Text and Causal Inference: A Review of Using Text to Remove Confounding from Causal Estimates

Katherine A. Keith, David Jensen, and **Brendan O'Connor**
College of Information and Computer Sciences
University of Massachusetts Amherst
{kkeith,jensen,brenocon}@cs.umass.edu

Abstract

Many applications of computational social science aim to infer causal conclusions from non-experimental data. Such *observational* data often contains *confounders*, variables that influence both potential causes and potential effects. Unmeasured or *latent* confounders can bias causal estimates, and this has motivated interest in measuring potential confounders from observed text. For example, an individual's entire history of social media posts or the content of a news article could provide a rich measurement of multiple confounders. Yet, methods and applications for this problem are scattered across different communities and evaluation practices are inconsistent. This review is the first to gather and categorize these examples and provide a guide to data-processing and evaluation decisions. Despite increased attention on adjusting for confounding using text, there are still many open problems, which we highlight in this paper.

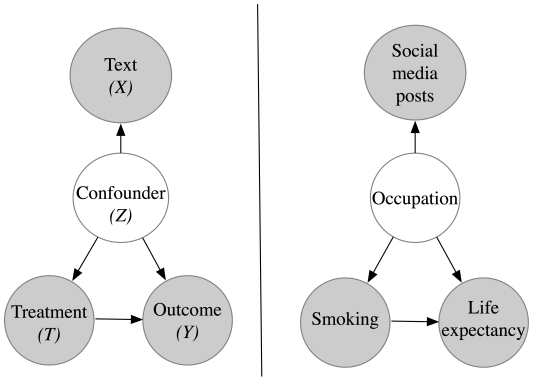

Figure 1: *Left:* A causal diagram for text that encodes causal confounders, the setting that is focus of this review paper. The major assumption is that latent confounders can be *measured* from text and those confounder measurements can be used in causal adjustments. *Right:* An example application in which practitioner does not have access to the confounding variable, *occupation*, in structured form but can measure confounders from unstructured text (e.g. an individual's social media posts).

1 Introduction

In contrast to descriptive or predictive tasks, causal inference aims to understand how *intervening* on one variable affects another variable (Holland, 1986; Pearl, 2000; Morgan and Winship, 2015; Imbens and Rubin, 2015; Hernán and Robins, 2020). Specifically, many applied researchers aim to estimate the size of a specific causal effect, the effect of a single *treatment* variable on an *outcome* variable. However, a major challenge in causal inference is addressing *confounders*, variables that influence both treatment and outcome. For example, consider estimating the size of the causal effect of smoking (treatment) on life expectancy (outcome). Occupation is a potential confounder that may influence both the propensity to smoke and life expectancy. Estimating the effect of treatment on outcome without accounting for this confounding could result in strongly biased estimates and thus invalid causal conclusions.

To eliminate confounding bias, one approach is to perform randomized controlled trials (RCTs) in which researchers randomly assign treatment. Yet, in many research areas such as healthcare, education, or economics, randomly assigning treatment is either infeasible or unethical. For instance, in our running example, one cannot ethically randomly assign participants to smoke since this could expose them to major health risks. In such cases, researchers instead use observational data and adjust for the confounding bias statistically with methods such as matching, propensity score weighting, or regression adjustment (§5).

In causal research about human behavior and society, there are potentially many latent confounding variables that can be measured from unstructured

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 5332–5344
July 5 - 10, 2020. ©2020 Association for Computational Linguistics

text data. Text data could either (a) serve as a surrogate for potential confounders; or (b) the language of text itself could be a confounder. Our running example is an instance of text as a surrogate: a researcher may not have a record of an individual's occupation but could attempt to measure this variable from the individual's entire history of social media posts (see Fig. 1). An example of text as a direct confounder: the linguistic content of social media posts could influence censorship (treatment) and future posting rates (outcome) (Roberts et al., 2020).

