Conceptualizing Treatment Leakage in Text-based Causal Inference

Adel Daoud[1,3,4] **Connor T. Jerzak**[1,3] **Richard Johansson**[2,4]
[1]Linköping University, [2]University of Gothenburg,
[3]Harvard University, [4]Chalmers University of Technology
adel.daoud@liu.se, connor.jerzak@liu.se, richard.johansson@gu.se

Abstract

Causal inference methods that control for text-based confounders are becoming increasingly important in the social sciences and other disciplines where text is readily available. However, these methods rely on a critical assumption that there is no *treatment leakage*: that is, the text contains only information about the confounder and no information about treatment assignment (leading to post-treatment bias). However, this assumption may be unrealistic in real-world situations involving text, as human language is rich and flexible.

We first define the leakage problem, discussing the identification and estimation challenges it raises. We also discuss the conditions under which leakage can be addressed by removing the treatment-related signal from the text in a pre-processing step we define as *text distillation*. Then, using simulation, we investigate the mechanics of treatment leakage on estimates of the average treatment effect (ATE).

1 Introduction

In observational settings, scholars need to collect information about potential confounders in order to estimate the causal effect (τ) of a treatment on an outcome. If we observed the set of confounders directly, we could condition on those quantities to recover unbiased causal effects. Yet, because some confounders U are difficult to measure directly, scholars are turning to alternative data sources, such as medical records, policy documents, or social media posts, to indirectly measure (proxy) confounders. Recent methodological frameworks supply ways of integrating high-dimensional text data into causal estimation (Mozer et al., 2020; Roberts et al., 2020; Feder et al., 2021).

However, prior literature has primarily assumed that documents only contain information about the confounder, and no information about the treatment—something we term the *no-treatment-leakage assumption*. When treatment leakage occurs after treatment assignment, its bias is equivalent to a post-treatment bias.

Treatment leakage leads to an identification challenge. The challenge is that $\boldsymbol{W}$ is both necessary for adjusting (as it is a proxy) yet it is also a post-treatment variable. Without treatment leakage, $\boldsymbol{W}$ would not be a post-treatment variable, as it does not harbour information about the treatment assignment. But because of leakage, scholars would have to accept bias arising from either adjusting on a post-treatment variable or bias arising from not adjusting for unobserved confounding that parts of $\boldsymbol{W}$ represent. Although several methodological studies develop and adapt causal-inference methods for text data (Keith et al., 2020), almost no studies examine the leakage bias dynamics.

Our work investigates the treatment-leakage challenge. It shows that if $\boldsymbol{W}$ is the only available text representing U and there exists a distillation method, f, that has the ability to transform (e.g. partition) $\boldsymbol{W}$ into its post-treatment $\boldsymbol{W}_T$ and proxy textual-components $\boldsymbol{W}_U$, then adjusting on $\boldsymbol{W}_U$ is then the best one can do in identifying τ. As $\boldsymbol{W}_U$ is not post-treatment, we can adjust for it to reduce the bias when estimating τ. These f functions can represent a human annotator, identifying and removing parts of text (e.g., words, sentences) that belong to $\boldsymbol{W}_T$ and curating $\boldsymbol{W}_U$; or, under additional assumptions, f can be based on supervised or unsupervised machine-learning machine learning models that transform the text or its representation.

In this paper, we define key assumptions and demonstrate the mechanics of text distillation in a simulated experiment. Using a language model, we generate synthetic documents $\boldsymbol{W}$ so that they contain information about the treatment assignment, T, and the unobserved confounding, U, imprinted paragraph-by-paragraph. Because we control which paragraph is affected by T (injecting post-treatment bias) or by U (infusing knowledge about the confounder), we have an oracle distillation function, f, that mimics human coding. This oracle method perfectly distills $\boldsymbol{W}$, and supplies $\boldsymbol{W}_U$. Then, when using $\boldsymbol{W}_U$ in our causal model, we reduce bias of $\hat{\tau}$ markedly. Although our oracle is idealized, it deepens intuition. In future work, we will develop automated methods to obtain f.

By conceptualizing the problem of treatment leakage in text data and investigating its impact, scholars developing causal methods can be better positioned to tailor their frameworks to reduce bias; domain scholars can better calibrate their data collection procedure to account for this leakage.

