Consistent Estimation of the Average Treatment Effect with Text as Confounder

Stanford CS224N Custom Project

Evan Munro
Graduate School of Business
Stanford University
name@stanford.edu

Abstract

I show that embedding representations of text can be used to construct a root-n consistent estimator of the average treatment effect under confounding. I explore using both GloVE embeddings as well as transformer-based document embeddings to integrate text data in the double machine learning framework for causal inference [1]. Using a large dataset of consumer complaints from 2018-2021 published by the CFPB, I estimate the causal effect of a complainant identifying themselves as an older American on the probability their complaint is resolved with monetary or non-monetary compensation. I show that including a representation of text reduces the treatment effect estimate.

1 Introduction

We are interested in estimating the average treatment effect when the treatment assignment is confounded by a latent variable that is estimable from text. We start by defining the average treatment effect, and formally defining the conditions under which the average treatment effect can be consistently estimated when text data is a confounder. The key condition is unconfoundedness, which assumes that the treatment assignment is random conditional on covariates and a representation of text. Then, we discuss various unsupervised representations of text, including word counts, topic models, GloVE embeddings, and embeddings from transformer models. We show how GloVE embeddings and transformer embeddings can be integrated into a framework for root-n consistent estimation of the average treatment effect. We discuss why using supervised text models is challenging in this setting and point to future avenues of work for using supervised models for text. For the transformer representation, we use S-ROBERTA [2]. We use data on complaints filed to the Consumer Financial Protection Bureau (CFBP) to estimate the causal effect of tagging a complaint as coming from an older American on receiving compensation from the company in response to the complaint. We find that controlling for confounders in this setting reduces the treatment effect estimate compared to a simple difference in means estimate.

2 Related Work

The closest work is [3], who use text embeddings from topic models and from BERT to estimate confounders of the average treatment effect on the treated. A novel contribution of this paper is to link embedding estimation with the results of [1]. This ensures the root-n consistency of the average treatment effect estimate when an unknown representation of text is a confounder. We choose SROBERTA rather than BERT, since BERT has not been shown to generate very good document embeddings.

Another related paper [4], who are interested in causal inference when text is the treatment or the outcome, in contrast to this paper which is interested in text as a confounder in causal inference settings. They apply a sample splitting approach to reduce overfitting when estimating a latent representation of text but use a topic model to estimate latent representations of text. They also use CFPB data, but rather than using text as a confounder, they estimate a model where the text content is the treatment. Our paper is the first to discuss how to construct a root-n consistent estimate of the average treatment effect when using text as confounder.

3 Approach

We are interested in estimating the average treatment effect when the treatment assignment is confounded by a latent variable that is estimable from text. We start by defining the average treatment effect, and formally defining the conditions under which the average treatment effect can be consistently estimated when text data is a confounder. A novel contribution of this paper is to explore the link of using text as confounders with the results of [1] on integrating high-dimensional data in causal inference, while maintaining root-n consistency.

We follow the potential outcomes approach to defining causal estimands. For $i \in \{1, \ldots, n\}$, we observe data (Y_i, W_i, X_i, G_i). G_i represents a raw sequence of text associated with individual i. $C_i \in \mathbb{R}^J$ are a set of observed covariates about individual i. $W_i \in \{0, 1\}$ is a binary treatment. We define potential outcomes $Y_i(1)$ and $Y_i(0)$ [5], which are the values of the outcome for each individual for each the two possible treatment statuses. However, we only observe the outcome for the assigned treatment: $Y_i = Y_i(W_i)$. We are interested in estimating the average treatment effect

$$\tau = \mathbb{E}[Y_i(1) - Y_i(0)]$$

when the treatment is not randomly assigned. Let $\gamma(G_i) \in \mathbb{R}^J$ be some lower dimensional representation of text. We can make progress in estimating the treatment effect by assuming **unconfoundedness**. For all $\gamma \in \Gamma$:

$$\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp W_i | X_i, \gamma(G_i) \tag{1}$$

The next step is to define the class Γ. We will discuss a few different options.