A challenging aspect of this research design is the high-dimensional nature of text. Other work has explored general methods for adjusting for high-dimensional confounders (D'Amour et al., 2017; Rassen et al., 2011; Louizos et al., 2017; Li et al., 2016; Athey et al., 2017). However, text data differ from other high-dimensional data-types because intermediate confounding adjustments can be read and evaluated by humans (§6) and designing meaningful representations of text is still an open research question.[1] Even when applying simple adjustment methods, a practitioner must first transform text into a lower-dimensional representation via, for example, filtered word counts, lexicon indicators, topic models, or embeddings (§4). An additional challenge is that empirical evaluation in causal inference is still an open research area (Dorie et al., 2019; Gentzel et al., 2019) and text adds to the difficulty of this evaluation (§7).

We narrow the scope of this paper to review methods and applications with text data as a causal *confounder*. In the broader area of text and causal inference, work has examined text as a mediator (Veitch et al., 2019), text as treatment (Fong and Grimmer, 2016; Egami et al.; Wood-Doughty et al., 2018; Tan et al., 2014), text as outcome (Egami et al.), causal discovery from text (Mani and Cooper, 2000), and predictive (Granger) causality with text (Balashankar et al., 2019; del Prado Martin and Brendel, 2016; Tabari et al., 2018).

Outside of this prior work, there has been relatively little interaction between natural language processing (NLP) research and causal inference. NLP has a rich history of applied modeling and diagnostic pipelines that causal inference could draw upon. Because applications and methods for text as a confounder have been scattered across many different communities, this review paper aims to gather and unify existing approaches and to concurrently serve three different types of researchers and their respective goals:

- **For applied practitioners,** we collect and categorize applications with text as a causal confounder (Table 1 and §2), and we provide a flowchart of analysts' decisions for this problem setting (Fig. 2).
- **For causal inference researchers working with text data,** we highlight recent work in representation learning in NLP (§4) and caution that this is still an open research area with questions of the sensitivity of effects to choices in representation. We also outline existing interpretable evaluation methods for adjustments of text as a causal confounder (§6).
- **For NLP researchers working with causal inference,** we summarize some of the most-used causal estimators that condition on confounders: matching, propensity score weighting, regression adjustment, doubly-robust methods, and causally-driven representation learning (§5). We also discuss evaluation of methods with constructed observational studies and semi-synthetic data (§7).

2 Applications

In Table 1, we gather and summarize applications that use text to adjust for potential confounding. This encompasses both (a) text as a surrogate for confounders, or (b) the language itself as confounders.[2]

As an example, consider Kiciman et al. (2018) where the goal is to estimate the size of the causal effect of alcohol use (treatment) on academic success (outcome) for college students. Since randomly assigning college students to binge drink is not feasible or ethical, the study instead uses observational data from Twitter, which also has the advantage of a large sample size of over sixty-three thousand students. They use heuristics to identify

[1] For instance, there have been four workshops on representation learning at major NLP conferences in the last four years (Blunsom et al., 2016, 2017; Augenstein et al., 2018, 2019).

[2] We acknowledge that Table 1 is by no means exhaustive. To construct Table 1, we started with three seed papers: Roberts et al. (2020), Veitch et al. (2019), and Wood-Doughty et al. (2018). We then examined papers cited by these papers, papers that cited these papers, and papers published by the papers' authors. We repeated this approach with the additional papers we found that adjusted for confounding with text. We also examined papers matching the query "causal" or "causality" in the ACL Anthology.

