Some Reflections on Drawing Causal Inference using Textual Data: Parallels Between Human Subjects and Organized Texts

Bo Zhang BOZHAN@WHARTON.UPENN.EDU
Department of Statistics and Data Science
University of Pennsylvania
Philadelphia, PA 19104, USA

Jiayao Zhang ZJIAYAO@WHARTON.UPENN.EDU
Cognitive Computation Group and
Department of Statistics and Data Science
University of Pennsylvania
Philadelphia, PA 19104, USA

Editors: Bernhard Schölkopf, Caroline Uhler and Kun Zhang

Abstract

We examine the role of textual data as *study units* when conducting causal inference by drawing parallels between human subjects and organized texts. We elaborate on key causal concepts and principles, and expose some ambiguity and sometimes fallacies. To facilitate better framing a causal query, we discuss two strategies: (i) shifting from immutable traits to perceptions of them, and (ii) shifting from some abstract concept/property to its constituent parts, i.e., adopting a constructivist perspective of an abstract concept. We hope this article would raise the awareness of the importance of articulating and clarifying fundamental concepts before delving into developing methodologies when drawing causal inference using textual data.

Keywords: Causal inference; Constructivism; Natural language processing; Potential outcomes framework; Pretreatment variables

1. Introduction: Causal Inference with Textual Data

With the unprecedented empirical success of transformers in various natural language processing tasks spanning from text comprehension to machine translation, extracting and representing semantic meanings from textual data have witnessed the rise to a new greatness (Devlin et al., 2019; Liu et al., 2019; Radford et al., 2019). Recently, there is an abundance of interest in drawing causal conclusions using high-dimensional, structured or less structured, contextual or less contextual, representations of textual data in one way or another. There are at least three[1] roles of textual data in a causal query (Feder et al., 2021; Sap et al., 2020).

(i) Textual data serve as vehicles for reasoning (commonsense) causal relations of semantic meanings (Ning et al., 2018; Sap et al., 2020; Shwartz et al., 2020; Zhang et al., 2022).

(ii) Textual data encode important covariates or outcome of interest in human populations research (Egami et al., 2018). For instance, in a study of the treatment effect of a new technology in cardiac surgery, patients' medical history hand-written by their physicians could encode patients' important clinical characteristics, which are likely to confound medical decisions and clinical outcomes and should be adjusted for.

1. For example, there is also much interest in *causal model explanation/interpretation* that aims to understand a notion of model sensitivity, which is different from causal inference in the context of this essay.

(iii) Textual data are themselves *study units* in a causal query, and interests lie in intervening in some aspect of textual data (Veitch et al., 2020; Feder et al., 2021). For instance, Veitch et al. (2020) studied the "causal effects" of number of theorems in a conference paper on its acceptance rate; Sridhar and Getoor (2019) investigated if the tone of replies results in a change of sentiment using transcripts from online debates; Roberts et al. (2020) studied if authors' gender affects articles' citation in a large cohort of academic articles on international relations. Keith et al. (2020, Table 1) surveyed many additional causal queries involving textual data as study units.

The primary goal of this article is to discuss the role of text data as *study units* (role (iii) discussed above) in causal inference by drawing some useful parallels between organized textual data and human subjects understood broadly, the most typical study units in classical causal inference literature; see, e.g., Cochran and Chambers (1965), Rosenbaum (2002b), and Imbens and Rubin (2015). In particular, we (i) examine key concepts, including causal variables and covariates, associated with textual data, (ii) discuss some ambiguity, inconsistency, and common fallacies we have observed in the machine learning (ML) and natural language processing (NLP) literature, and (iii) summarize two useful strategies to help researchers better frame their causal queries when study units are textual data.

It perhaps comes as no surprise that much ambiguity we identified in the ML/NLP literature when study units are textual data was also encountered in human populations research in classical causal inference literature; hence, we found it valuable to bridge the classical causal inference literature, including works in statistics, biostatistics, economics, political science, etc, and the ML/NLP literature, so that some time-honored wisdom from pioneers in causal inference would benefit researchers in the ML/NLP field, and novel ideas and recent progress from ML/NLP and the abundance and ubiquity of textual data could also benefit causal inference practitioners in biomedical and social sciences. We will ground our discussion in the potential outcomes framework (Neyman, 1923; Rubin, 1974; Imbens, 2020), though many parallels between textual data and human subjects and caveats we derived should also be useful and readily available to other causal paradigms (Pearl, 1995, 2009; Heckman, 2005; Peters et al., 2017). We will use two examples discussed in Veitch et al. (2020) to illustrate basic concepts, principles, and sometimes fallacies. We have nothing against the authors; on the contrary, we believe that these authors, in Veitch et al. (2020) and many related works, have made many pioneering and useful contributions to expanding the boundary of causal inference to textual data.