2 Treatment Leakage in Text Data

While the literature on dealing with confounding in observational studies is established (Rubin, 1974), recent advances have been made in the analysis of text-based causal inference. Indeed, text $\boldsymbol{W}$ is widely available in the health and social sciences (Gentzkow et al., 2019; Kino et al., 2021), and can be used to proxy for some confounders, U, that would otherwise remain unobserved (Keith et al., 2020). If the text only contains information about U and no other factors, then $\boldsymbol{W}$ is a faithful representation of U and we denote it as $\boldsymbol{W}_U$. However, text, by its nature as a medium of creativity, rarely has fixed boundaries, and can contain information not only about confounders, but also leak information about the treatment assignment and its effects.

The future- and backward-looking nature of text can exacerbate treatment leakage. Documents that often contain backward looking temporally (e.g. in much of journalism) or has an unknown production date, will like contain information about the treatment and its effects. Using these documents directly for causal inference would inject post-treatment bias. Conversely, documents that reference the future (e.g., many public-policy documents in the economy and polity) may also lead to unfavorable RMSE if they predict the future well (see §2.1.2). As a result, a substantial amount of real-world text containing rich information about confounding factors might be affected by that language can reference the future, post-treatment state.

2.1 Characterizing Treatment Leakage

We define *treatment leakage* as when the text, $\boldsymbol{W}$, is affected by treatment status, T.

$$\textit{Treatment leakage:}\ \exists\, \boldsymbol{W}_T, \boldsymbol{W}_U \subseteq \boldsymbol{W},$$

so that there is a part of the text affected by T (i.e., $\boldsymbol{W}_T$) and a part affected by U (i.e., $\boldsymbol{W}_U$). In general, these two parts could be intersecting, if, for example, both the confounder and the treatment affect the overall tone (sentiment) of a document.

The degree of treatment leakage can vary. First, it depends on how much signal T injects into $\boldsymbol{W}_T$. Second, it also depends on the relative (distributional) strength of each partition $\boldsymbol{W}_U$ and $\boldsymbol{W}_T$ in jointly composing the entire document $\boldsymbol{W}$. Additionally, there will be random noise R irrelevant for T and U, yet present in composing $\boldsymbol{W}$. If we simplify and assume that each partition carries strength equal to the number of its elements (e.g., words) and each element has the same strengths, then the magnitude of treatment leakage is: $\frac{|\boldsymbol{W}_T|}{|\boldsymbol{W}_U \cup \boldsymbol{W}_T \cup R|}$, taking values $(0, 1)$.

2.1.1 Case 1: Text is Post-treatment

In one form of this phenomenon, the text is conditionally dependent on the treatment status T given a hidden confounder, U. Figure 1, panel *a.*, shows a directed acyclic graph (DAG) representing this scenario where the text affected by the treatment status. This sort of treatment leakage induces post-treatment bias: when the text is affected by the treatment, conditioning on the text (which is a collider) opens the path from T to Y through $\boldsymbol{W}$ and U, will in general yield biased estimates (in the notation of Pearl (2015), $(Y \not\perp T|\boldsymbol{W})_{G_{\underline{T}}}$).

Identification assumptions may also be hard to maintain, with the treated/control units having distinct text features (e.g. if all treated units have associated texts referring to the treatment). This lack of overlap would violate the identification assumptions of causal estimators such as Inverse Propensity Score Weighting (IPW) (Heinrich et al., 2010), and could lead to extreme estimated probabilities, something we see empirically in Figure 3).

2.1.2 Other Cases

Figure 1 shows a case when text is post-treatment, but in other cases the precise DAG structure may not be known. For example, text may represent a mediator if the document includes post-treatment information and also affected the outcome (if, for example, the text is congressional speech and the outcome is a roll call vote). If the proxy text is pre-treatment and directly affects the treatment, conditioning on the treatment-related portion of the text could increase the variance of estimation, leading to unfavorable RMSE (Myers et al., 2011).

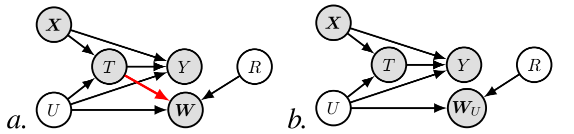

Figure 1: A causal model consisting of observed variables (shaded): confounders ($\boldsymbol{X}$), treatment (T), outcome (Y), document ($\boldsymbol{W}$), and unobserved variables (unshaded): confounder (U) and residual factors (R). The red-colored edge in *a.* represents the treatment leakage. In *b.*, A distillation function f has removed the treatment information in the text, leaving only information from the confounder. A perfect intervention of f is equivalent with deleting the red arrow; a less than perfect intervention reduces at least its dependence.

