings of NAACL*.
- Ramit Sawhney, Piyush Khanna, Arshiya Aggarwal, Taru Jain, Puneet Mathur, and Rajiv Ratn Shah. 2020. [VolTAGE: Volatility forecasting via text audio fusion with graph convolution networks for earnings calls](#). In *Proceedings of EMNLP*.
- Raj Shah, Kunal Chawla, Dheeraj Eidnani, Agam Shah, Wendi Du, Sudheer Chava, Natraj Raman, Charese Smiley, Jiaao Chen, and Diyi Yang. 2022. [When FLUE meets FLANG: Benchmarks and large pre-trained language model for financial domain](#). In *Proceedings of EMNLP*.
- Uri Shalit, Fredrik D. Johansson, and David A. Sontag. 2017. [Estimating individual treatment effect: generalization bounds and algorithms](#). In *Proceedings of ICML*.
- Claudia Shi, David M. Blei, and Victor Veitch. 2019. [Adapting neural networks for the estimation of treatment effects](#). In *Proceedings of NeurIPS*.
- J. Sława-Neyman, Dorota Dabrowska, and Terence P. Speed. 1990. On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*.
- Dhanya Sridhar and Lise Getoor. 2019. [Estimating causal effects of tone in online debates](#). In *Proceedings of IJCAI*.
- Victor Veitch, Dhanya Sridhar, and David M. Blei. 2020. [Adapting text embeddings for causal inference](#). In *Proceedings of UAI*.
- Matthew James Vowels, Necati Cihan Camgoz, and R. Bowden. 2020. Targeted vae: Variational and targeted learning for causal inference. *IEEE International Conference on SMDS*.
- William Yang Wang and Zhenhao Hua. 2014. [A semi-parametric Gaussian copula regression model for predicting financial risks from earnings calls](#). In *Proceedings of ACL*.
- Galen Cassebeer Weld, Peter West, Maria Glenski, David T. Arbour, Ryan A. Rossi, and Tim Althoff. 2020. [Adjusting for confounders with text: Challenges and an empirical evaluation framework for causal inference](#). *Proceedings of AAI*, abs/2009.09961.
- Zach Wood-Doughty, Ilya Shpitser, and Mark Dredze. 2018. [Challenges of using text classifiers for causal inference](#). In *Proceedings of EMNLP*.
- Zach Wood-Doughty, Ilya Shpitser, and Mark Dredze. 2021. [Generating synthetic text data to evaluate causal inference methods](#). *ArXiv preprint*.
- Jeffrey M. Wooldridge. 2016. Should instrumental variables be used as matching variables. *Research in Economics*.
- Anpeng Wu, Kun Kuang, Junkun Yuan, Bo Li, Pan Zhou, Jianrong Tao, Qiang Zhu, Yueting Zhuang, and Fei Wu. 2020. [Learning decomposed representation for counterfactual inference](#). *ArXiv*, abs/2006.07040.
- Linyi Yang, Jiazheng Li, Ruihai Dong, Yue Zhang, and Barry Smyth. 2022. [Numhtml: Numeric-oriented hierarchical transformer model for multi-task financial forecasting](#). In *Proceedings of AAI*.
- Mengyue Yang, Furui Liu, Zhitang Chen, Xinwei Shen, Jianye Hao, and Jun Wang. 2021. [Causalvae: Disentangled representation learning via neural structural causal models](#). In *IEEE Conference on CVPR*.
- Weijia Zhang, Lin Liu, and Jiuyong Li. 2021. [Treatment effect estimation with disentangled latent factors](#). In *Proceedings of AAI*.

A Data Statistics

Table A1 shows the detailed statistics of each scenario.

Treatment	Train		Dev		Test	
	# Treat.	# Ctrl.	# Treat.	# Ctrl.	# Treat.	# Ctrl.
<i>Stock Volatility</i>						
Political Risk	8,000	8,000	1,000	1,000	6,000	6,000
Sentiment	8,000	8,000	1,000	1,000	6,000	6,000
<i>Stock Movement</i>						
Political Risk	8,000	8,000	1,000	1,000	6,000	6,000
Sentiment	8,000	8,000	1,000	1,000	6,000	6,000

Table A1: Data statistics.

B Hyper-parameters

Table A2 shows the detailed hyper-parameters setting of DIVA under all scenarios.

Hyper-parameter	
Framework	Pytorch
GPUs	1 A100
Batch Size	86
Epoch	30
Warmup Steps	10%
Learning Rate	5.00E-05
Optimizer	AdamW
Adam ϵ	1E-08
Max Sequence Length	512
Hidden Size	798
Hidden Layer	12
Dropout probability	0.2
Latent Dimension	200
Coefficient α	1
Coefficient β	1
Coefficient γ	0.1
Coefficient η	0.1
Coefficient λ	0.01

Table A2: Hyper-parameters of DIVA.