2. Study Units and Covariates

One caveat Donald Rubin gave to practitioners of causal inference (and statistics at large) is that researchers should articulate the quantity of interest, i.e., *estimands*, prior to describing an algorithm and its associated output (Rubin, 2005).

In causal inference problems, a unit-level causal effect refers to a contrast in potential outcomes of a study unit, and a summary causal effect, e.g., the sample average causal effect, refers to causal effects averaged over a collection of units. The first key concept involved in defining a causal effect is the notion of "study units." For instance, in an analysis of the effect of utilizing a certain technology in cardiac surgery using the U.S. Medicare and Medicaid data, units involved in the study are mostly senior Americans aged over 65 (MacKay et al., 2021; Zhang et al., 2021a). A causal query where each study unit is clearly a complete piece of article, including all words, figures,

tables, references, and metadata like authors, their affiliations, etc, can be found in the following example in Veitch et al. (2020).

Example 1 Consider a corpus of scientific papers submitted to a conference. Some have theorems; others do not. The goal is to infer the causal effect of including a theorem on paper acceptance. The effect is confounded by the subject of the paper: more technical topics demand theorems, but may have different rates of acceptance. The data do not explicitly list the subject, but it does include each paper's abstract. It is hoped to use the text to adjust for the subject and estimate the causal effect.

Let us recall that associated with each study unit is a vector of, possibly high-dimensional, "covariates." According to Rubin (2005), covariates refer to

> *"variables that take their values **before** the treatment assignment or, more generally, simply **cannot be affected** by the treatment, such as preaspirin headache pain or sex of the unit."*

Not all variables can be regarded as "covariates" in the above sense. A variable is a "covariate" when the study unit takes on the value *before* the treatment assignment. For instance, in the Medicare example, senior Americans' covariates include their race/ethnicity, age, preexisting comorbidities, etc, and these variables qualify as "covariates" because their values cannot be modified by the new technology or the absence of it in the cardiac surgery. To stress the "temporal" nature of "covariates," researchers in statistics, epidemiology, and social sciences often refer to them as "pretreatment variables." It is widely appreciated that "pretreatment variables" need to be adjusted for (Rosenbaum, 2002a), via matching, subclassification, or model-based methods, in order to obtain an unbiased causal estimate of a meaningful causal estimand; on the contrary, concomitant variables affected by the treatment, or the so-called "posttreatment variables" should not be adjusted for. In fact, adjusting for "posttreatment variables" often leads to "posttreatment bias" (Rosenbaum, 1984).

A central question then emerges when we start drawing parallels between two types of study units: human subjects and organized textual data:

> *What qualify as "covariates" or "pretreatment variables" when the study units involved in the analysis are textual data organized in one way or another?*

3. Immutable Characteristics and Causal Variables

To understand what variables quality as "pretreatment variables," we first need to articulate the "treatment" or more generally the "causal variable" under consideration. We would like to argue that many "attributes" of textual data, e.g., number of theorems in a conference paper in Example 1, are *not* appropriate causal variables *per se*.

To continue our analogy between human subjects and textual data, consider the role of "race", "ethnicity", or "gender identity" in human populations research. In a well-argued article, Holland (2008) pointed out that attributes of study units are not "causal variables" if they are immutable and do not lend themselves to "plausible states of counterfactuality." Put a different way, "it is critical that each unit be potentially exposable to any of the causes," and that it is not appropriate to talk about"causation without manipulation" (Holland, 1986). Moreover, if we imagine race/ethnicity, understood biologically as opposed to a social construct, as being "assigned" at a human subject's

conception, then almost every aspect of the human subject, including socioeconomic status, preexisting comorbid conditions, etc, is a "posttreatment variable" with gender being possibly the only exception (King, 1991; Gelman and Hill, 2006).