3 Text Distillation as Preprocessing

Text distillation is a form of text preprocessing. It has to target any text (e.g., tone, words, sentences) that belongs to W_T, and remove it from W. Thus, distillation ensures that the treatment signal is negated. As Figure 1, panel *b.* shows, if distillation is perfectly successful, it results in cutting the red arrow (from T to $\boldsymbol{W}$). The arrow is cut, because the distillation function has removed $\boldsymbol{W_T}$ from $\boldsymbol{W}$, supplying $\boldsymbol{W_U}$ for causal analysis.

3.1 Assumptions for Valid Distillation

Distillation relies on $\boldsymbol{W}$ and its key components being *separable*: that it can be decomposed into two portion (e.g., sets of paragraphs), where the first is only determined by U and the second only by T. That is,

$$\textit{Separability Assumption:} \quad \boldsymbol{W_U} \cap \boldsymbol{W_T} = \varnothing$$

Assuming separability, a perfect distillator will produce $\boldsymbol{W}^* = f(\boldsymbol{W})$ that is equivalent to the confounder, $\boldsymbol{W_U}$. *Perfect* distillation means that the distillator f identified text that contains the same information about U as $\boldsymbol{W_U}$ has. Thus, $\boldsymbol{W}^* = \boldsymbol{W_U}$, and if $\boldsymbol{W_U}$ is a valid adjustment set, then $\boldsymbol{W}^*$ is that as well. The separability assumption is appealing because it implies that researchers only need to find a valid partition of the text (and do not need to consider all possible text transformations).

This separability assumption is particularly plausible for text data, which by its nature consists of a sequence of linguistic signifiers which can be decomposed into smaller units (e.g. paragraphs).

While plausible for many circumstances, in some cases, separability may not hold, as when the entire tone of the text is affected by the treatment. In this more complicated setting, we need a more general assumption, that the transformed text, $\boldsymbol{W}^*$, is conditionally independent of T given U. That is, the conditional mutual information between $\boldsymbol{W}^*$ and T given U is zero, while information about U in $\boldsymbol{W}^*$ is maintained. Despite the benefits of this more general framing, because U is unobserved, it may be difficult for investigators to assess whether the assumption is satisfied or whether ethically problematic information has been included in the f function (e.g., race; Menon and Williamson (2018)). Unlike numerical data, as text data is readable, scholars can examine and validate whether $\boldsymbol{W}^*$ still contains information about T.

4 Experimental Setup

We use simulation to illustrate the dynamics of text distillation and build on the framework for evaluating text-based causal inference methods introduced by Wood-Doughty et al. (2021). We generate numerical covariates from the model in Figure 1; the general procedure is described in §A, with implementation details in §B. Parameters are selected so that ATE estimates $\hat{\tau}$ are biased if the estimator does not account for the unobserved confounder U.

Following Wood-Doughty et al. (2021), we generate documents, $\boldsymbol{W}$, by sampling from an English-language GPT-2 model (Radford et al., 2019). In contrast to their approach, text generation is conditioned not only on U but also on T. As described in detail in §A, we define paragraph-level topics, where some topics are associated with U, some with T, and some with a residual topic related only to other background variables (R in Figure 1). For a given paragraph topic, we define a number of prompts and a distribution shift that increases the probability of generating topic-related keywords.

As we simulate and record which paragraphs are affected by T and by U, our distillator f has oracle properties. We can then use f to investigate three idealized distillation scenarios. The first is when a distillator was not applied or the distillator failed to do any distillation $f(\boldsymbol{W}) = \boldsymbol{W}$. It outputs the same corpus. The second is when it perfectly distills $\boldsymbol{W}$, excluding all paragraphs affected by T. That is, apply $f(\boldsymbol{W}) = \boldsymbol{W}^*$ such that $\boldsymbol{W}^* = \boldsymbol{W_U}$. The third scenario is when f was overly aggressive and "accidentally" removed not only $T-$ but also U-related paragraphs, resulting in $\boldsymbol{W}^{**}$. This corpus violates the proxy-faithfulness assumption that

$\boldsymbol{W}^{**}$ fully measures U. Then, we use the three corpora, one at a time, for causal inference. We use an Inverse Propensity Weighting (IPW) esitmator, fully described in §C.

5 Experiments and Results

Based on the setting described in §4, our analysis produces six estimates, three based on distillation and three based on facts about the data-generating process. Figure 2 shows all estimates.