1. The simplest representation of text is word-frequencies from a hand-curated dictionary, which is commonly used in economics, see [6] and [7]. Capturing the meaning of text that is relevant for removing confounding of the treatment assignment, however, is a complex task, and hand-curating a dictionary for this task would require a new dictionary for each new causal inference task.

2. Topic models are another possibility. Topic models do not consider sentence-level context, however, that may be important for understanding meaning in a text that may be important for controlling for confounding.

3. Document vectors from averaging word vectors from word2vec [8] or GLoVE [9] are a better option; word vectors take into account some context that may be relevant for confounding, so we use a document embedding from GloVE as a baseline method. However, there are long range dependencies in a document that may be relevant for causal inference and require a more complex model to capture.

4. Our final approach uses embeddings from SROBERTA. The pretraining makes the 500-dimensional embedding constructed from an input sequence meaningful for general language understanding and document similarity tasks. This approach is shown in Figure 3. Given the general purpose nature of the embeddings from SROBERTA, it is reasonable that SROBERTA embeddings pretrained on any large corpus of text define the Γ that are most likely to meet the unconfoundedness assumption in (1).

We assume that treatment assignment is random conditioning on observed covariates and a lower dimensional representation of text for all mappings γ in some class Γ. Using sample splitting to estimate nuisance functions, following the approach in [1], results in consistent estimation of the average treatment effect when confounders are high-dimensional. First, split the data into two random halfs I_1 and I_2. We have that $\hat{\tau} = \frac{|I_1|}{n}\hat{\tau}_1 + \frac{|I_2|}{n}\hat{\tau}_2$, where:

$$\hat{\tau}_1 = \frac{1}{|I_1|}\sum_{i \in I_1}(\hat{\mu}^{I_2}_{(1)}(X_i, \gamma(G_i)) - \hat{\mu}^{I_2}_{(0)}(X_i, \gamma(G_i))$$
$$+ W_i \frac{Y_i - \hat{\mu}^{I_2}_{(1)}(X_i, \gamma(G_i))}{\hat{e}^{I_2}(X_i, \gamma(G_i))} - (1 - W_i)\frac{Y_i - \hat{\mu}^{I_2}_{(0)}(X_i, \gamma(G_i))}{1 - \hat{e}^{I_2}(X_i, \gamma(G_i))})$$

$$\hat{\tau}_2 = \frac{1}{|I_2|}\sum_{i \in I_2}(\hat{\mu}^{I_1}_{(1)}(X_i, \gamma(G_i)) - \hat{\mu}^{I_1}_{(0)}(X_i, \gamma(G_i))$$
$$+ W_i \frac{Y_i - \hat{\mu}^{I_1}_{(1)}(X_i, \gamma(G_i))}{\hat{e}^{I_1}(X_i, \gamma(G_i))} - (1 - W_i)\frac{Y_i - \hat{\mu}^{I_1}_{(0)}(X_i, \gamma(G_i))}{1 - \hat{e}^{I_1}(X_i, \gamma(G_i))})$$

$\hat{\tau}_1$, the treatment effect on data split I_1, is constructed from the predictions of outcomes for individuals who are treated and not treated ($\hat{\mu}^{I_2}_{(1)}$, $\hat{\mu}^{I_2}_{(0)}$) and propensity scores $\hat{e}^{I_2}$, which is the prediction of treatment assignment conditional on context and embeddings. These functions are estimated using data split I_2. We call the outcome and propensity score functions nuisance functions. Under unconfoundedness, and suitable conditions on the estimators of the nuisance functions, [1] proves that $\sqrt{n}(\hat{\tau} - \tau) \rightarrow_p 0$ and that the estimator $\hat{\tau}$ is semi-parametrically efficient (not discussed in this report).

Theorem 1. (*Chernozhukov et. al, 2018*) [1] Under the following conditions:

1. **Overlap:** For some $\eta > 0$, $\eta < e(x, \gamma(g)) < 1 - \eta$ for all $x \in \mathcal{X}, g \in \mathcal{G}$.
2. **Consistency:** Regression estimates are sup-norm consistent.
$$\sup_{x \in \mathcal{X}, g \in \mathcal{G}} |\hat{\mu}_{(w)}(x, \gamma(g))| \rightarrow_p 0$$
$$\sup_{x \in \mathcal{X}, g \in \mathcal{G}} |\hat{e}(x, \gamma(g))| \rightarrow_p 0$$
3. **Risk decay:** All regression estimates are $o_p(n^{-1/4})$ consistent in root-mean-squared error.