Organized textual data are different from human subjects in many ways; however, similar concerns persist. Take as an example the number of theorems in a conference paper. In their original article, Veitch et al. (2020) regarded "the sequence of words" in a conference paper as "covariates." From our discussion of race/ethnicity in human populations research, it becomes obvious that this notion of covariates is, at best, untenable and demands much clarification: many words in a conference paper are "affected" by "whether there is a theorem" or the "number of theorems," e.g., those discussing implications of the theorems in the abstract and the conclusion section. The same concern persists and clarification is much needed when embeddings from language models of "the sequence of words" are used as "covariates." Moreover, it is very difficult to even conceptualize exposing each study unit (i.e., a conference paper) to this "treatment:" it is nearly nonsensical to ask what would have happened to a pure empirical/dataset/benchmark paper in, say, EMNLP or ICLR, had it proved one or more theorems. Even when we restrict our attention to methodological papers backed up by simulations but not theorems, it is not clear what theorems these methodological papers would have proved in the first place. In fact, there are likely many different *versions* of this "treatment" of "adding one theorem," e.g., one version being "proving one theorem on the convergence rate of the proposed algorithm" and another version being "proving one theorem on the minimax lower bound of the problem," thus violating the *stable unit treatment value assumption* (SUTVA) as discussed in more detail in Rubin (1980, 1986). The key here is to think twice whether the "causal variable" under consideration lends itself to "plausible states of counterfactuality" (Holland, 2008). Even in some scenarios where adding or subtracting a theorem of a particular type can be envisioned, researchers need to articulate these scenarios and properly restrict their attention to meaningful study units.

Again, borrowing wisdom from classical causal inference reasoning, one acceptable, albeit not entirely satisfying, solution is to adjust for attributes that are definitely not affected by the number of theorems and drop any aspect of the article that are likely to be affected by it. In this spirit, researchers could adjust for certain "metadata" of the conference paper, e.g., topic listed, authors, and their affiliations, but *not* attributes of the article like the number of references, number of equations, sentiment and flow of the article, etc: these aspects are all likely to be affected by the number of theorems. This strategy is sometimes used in human populations research when the causal variable of interest is some immutable trait like "race/ethnicity" or "gender identity" (King, 1991), although better strategies (in our opinion) exist as we are ready to discuss.

4. Two Strategies Facilitating Framing Proper Causal Queries

We discuss two strategies that could be useful when researchers would like to properly frame the causal queries when study units are textual data.

4.1. Shifting from actual traits to perceptions of them

In a seminal paper, Greiner and Rubin (2011) discussed some prerequisites for the design and analysis of observational studies when the causal variable of interest is immutable. Greiner and Rubin (2011)'s key insight is that there are often *two actors* in causal inquiries concerning immutable traits: a study unit that possesses such immutable traits and a "decider" that *perceives* such traits.

The causal query is often concerned about decider's behaviors after perceiving study units' certain immutable trait. While it is not appropriate to "manipulate" study units' immutable traits, it is possible, both conceptually and in many cases operationally, to manipulate a decider's perception of such immutable traits; similar ideas and arguments were also presented by Fienberg and Haviland (2003) and Kaufman (2008). We illustrate this point using Veitch et al. (2020)'s second example.

Example 2 Consider comments from `reddit.com`, an online forum. Each post has a popularity score and the author of the post may (optionally) report their gender. The goal is to examine whether there is a direct effect of a "male" label on the score of the post. However, the author's gender may affect the text of the post, e.g., through tone, style, or topic choices, which also affects its score. Again, it is hoped to use the text (post content) to estimate the causal effect.

In this example, Veitch et al. (2020) imagine "manipulating" the gender identity of the *author* of a post and consider a mediation-type analysis (Imai et al., 2010). This is not a well-posed causal query in our opinion for similar reasons discussed before: it is highly speculative to ask what tone, style, or topic a person would have written in a post at a given time in the life course had the person had a different gender identity.