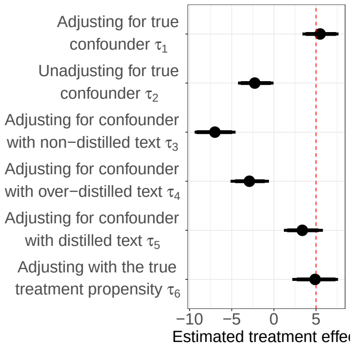

Figure 2: Estimates under different distillation regimes.

The ***first estimate***, $\hat{\tau}_1 = 5.5$, is the baseline where all information is known to the outcome model, including U. Because this linear model adjusting for U and $\boldsymbol{X}$ is equivalent to the data-generating model, and the estimated effect would be equal to the true value of 5 without sampling noise. The bootstrapped 95% confidence interval (CI) is 3.4 to 7.6. The ***second estimate***, $\hat{\tau}_2$ = -2.3, is obtained when U is omitted from the model to induce omitted variable bias (CI: -4.2, -0.1).

The ***third estimate***, $\hat{\tau}_3$, uses IPW to estimate the ATE (see §C). Here, we use the non-disitilled documents, $\boldsymbol{W}$, to estimate propensities. As Figure 2 shows, in the absence of distillation, the bias *increases* compared to conditioning on $\boldsymbol{X}$ alone, producing $\hat{\tau}_3 = -7.0$ (CI: -9.4, -4.6). The ***fourth estimate***, $\hat{\tau}_4$, applies overly aggressive distillation. This approach gives a result similar to the unadjusted estimate: $\hat{\tau}_4 = -2.9$ (CI: -5.1, -0.6).

The ***fifth estimate***, $\hat{\tau}_5$, applies oracle distillation by removing the paragraphs we know were affected by T. Using $\boldsymbol{W}^*$, the bias is reduced substantially, yielding an estimate $\hat{\tau}_5 = 3.5$ (CI: 1.2, 5.8). As the CI of this $\hat{\tau}$ includes the true $\tau = 5$, we conclude that distillation successfully recovers τ. However, we note that this recovery is not perfect and will be affected by sampling and modeling parameters.

The ***sixth estimate***, $\hat{\tau}_6$, demonstrates the impact of model selection for the propensity estimator. Using the *true* (simulated) propensity, the IPW estimate is $\hat{\tau}_6 = 4.9$ (CI: 2.2, 7.6). This result shows that further gains could be made by careful model selection (Chernozhukov et al., 2018).

Figure 3 shows distributions of propensity values for $\hat{\tau}_3$, $\hat{\tau}_5$, and $\hat{\tau}_6$. Without distillation (red), the estimated propensities cluster near 0 and 1. T is predicted almost perfectly, as mentioned in §2.1.1, causing the IPW estimate to to be similar to the unweighted one. Conversely, with distillation, the predicted probabilities are now similar to the data-generating propensities, and thereby, the resulting causal estimate is improved.

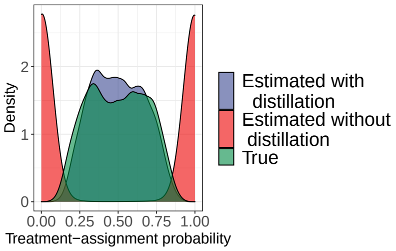

Figure 3: Estimated and true assignment probabilities.

6 Discussion

This paper shows the critical role of the no-treatment-leakage assumption when using text for causal inference. While text is becoming an established data source, it may harbour valuable information about a confounder but also contaminating information about post-treatment effects. This issue has seen little discussion in text-based causal inference literature, but has the potential to severely bias causal estimates, potentially leading to false discoveries or invalid policy recommendations.

Our study has limitations. First, more work is required to show how the no-treatment-leakage assumption operates under different covariance structures. Second, a larger simulation framework is needed to decompose estimator bias and variance.

For extensions, we recommend two paths. First, while this paper focuses on treatment leakage, there are other types of leakage when a single document is a function of multiple causal nodes. Thus, a generalization of the no-treatment-leakage assumption is the *no-node-leakage assumption*. Second, researchers need a framework when human partitioning of text is not possible due to corpus size. *Automatic* distillation could be attempted with additional assumptions, perhaps building from the literature on removing sensitive information in data (Bolukbasi et al., 2016; Ravfogel et al., 2020).

Acknowledgements

Richard Johansson was supported by the projects *Interpreting and Grounding Pre-trained Representations for NLP* and *Representation Learning for Conversational AI*, both funded by Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation. Adel Daoud would like to acknowledge a grant from The Royal Swedish Academy of Letters, History and Antiquities.