Then we have that $\sqrt{n}(\hat{\tau} - \tau) \rightarrow_p 0$.

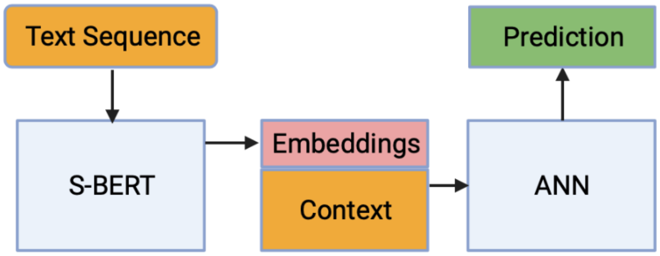

Figure 1: Neural Network Architecture for Causal Inference with Text

We next evaluate how the approach in Figure 3 can meet the criteria of Theorem 1.

1. Overlap is a condition that depends on the dataset used. It means that for any joint set of covariates and text representations, it is possible that the individual could have $W_i = 1$ or $W_i = 0$. In the data section, we will briefly discuss why this likely holds for our example.

[1] The exposition of these results follows the treatment in Stefan Wager's STAT 361 course notes.

2. Consistency and risk decay depends on the machine learning estimator used to construct the outcome and propensity score functions from x_i and $\gamma(g_i)$. Figure 3 contains an ANN, but we also use a random forest in place of this. We are restricted to machine learning estimators that meet these conditions. A recent paper has shown these conditions hold for deep neural networks, see [10]. The conditions also hold for random forests, see [11].

As long as our set of covariates and our text representation γ satisfies the unconfoundedness condition, then our estimate of the average treatment effect using the approach in Figure 3 is a root-n consistent estimation of the average treatment effect.

3.1 Challenges with Estimating a Text Representation

In all of the models we considered, we used unsupervised text representations, where the embedding mapping was either deterministic and specified by the researcher (in the case of word counts), or estimated using some dataset unrelated to the causal inference task at hand (e.g. GloVE and S-ROBERTA embeddings). As a result, we considered the text representation γ to be fixed rather than estimated from the data. It is worth discussing briefly why this choice was made for this paper. An alternative would be to estimate $\hat{\gamma}$ as well as $\hat{\mu}$ and $\hat{e}$. This can be very useful for generating embeddings that are more likely to satisfy the unconfoundedness condition in Equation 1. We had originally planned to fine-tune a transformer network to construct embeddings for causal inference, but it turns out there are theoretical issues with doing that. Models for supervised learning from text, such as supervised topic models or fine-tuned transformer models do not have strong convergence results available. The learned embedding representation $\hat{\gamma}$ will depend on the model architecture and may not converge to some unique embedding representation γ that we can assume fulfils the unconfoundedness condition. In contrast, if we take a pretrained model as given to build our embeddings in an unsupervised way, then we can make assumptions about the fixed embedding representation γ given by that pretrained model. For many classes of supervised text models, it is not straightforward to verify that models for the prediction of the conditional mean and the propensity score functions are sup-norm consistent and that they converge at an appropriate rate. A promising avenue for future work is to either prove some convergence rates for more complex text models or to formulate some weaker conditions for inference using high-dimensional confounders.

4 Experiments

4.1 Data

The Consumer Financial Protection Bureau (CFPB) was created in 2011. Its mission is to regulate the offering and provision of consumer financial products or services under federal consumer financial laws and educate and empower consumers to make better informed financial decisions[2]. As part of this mission, the CFPB collects complaints data on individuals about financial products and services. They use the complaints in the following way [3]:

1. Complaints are forwarded to the appropriate company for response
2. Complaints are shared with state and federal agencies, and reports are prepared for Congress.
3. Complaints data are analyzed to help enforce federal consumer financial laws, supervise companies, and write improved rules and regulations
4. De-identified complaints data are published publicly.