In fact, Example 2 could be re-formulated under the "perceived traits" framework outlined by Greiner and Rubin (2011). Suppose that some posts are associated with a hashtag or an icon indicating writers' gender identity, then we may stipulate that the treatment happens when the viewer of the post first perceives or reads the post. To stress, the gender identity of the writer of the post is *not* manipulated; viewers' perceptions are manipulated, presumably by manipulating the hashtag or icon attached to the post. By shifting from the actual trait of the study unit, i.e., gender identity of the author of the post, to viewers' perceptions of it, we are able to articulate the timing of the treatment, and it immediately becomes clear what variables qualify as "covariates" for this treatment: any aspect of the post that remains unchanged is a covariate that needs to be adjusted for. This includes pretty much every integral part of a post: words, emojis, pictures, tone, topic, etc. Moreover, it implies what variables do *not* qualify as "covariates:" in short, anything happening after viewers' first perception is a posttreatment and should not be adjustment for, e.g., any aspect of the comments left by the viewers. We note that many authors in the ML/NLP literature have been switching from traits of textual data to perceptions of them; see, e.g., Roberts et al. (2020), Feder et al. (2021), and Pryzant et al. (2021).

4.2. Shifting from one trait to a property/concept including many traits

The question remains as how to frame the causal query in Example 1. Shifting from the number of theorems to viewers' perception of theorems still seems insufficient in this example: it is unclear how to manipulate viewers' perception of theorems without also manipulating viewers' perceptions of other related textual data such as the number of mathematical equations or the complexity of mathematical notation. In this example, presumably researchers are interested in manipulating viewers' perceptions of an article's "mathematical rigor," and "number of theorems" is merely one aspect of this rather general concept.

Let us take a step back and think again the role of "race/ethnicity" in human populations research. A strict biological perspective would regard "race/ethnicity" as being "assigned" at a person's conception; on the contrary, an arguably more appropriate theory of race would hold that

distinction between so-called races are the products of many social forces including cultural, geographical, social, historical, and many other influences (Rutter and Tienda, 2005; Holland, 2008; Sen and Wasow, 2016). The concept of "race/ethnicity," under this "constructivist framework," is a complex consisting of many readily mutable traits, or "a bundle of sticks" (Sen and Wasow, 2016), and causal queries concerning "race/ethnicity" become better-posed from this constructivist perspective of race/ethnicity.

This strategy of further switching from the perception of a concept or one aspect of it to the perception of all constituent parts of the concept could be beneficial to framing causal queries in Example 1 and inferring the causal effect of a certain linguistic property of textual data, e.g., "politeness" as discussed in Pryzant et al., 2021. In Example 1, it seems most appropriate to define the causal variable as viewers' perceptions of an article's mathematical rigor. The notion of mathematical rigor could incorporate many aspects of the article, with the number of theorems being one component. Under this perspective, variables like "number of graphs," "flow of the paper," and "number of references to other theory papers" should be regarded as components of the "causal variable" rather than covariates to be adjusted for. As another example, suppose that we are interested in the causal effect of viewers' perceptions of a linguistic property of textual data, say "politeness." The key issue here is to identify and articulate constituent parts, i.e., "sticks" metaphorically, of the theory of politeness, i.e., the "politeness bundle." For instance, viewers' perceptions of an article/post/email's politeness may consist of their perceptions of its sentiment and formality in wording, among other things. Covariates to be adjusted for may include topic, authorship, brevity, among other things. This task of separating causal variables from covariates should be conducted on a case-by-case basis, and could benefit greatly from linguists' domain knowledge; see, e.g., Brown and Levinson (1978) for a comprehensive theory of politeness.

In any circumstances, we find it important to clearly define the causal variable, including its timing and/or constituent parts when necessary, and articulate what variables qualify as "covariates" for the causal variable under consideration. We recognize that some variables may fall into the grey areas of this "causal variable/covariates" dichotomy; however, acknowledging and articulating the ambiguity is itself a crucial step towards any meaningful scientific communication.

5. A Dichotomy of Covariates

Thus far, we have focused mostly on the timing of a treatment and the "causal variables/covariates" dichotomy when study units are textual data. In this section, we assume an appropriate causal query is framed, e.g., how perception of authors' gender identity affects the acceptance of a conference paper, and briefly discuss different types of covariates to be included in such an analysis.