References

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. 2018. Double/debiased machine learning for treatment and structural parameters.

Adel Daoud, Bernhard Reinsberg, Alexander E. Kentikelenis, Thomas H. Stubbs, and Lawrence P. King. 2019. The International Monetary Fund's interventions in food and agriculture: An analysis of loans and conditions. *Food Policy*, 83:204–218.

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. 2021. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *arXiv preprint arXiv:2109.00725*.

Michele Jonsson Funk, Daniel Westreich, Chris Wiesen, Til Stürmer, M Alan Brookhart, and Marie Davidian. 2011. Doubly robust estimation of causal effects. *American journal of epidemiology*, 173(7):761–767.

Matthew Gentzkow, Bryan Kelly, and Matt Taddy. 2019. Text as data. *Journal of Economic Literature*, 57(3):535–74.

Carolyn Heinrich, Alessandro Maffioli, Gonzalo Vazquez, et al. 2010. A primer for applying propensity-score matching. *Inter-American Development Bank*.

Katherine Keith, David Jensen, and Brendan O'Connor. 2020. Text and causal inference: A review of using text to remove confounding from causal estimates. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5332–5344, Online. Association for Computational Linguistics.

Shiho Kino, Yu-Tien Hsu, Koichiro Shiba, Yung-Shin Chien, Carol Mita, Ichiro Kawachi, and Adel Daoud. 2021. A scoping review on the use of machine learning in research on social determinants of health: Trends and research prospects. *SSM - Population Health*, 15:100836.

Aditya Krishna Menon and Robert C Williamson. 2018. The cost of fairness in binary classification. In *Conference on Fairness, Accountability and Transparency*, pages 107–118. PMLR.

Reagan Mozer, Luke Miratrix, Aaron Russell Kaufman, and L. Jason Anastasopoulos. 2020. Matching with text data: An experimental evaluation of methods for matching documents and of measuring match quality. *Political Analysis*, 28(4):445–468.

Jessica A Myers, Jeremy A Rassen, Joshua J Gagne, Krista F Huybrechts, Sebastian Schneeweiss, Kenneth J Rothman, Marshall M Joffe, and Robert J Glynn. 2011. Effects of adjusting for instrumental variables on bias and precision of effect estimates. *American journal of epidemiology*, 174(11):1213–1222.

Judea Pearl. 2015. Conditioning on post-treatment variables. *Journal of Causal Inference*, 3(1):131–137.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7237–7256. Association for Computational Linguistics.

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

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

Donald B. Rubin. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701.

Zach Wood-Doughty, Ilya Shpitser, and Mark Dredze. 2021. Generating synthetic text data to evaluate causal inference methods. *arXiv preprint arXiv:2102.05638*.

A Synthetic Data Generation

We first summarize the general approach in this section and provide details for the simulation in §5 in the next section.

For each document i, we first draw observed and unobserved confounders $\boldsymbol{X}_i$ and U_i, and then the treatment T_i. For each paragraph j in the document, we draw a paragraph topic Z_{ij}, depending on the values of U_i and T_i, and then a prompt W_{ij}^0 depending on the value of Z_{ij}. Finally, we sample from the GPT-2 language model[1] to generate the paragraph text $\boldsymbol{W}_{ij}$, starting from the prompt W_{ij}^0 and with a vocabulary distribution shift defined by Z_{ij}. Algorithm 1 shows the pseudocode.

Algorithm 1 Generation of synthetic data.

```
for $i \in 1, \ldots, N$
    $\boldsymbol{X}_i \sim f_X$
    $U_i \sim f_U$
    $T_i \sim \text{Bernoulli}(\text{sigmoid}(f_T(\boldsymbol{X}_i, U_i)))$
    $Y_i \sim f_Y(\boldsymbol{X}_i, U_i, T_i)$
    for $j \in 1, \ldots, K$
        $Z_{ij} \sim \text{Categorical}(f_Z(U_i, T_i))$
        $W_{ij}^0 \sim \text{Categorical}(f_{W^0}(Z_{ij}))$
        $\boldsymbol{W}_{ij} \sim \text{LM}(W_{ij}^0, Z_{ij})$
```

In the pseudocode above, the functions f_X, f_U, f_T, and f_Y define the distributions of the observed confounders, unobserved confounder, treatment and outcome, respectively. On the paragraph level, the function f_Z defines a categorical distribution over paragraph topics, and f_{W^0} a categorical distribution over prompts.