Paper	Treatment	Outcome(s)	Confounder	Text data	Text rep.	Adjustment method
Johansson et al. (2016)	Viewing device (mobile or desktop)	Reader's experience	News content	News	Word counts	Causal-driven rep. learning
De Choudhury et al. (2016)	Word use in mental health community	User transitions to post in suicide community	Previous text written in a forum	Social media (Reddit)	Word counts	Stratified propensity score matching
De Choudhury and Kiciman (2017)	Language of comments	User transitions to post in suicide community	User's previous posts and comments received	Social media (Reddit)	Unigrams and bigrams	Stratified propensity score matching
Falavarjani et al. (2017)	Exercise (Foursquare checkins)	Shift in topical interest on Twitter	Pre-treatment topical interest shift	Social media (Twitter, Foursquare)	Topic models	Matching
Olteanu et al. (2017)	Current word use	Future word use	Past word use	Social media (Twitter)	Top unigrams and bigrams	Stratified propensity score matching
Pham and Shen (2017)	Group vs. individual loan requests	Time until borrowers get funded	Loan description	Microloans (Kiva)	Pre-trained embeddings + neural networks	A-IPTW, TMLE
Kiciman et al. (2018)	Alcohol mentions	College success (e.g. study habits, risky behaviors, emotions)	Previous posts	Social media (Twitter)	Word counts	Stratified propensity score matching
Sridhar et al. (2018)	Exercise	Mood	Mood triggers	Users' text on mood logging apps	Word counts	Propensity score matching
Saha et al. (2019)	Self-reported usage of psychiatric medication	Mood, cognition, depression, anxiety, psychosis, and suicidal ideation	Users' previous posts	Social media (Twitter)	Word counts + lexicons + supervised classifiers	Stratified propensity score matching
Sridhar and Getoor (2019)	Tone of replies	Changes in sentiment	Speaker's political ideology	Debate transcripts	Topic models + lexicons	Regression adjustment, IPTW, A-IPTW
Veitch et al. (2019)	Presence of a theorem	Rate of acceptance	Subject of the article	Scientific articles	BERT	Causal-driven rep. learning + Regression adjustment, TMLE
Roberts et al. (2020)	Perceived gender of author	Number of citations	Content of article	International Relations articles	Topic models + propensity score	Coarsened exact matching
Roberts et al. (2020)	Censorship	Subsequent censorship and posting rate	Content of posts	Social media (Weibo)	Topic models + propensity score	Coarsened exact matching

Table 1: Example applications that infer the causal effects of treatment on outcome by measuring confounders (unobserved) from text data (observed). In doing so, these applications choose a representation of text (text rep.) and a method to adjust for confounding.

the Twitter accounts of college-age students and extract alcohol mentions and indicators of college success (e.g., study habits, risky behaviors, and emotions) from their Twitter posts. They condition on an individual's previous posts (temporally previous to measurements of treatment and outcome) as confounding variables since they do not have demographic data. They represent text as word counts and use stratified propensity score matching to adjust for the confounding bias. The study finds the effects of alcohol use include decreased mentions of study habits and positive emotions and increased mentions of potentially risky behaviors.

Text as a surrogate for confounders. Traditionally, causal research that uses human subjects as the unit of analysis would infer demographics via surveys. However, with the proliferation of the web and social media, social research now includes large-scale observational data that would be challenging to obtain using surveys (Salganik, 2017). This type of data typically lacks demographic information but may contain large amounts of text written by participants from which demographics can be extracted. In this space, some researchers are specific about the confounders they want to extract such as an individual's ideology (Sridhar and Getoor, 2019) or mood (Sridhar et al., 2018). Other researchers condition on all the text they have available and assume that low-dimensional summaries capture all possible confounders. For example, researchers might assume that text encodes all possible confounders between alcohol use and college success (Kiciman et al., 2018) or psychiatric medication and anxiety (Saha et al., 2019). We dissect and comment on this assumption in Section 8.

Open problems: NLP systems have been shown to be inaccurate for low-resource languages (Duong et al., 2015), and exhibit racial and gender disparity (Blodgett and O'Connor, 2017; Zhao et al., 2017). Furthermore, the ethics of predicting psychological indicators, such as mental health status, from text are questionable (Chancellor et al., 2019). It is unclear how to mitigate these disparities when trying to condition on demographics from text and how NLP errors will propagate to causal estimates.