In human populations research, covariates or pretreatment variables are often divided into "observed" and "unobserved" (Rosenbaum, 2002b). When study units are organized textual data, we find it meaningful to further divide observed covariates into two broad categories: "explicit observed covariates" that could be derived from the organized textual data at face value, e.g., the number of theorems/equations/figures in a conference paper, and "implicit observed covariates" that capture deeper aspects intrinsic to the textual data. Some concrete examples of implicit covariates include: bag-of-words embeddings such as Word2Vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014), and contextual embeddings such as BERT (Devlin et al., 2019) and SentenceBERT (Reimers and Gurevych, 2019); perceived sentiments, tones, and emotions from the text (Barbieri et al., 2020; Pérez et al., 2021); topic modeling and keyword summarizing (Xie et al., 2015; Blei and Lafferty,

2007; Ramage et al., 2009; Wang et al., 2020; Santosh et al., 2020); evaluated trustworthiness of the claims made (Nadeem et al., 2019; Zhang et al., 2021b); temporal relationships and semantic relationships of events mentioned (Zhou et al., 2021; Han et al., 2021); commonsense knowledge reasoning (such as complex relations between events, consequences, and predictions) based on the text (Chaturvedi et al., 2017; Speer et al., 2017; Hwang et al., 2021; Jiang et al., 2021). These are by no means exhaustive; nor are they necessary for each and every causal query. It is always essential to incorporate domain knowledge before determining an appropriate set of covariates.

The central assumption to drawing causal inference from observational data, namely the "treatment ignorability assumption" (Rosenbaum and Rubin, 1983), also known as the "no unmeasured confounding assumption," needs to take into account both explicit and implicit observed covariates, so that the treatment assignment is closer to being randomly assigned within strata defined by both explicit and implicit observed covariates. Moreover, there is an additional layer of complexity as some "implicit observed covariates" (e.g., sentiment of the article) are themselves output from some pretrained language models on certain domains of texts (e.g., twitters) and likely to be measured with error or summarized insufficiently.

Unobserved covariates have a particularly important role in drawing causal inference from non-randomized observational data as Rosenbaum (2018) reminds us that "at the end of the day, scientific arguments about what causes what are almost invariably arguments about some covariate that was not measured or could not be measured." The primary appeal of a randomized controlled experiment is that randomization stochastically balances both observed and unobserved covariates, while conclusions derived from observational data can always be challenged on the basis of unmeasured confounding bias. For instance, Sir Ronald Fisher challenged the causal interpretation of the association between smoking and lung cancer by pointing to the possibility of a genetic variant making a person simultaneously prone to smoking and susceptible to lung cancer (Fisher, 1958). In the context of causal inference with textual data, this concern of unmeasured confounding is even more pronounced as relevant "covariates" like semantics may be buried under the ocean of words and difficult to be fully recovered. Methods that aim to address unmeasured confounding concerns, e.g., sensitivity analysis (Rosenbaum, 2002b, 2010; VanderWeele and Ding, 2017; Veitch and Zaveri, 2020), negative control methods (Lipsitch et al., 2010), and instrumental variable methods (Angrist et al., 1996), should be explored and more actively employed in the context of causal inference with textual data.

6. Conclusion

Properly framing a causal query with textual data is challenging. One general caveat that Angrist and Pischke (2008) gave to practitioners of causal inference is that research questions for which there are no experimental analogies (even hypothetical ones in a world with unlimited time, budgets, and omniscient powers) may be fundamentally unidentified or at least ill-posed. We encourage researchers working with textual data to always (i) identify study units, (ii) articulate the treatment or causal variable including its timing or its constituent parts when appropriate, and (iii) make an argument for each covariate to be adjusted for in the analysis. Causal inference is always an ambitious task; carefully going through steps (i)−(iii), though not a panacea, helps at least expose potential fallacies and facilitate conversations and critiques. We hope that our discussion in the article could raise awareness that although methodological development is important, it is equally

important to pay attention to many fundamentals, including key concepts and basic principles, when conducting causal inference with textual data.

Acknowledgements

This work was supported in part by NSF through CCF-1934876 and ONR Contract N00015-19-1-2620. We would like to thank Dan Roth, Dylan S. Small, and Weijie J. Su for stimulating discussions and helpful feedback on this manuscript.

References

Joshua D Angrist and Jörn-Steffen Pischke. *Mostly Harmless Econometrics*. Princeton University Press, 2008.

Joshua D Angrist, Guido W Imbens, and Donald B Rubin. Identification of causal effects using instrumental variables. *Journal of the American statistical Association*, 91(434):444–455, 1996.

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa-Anke, and Leonardo Neves. TweetEval:Unified Benchmark and Comparative Evaluation for Tweet Classification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2020.

David M. Blei and John D. Lafferty. A correlated topic model of science. *The Annals of Applied Statistics*, 1:17–35, 2007.