The dataset used in this paper is a set of 650,886 complaints submitted to the Consumer Financial Protection Bureau between January 01, 2019 and December 31, 2020. The dataset was collected from the CFPB website. The variables collected on each complaint include the text of the complaint, some optional tags selected by the complainant (e.g. if they are a veteran or an older American), the product category for the complaint, a categorization of the response of the company, and the date the complaint was submitted. Table 1 contains product categories for the complaint, which makes up the context space $X_i \in \mathcal{X}$ for our causal inference task. Table 2 contains a mapping of

[2] https://www.consumerfinance.gov/about-us/budget-strategy/

[3] https://www.consumerfinance.gov/complaint/data-use/

outcome categories for the complaints to the binary outcome $Y_i \in \{0, 1\}$. The treatment $W_i \in \{0, 1\}$ is equal to one if the complaint is tagged by whoever submitted the complaint as coming from an older American. The text of the complaint G_i is also a potential confounder that we would like to control for.

Complaint Product Categories
Debt Collection
Mortgage
Credit or prepaid card
Payday loan, title loan, or personal loan
Vehicle loan or lease
Student loan

Table 1: Categories for the complaint context $\mathcal{X}$

Complaint Outcomes
Closed with monetary relief ($Y_i = 1$)
Closed with non-monetary relief ($Y_i = 1$)
Closed with explanation ($Y_i = 0$)
Untimely response ($Y_i = 0$)
In progress ($Y_i = 0$)

Table 2: Mapping of complaint outcome categories to binary outcome

4.2 Evaluation

The parameter of interest is the average treatment effect of tagging a post as one submitted by an older American on the probability of receiving monetary or non-monetary compensation. The CFPB uses complaints data to enforce regulations and to write new regulations. Since older Americans are considered a more vulnerable group by the CFPB, it is possible that companies may treat complaints by older Americans more leniently. The CFPB may enforce regulations more strictly against companies with high rates of unresolved issues with older Americans. Alternatively, it is possible that complaints are dealt with by companies independently of who submits them. We are interested in using the framework described in Section 3 to analyze how the estimated treatment effect changes when different combinations of potential confounders are controlled for in the treatment effect estimate. Our evaluation metric is qualitative, rather than quantitative, and will examine how the estimated treatment effect changes when various combinations of controls are included.

The first estimate we will compute is a simple difference in means estimate, which examines how the resolution of complaints differ for complaints submitted by older Americans compared to younger Americans.

$$\hat{\tau}^{Diff-Means} = \sum_{i:W_i=1} Y_i/N_1 - \sum_{i:W_i=0} Y_i/N_0$$

This is a causal estimate only if the tagging of a complaint is random without conditioning on any information. This is not likely to be true. For example, older Americans may submit complaints for different products than younger Americans at a higher rate, and different products may be associated with different kinds of outcomes. As a result, our second estimate is the treatment effect estimate using the Double Machine Learning framework when we control only for the product category X_i. This is still not likely to be a causal estimate, as it is possible that the language and persuasion of the complaint text submitted by older Americans compared to younger Americans is different. As a result, it is necessary to control for differences in the complaint text to examine the causal effect of the tag only on outcomes. Our third type of estimate is a causal estimate controlling for both the product category and an embedding representation.

We should also note that for this dataset the overlap assumption in the DML framework holds; for any given set of complaint text and product category, it is certainly possible that the complaint was submitted by an older American or younger American.

4.3 Experimental Details

For the text analysis component of the paper, we use pretrained models. The document vector for the GloVE embedding representation is composed of averaging word vectors for every word in the complaint that is not a stop word. The GloVE embedding is a 50-dimensional vector from the Wikipedia 2014 + Gigaword 5 training dataset.

For the S-RoBERTA representation, we use the `sentence-transformers` package code and load the pretrained model `stsb-roberta-base` and use a GPU to estimate embeddings for each input sequence, truncated to 128 tokens, which captures most of the text for most complaints. Then, I estimate $\hat{\tau}$ using the DML framework with sample splitting for the outcome and propensity score estimation functions.

For the nuisance function estimation in the DML framework, we use two options.