Similarly to Wood-Doughty et al. (2021), we use two mechanisms to condition the generation of a paragraph on a topic Z: a prompt and a vocabulary distribution shift. The distribution shift is designed to promote a set of *keywords* related to the topic and we implement it by multiplying the language model probabilities by a topic-specific vector θ_Z of scale factors:

$$P'(w|\text{context}, Z) \propto P_{\text{LM}}(w|\text{context}) \cdot \theta_Z(w)$$

B Parameterization Used in §5

In §5, we generated N = 10,000 instances, each consisting of numerical values and a document. We used the following distributions to generate the document-level variables: f_X was a 3-dimensional isotropic Gaussian; f_U was an even coin toss; f_T was linear in $\boldsymbol{X}_i$ and U_i; f_Y was Gaussian with a mean defined by a linear function of $\boldsymbol{X}_i$, U_i, and T_i and a fixed standard deviation.

Each document consisted of K = 20 paragraphs. For the paragraph generation, we defined five different topics: two corresponding to positive and negative treatment values; two corresponding to positive and negative values of the unobserved confounder; one general background topic that was unrelated to U or T (but conceptually thought of as controlled by other “residual” variables R). For a document with given values of U and T, we set the topic distribution f_Z to select the U topic with a probability of 0.2, the T topic with a probability of 0.2, and the general topic with a probability of 0.6.

The generated texts were designed to simulate a hypothetical use case where the researchers want to investigate the effect of IMF programs on some country-level indicator (cf. Daoud et al., 2019). The treatment variable T represents the presence or absence of an IMF program; the unseen confounder U represents the political situation of the country with respect to the IMF. For each topic except the general topic, we define four different prompts: for instance, for a positive treatment value, one of the prompts was *The International Monetary Fund mandates the deregulation of [COUNTRY]’s labor market*. In the analysis, *“[COUNTRY]”* is substituted by randomly sampled country names.

All topics except the general topic defined a distribution shift used when generating from the language model. We used 8 topic keywords for each of these topics. For these keywords, the corresponding entries in the vocabulary distribution shift vector $\log \theta_Z$ were set to a value that defines the strength of the effect of T on $\boldsymbol{W}$; for all other words except these keywords, $\log \theta_Z$ was 0. Since our focus in this paper is on a clear-cut use case where the effects are strong, we set the strength parameter to a value of 4, which gives a noticeable effect on the generated texts.

The text generation model was run on a single GPU (NVIDIA GeForce GTX TITAN X). Generating the 10,000 documents took around 10 hours. The generation of random text is within the intended use of the GPT-2 model.[2]

[1] We used the implementation from the HuggingFace repository, https://huggingface.co/gpt2.

[2] https://huggingface.co/gpt2

C IPW Details

C.1 Background

The ATE is defined as $\tau = \mathbb{E}[Y_i(1) - Y_i(0)]$, where $Y_i(t)$ is the potential outcome for unit i under treatment t. It can be identified in randomized experiments (Rubin, 1974). However, the situation is more complicated in the observational setting, where the treatment is not randomized to units but could be correlated with confounders, $\boldsymbol{X}_i$, that are associated with the treatment and the outcome. In that setting, we can, with additional assumptions, still recover the ATE using Inverse Propensity Weighting (IPW) or related robust methods (Funk et al., 2011), where observations are weighted by the inverse of their estimated treatment probabilities $\hat{\pi}(\boldsymbol{X}_i) = \widehat{\Pr}(T_i = 1|\boldsymbol{X}_i)$ (Rosenbaum and Rubin, 1983): $\widehat{\tau} = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{T_iY_i}{\hat{\pi}(\boldsymbol{X}_i)} - \frac{(1-T_i)Y_i}{1-\hat{\pi}(\boldsymbol{X}_i)}\right\}$.

C.2 Estimation

ATE estimates based on Inverse Propensity Weighting (see §C.1) require the estimation of the propensity scores, $\widehat{\Pr}(T|\boldsymbol{X}, \boldsymbol{W})$. To estimate these scores, we applied a L_1-regularized logistic regression model using the `glmnet` package in `R`. The regularization strength (λ) was set automatically via 10-fold cross-validation. When estimating propensities, we represented the (non-distilled or distilled) document as an L_2-normalized TF-IDF vector using the 256 most frequent terms in the vocabulary, while the numerical covariates $\boldsymbol{X}$ were standardized.