Penelope E. Brown and S. Levinson. Universals in Language Usage: Politeness Phenomena. 1978.

Snigdha Chaturvedi, Haoruo Peng, and Dan Roth. Story Comprehension for Predicting What Happens Next. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2017.

William G Cochran and S Paul Chambers. The planning of observational studies of human populations. *Journal of the Royal Statistical Society. Series A (General)*, 128(2):234–266, 1965.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 2019.

Naoki Egami, Christian J. Fong, Justin Grimmer, Margaret E. Roberts, and Brandon M. Stewart. How to make causal inferences using texts, 2018.

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. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond, 2021.

Stephen E Fienberg and Amelia M Haviland. Discussion of statistics and causal inference: A review. *Test*, 12:319–327, 2003.

Ronald A Fisher. Cancer and smoking. *Nature*, 182(4635):596–596, 1958.

Andrew Gelman and Jennifer Hill. *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press, 2006.

James D Greiner and Donald B Rubin. Causal effects of perceived immutable characteristics. *Review of Economics and Statistics*, 93 (3):775–785, 2011.

Rujun Han, I-Hung Hsu, Jiao Sun, Julia Baylon, Qiang Ning, Dan Roth, and Nanyun Peng. ESTER: A Machine Reading Comprehension Dataset for Reasoning about Event Semantic Relations. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2021.

James J Heckman. Rejoinder: Response to Sobel. *Sociological Methodology*, 35(1):135–150, 2005.

Paul W Holland. Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960, 1986.

Paul W Holland. Causation and race. *White logic, white methods: Racism and methodology*, pages 93–109, 2008.

Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. Comet-atomic 2020: On symbolic and neural commonsense knowledge graphs. In *Proceedings of the Conference on Artificial Intelligence*, 2021.

Kosuke Imai, Luke Keele, and Dustin Tingley. A general approach to causal mediation analysis. *Psychological Methods*, 15(4):309, 2010.

Guido W. Imbens. Potential outcome and directed acyclic graph approaches to causality: Relevance for empirical practice in economics. *Journal of Economic Literature*, 58(4):1129–79, December 2020. doi: 10.1257/jel.20191597.

Guido W Imbens and Donald B Rubin. *Causal inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press, 2015.

Liwei Jiang, Antoine Bosselut, Chandra Bhagavatula, and Yejin Choi. "I'm not mad": Commonsense implications of negation and contradiction. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4380–4397, Online, June 2021. Association for Computational Linguistics.

Jay S Kaufman. Epidemiologic analysis of racial/ethnic disparities: some fundamental issues and a cautionary example. *Social Science & Medicine*, 66(8):1659–1669, 2008.

Katherine A Keith, David Jensen, and Brendan O'Connor. Text and causal inference: A review of using text to remove confounding from causal estimates. *arXiv preprint arXiv:2005.00649*, 2020.

Gary King. "Truth" is stranger than prediction, more questionable than causal inference. *American Journal of Political Science*, 35(4): 1047–1053, 1991.

Marc Lipsitch, Eric Tchetgen Tchetgen, and Ted Cohen. Negative controls: a tool for detecting confounding and bias in observational studies. *Epidemiology (Cambridge, Mass.)*, 21(3):383, 2010.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *ArXiv*, abs/1907.11692, 2019.

Emily J MacKay, Bo Zhang, Siyu Heng, Ting Ye, Mark D Neuman, John G Augoustides, Jared W Feinman, Nimesh D Desai, and Peter W Groeneveld. Association between transesophageal echocardiography and clinical outcomes after coronary artery bypass graft surgery. *Journal of the American Society of Echocardiography*, 34(6):571–581, 2021.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

Moin Nadeem, Wei Fang, Brian Xu, Mitra Mohtarami, and James R. Glass. Fakta: An automatic end-to-end fact checking system. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 2019.

Jerzy S Neyman. On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Annals of Agricultural Sciences*, 10:1–51, 1923.

Qiang Ning, Zhili Feng, Hao Wu, and Dan Roth. Joint Reasoning for Temporal and Causal Relations. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 2278–2288, Melbourne, Australia, 7 2018. Association for Computational Linguistics.

Judea Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995. ISSN 00063444. URL http://www.jstor.org/stable/2337329.

Judea Pearl. *Causality*. Cambridge University Press, 2009.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. GloVe: Global vectors for word representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1532–1543, 2014.