1. The first is a random forest from `scikit-learn` for the estimation of the outcome and propensity scores. The random forest has a maximum depth of eight for each tree.
2. The second is a multi-layer perceptron from `scikit-learn` for the estimation of the outcome and propensity scores. The MLP regressor has hidden layers of size (100, 50) and is trained using the Adam optimization algorithm.

4.4 Results and Analysis

Treatment Effect Estimate	$\hat{\tau}$ (s.e.)
Difference in Means	0.0206 (0.0040)
Product Category Only + RF	0.0082 (0.0039)
GloVE vectors + RF	0.0063 (0.0038)
Product Category + ANN	0.0082 (0.0038)
GloVE vectors + ANN	0.0027 (0.004)
SROBERTA + ANN	0.0057 (0.004)

Table 3: Treatment Effect Estimates with Complaint Text as Confounders

The reason we should control for text content in this setting is that complaints submitted by older Americans may differ systematically in that they use different language and different persuasian tactics. Examining the causal effect of the tag alone requires us to control for differences in the text content. We find that the causal effect estimate when we control only for the type of the complaint is quite a bit lower than the causal effect when we do not control for any confounders. When we include both the product category and the document embeddings vector as a potential confounder, the estimate is is even lower, for both random forest and neural network nuisance function estimators, and is close to zero. The estimate using GloVE compared to RoBERTA embeddings for the random forest and for the neural network are all very similar. This suggests that there is not a significant causal effect of tagging a complaint as coming from an older American on the probability of receiving a positive outcome. Any difference in observed outcomes are due to differences in types and content of complaints that older Americans submit, rather than preferential treatment by companies due to the tag.

I conclude this section by highlighting a major limitation of the quantitative analysis performed. In this section we provided some quantitative evidence that including a representation of text meaningfully lowers the average treatment effect estimate. This indicates that the higher rate of complaint resolution for older Americans is partially explained by the differences in the types of complaints, as measured by the complaint category and the context of the text of the complaint. The ideal analysis in this setting would be to be able to conclude that using a certain representation of text leads to an estimated average treatment effect that is close to the true treatment effect. But, for real datasets where we do not know what the average treatment effect is or what the causal graph is, then this is not possible. One option for obtaining some more concrete quantitative results would be to use a synthetic dataset that includes confounding text and context variables, outcomes, treatments and a known average treatment effect. It is very challenging, however, to generate a synthetic dataset that would include realistic text sequences, without favoring one type of text embedding model over another. One interesting avenue for future work would be to examine whether a GAN approach could be used to generate a synthetic dataset that included text confounders from real datasets that included text. See [12] for a conditional GAN method for evaluating causal inference methods on datasets that do not include text.

5 Conclusion

In this paper, we have shown how text data can be incorporated into the Double Machine Learning framework of [1] using unsupervised models and pretrained models. In an example using CFPB data, we show that the causal effect of tagging a complaint as coming from an older american has close to zero effect on the probability of receiving monetary or non-monetary compensation.

We have identified significant challenges for improving methodology beyond the basic techniques used in this paper. Important future work for methodological advancements in causal inference with text data includes generating synthetic datasets that include text, perhaps using GANs. In addition, it would be helpful to understand how a supervised machine learning method can be used to estimate text embeddings that are predictive of the outcome and propensity score, without causing convergence issues with asymptotic analysis of the average treatment effect estimator.

References

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

[2] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.

[3] Victor Veitch, Dhanya Sridhar, and David M Blei. Using text embeddings for causal inference. *arXiv preprint arXiv:1905.12741*, 2019.

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

[5] Donald B Rubin. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331, 2005.

[6] Tim Loughran and Bill McDonald. When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of finance*, 66(1):35–65, 2011.

[7] Measuring economic policy uncertainty.

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

[9] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[10] Max H Farrell, Tengyuan Liang, and Sanjog Misra. Deep neural networks for estimation and inference. *Econometrica*, 89(1):181–213, 2021.

[11] Stefan Wager. Asymptotic theory for random forests. *arXiv preprint arXiv:1405.0352*, 2014.

[12] Susan Athey, Guido W Imbens, Jonas Metzger, and Evan M Munro. Using wasserstein generative adversarial networks for the design of monte carlo simulations. Technical report, National Bureau of Economic Research, 2019.