Jonas Peters, Dominik Janzing, and Bernhard Schlkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. The MIT Press, 2017. ISBN 0262037319.

Reid Pryzant, Dallas Card, Dan Jurafsky, Victor Veitch, and Dhanya Sridhar. Causal effects of linguistic properties. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4095–4109, 2021.

Juan Manuel Pérez, Juan Carlos Giudici, and Franco Luque. pysentimiento: A python toolkit for sentiment analysis and social nlp tasks, 2021.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D. Manning. Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2009.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019.

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

Paul Rosenbaum. *Observation and experiment*. Harvard University Press, 2018.

Paul R Rosenbaum. The consequences of adjustment for a concomitant variable that has been affected by the treatment. *Journal of the Royal Statistical Society: Series A (General)*, 147(5):656–666, 1984.

Paul R Rosenbaum. Covariance adjustment in randomized experiments and observational studies. *Statistical Science*, 17(3):286–327, 2002a.

Paul R Rosenbaum. *Observational Studies*. Springer, 2002b.

Paul R Rosenbaum. *Design of Observational Studies*. Springer, 2010.

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

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

Donald B Rubin. Randomization analysis of experimental data: The fisher randomization test comment. *Journal of the American Statistical Association*, 75(371):591–593, 1980.

Donald B. Rubin. Statistics and causal inference: Comment: Which ifs have causal answers. *Journal of the American Statistical Association*, 81(396):961–962, 1986. ISSN 01621459.

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

Michael Rutter and Marta Tienda. The multiple facets of ethnicity. *Ethnicity and Causal Mechanisms*, 50:79, 2005.

T.y.s.s Santosh, Debarshi Kumar Sanyal, Plaban Kumar Bhowmick, and Partha Pratim Das. SaSAKE: Syntax and semantics aware keyphrase extraction from research papers. In *Proceedings of the International Conference on Computational Linguistics*, pages 5372–5383. International Committee on Computational Linguistics, 2020.

Maarten Sap, Vered Shwartz, Antoine Bosselut, Yejin Choi, and Dan Roth. Commonsense reasoning for natural language processing. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 27–33, 2020.

Maya Sen and Omar Wasow. Race as a bundle of sticks: Designs that estimate effects of seemingly immutable characteristics. *Annual Review of Political Science*, 19:499–522, 2016.

Vered Shwartz, Peter West, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Unsupervised commonsense question answering with self-talk. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2020.

Robyn Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the Conference on Artificial Intelligence*, 2017.

Dhanya Sridhar and Lise Getoor. Estimating causal effects of tone in online debates. In *IJCAI*, 2019.

Tyler J VanderWeele and Peng Ding. Sensitivity analysis in observational research: introducing the e-value. *Annals of Internal Medicine*, 167(4):268–274, 2017.

Victor Veitch and Anisha Zaveri. Sense and sensitivity analysis: Simple post-hoc analysis of bias due to unobserved confounding. *arXiv preprint arXiv:2003.01747*, 2020.

Victor Veitch, Dhanya Sridhar, and David Blei. Adapting text embeddings for causal inference. In *Conference on Uncertainty in Artificial Intelligence*, pages 919–928. PMLR, 2020.

Yansen Wang, Zhen Fan, and Carolyn Rose. Incorporating multimodal information in open-domain web keyphrase extraction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1790–1800. Association for Computational Linguistics, 2020.

Pengtao Xie, Diyi Yang, and Eric Xing. Incorporating word correlation knowledge into topic modeling. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 725–734. Association for Computational Linguistics, 2015.

Bo Zhang, Siyu Heng, Emily J MacKay, and Ting Ye. Bridging preference-based instrumental variable studies and cluster-randomized encouragement experiments: Study design, noncompliance, and average cluster effect ratio. *Biometrics (in press)*, 2021a.

Jiayao Zhang, Hongming Zhang, Dan Roth, and Weijie J. Su. Causal inference principles for reasoning about commonsense causality, 2022. URL https://arxiv.org/abs/2202.00436.

Yi Zhang, Zachary G. Ives, and Dan Roth. What is Your Article Based On? Inferring Fine-Grained Provenance. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2021b.

Ben Zhou, Kyle Richardson, Qiang Ning, Tushar Khot, Ashish Sabharwal, and Dan Roth. Temporal Reasoning on Implicit Events from Distant Supervision. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 2021.