Language as confounders. There is growing interest in measuring language itself (e.g. the sentiment or topical content of text) as causal confounders. For example, Roberts et al. (2020) examine how the perceived gender of an author affects the number of citations that an article receives. However, an article's topics (the confounders) are likely to influence the perceived gender of its author (reflecting an expectation that women write about certain topics) and the number of citations of that article ("hotter" topics will receive more

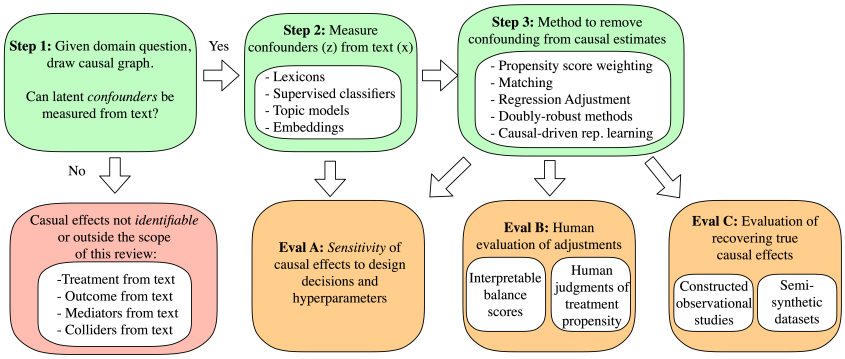

Figure 2: This chart is a guide to design decisions for applied research with causal confounders from text. *Step 1:* Encode domain assumptions by drawing a causal diagram (§3). If the application does not use text to measure latent *confounders*, the causal effects are not identifiable or the application is outside the scope of this review. *Step 2:* Use NLP to measure confounders from text (§4). *Step 3:* Choose a method that adjusts for confounding in causal estimates (§5). Evaluation should include *(A)* sensitivity analysis (§4), *(B)* human evaluation of adjustments when appropriate (§6), and *(C)* evaluation of recovering the true causal effects (§7).

citations). Other domains that analyze language as a confounder include news (Johansson et al., 2016), social media (De Choudhury et al., 2016; Olteanu et al., 2017), and loan descriptions (Pham and Shen, 2017). See Section 4 for more discussion on the challenges and open problems of inferring these latent aspects of language.

3 Estimating causal effects

Two predominant causal inference frameworks are *structural causal models (SCM)* (Pearl, 2009b) and *potential outcomes* (Rubin, 1974, 2005), which are complementary and theoretically connected (Pearl, 2009b; Richardson and Robins, 2013; Morgan and Winship, 2015). While their respective goals substantially overlap, methods from structural causal models tend to emphasize conceptualizing, expressing, and reasoning about the effects of possible causal relationships among variables, while methods from potential outcomes tend to emphasize estimating the size or strength of causal effects.

3.1 Potential outcomes framework

In the ideal causal experiment, for each each unit of analysis, i (e.g., a person), one would like to measure the outcome, y_i (e.g., an individual's life expectancy), in both a world in which the unit received treatment, $t_i = 1$ (e.g., the person smoked), as well as in the counterfactual world in which the same unit did not receive treatment, $t_i = 0$ (e.g the same person did not smoke).[3] A fundamental challenge of causal inference is that one cannot simultaneously observe treatment and non-treatment for a single individual (Holland, 1986).

The most common population-level estimand of interest is the *average treatment effect (ATE)*.[4] In the absence of confounders, this is simply the difference in means between the treatment and control groups, $\tau = \mathbb{E}(y_i|t_i = 1) - \mathbb{E}(y_i|t_i = 0)$, and the "unadjusted" or "naive" estimator is

$$\hat{\tau}_{\text{naive}} = \frac{1}{n_1} \sum_{i:t_i=1} y_i - \frac{1}{n_0} \sum_{j:t_j=0} y_j \qquad (1)$$

where n_1 is the number of units that have received treatment and n_0 is the number of units that have not received treatment. However, this equation will be biased if there are confounders, z_i, that influence both treatment and outcome.

3.2 Structural causal models framework

Structural causal models (SCMs) use a graphical formalism that depicts nodes as random variables and directed edges as the direct causal dependence between these variables. The typical estimand of choice for SCMs is the probability distribution of an outcome variable Y given an intervention on a treatment variable T:

$$P(Y \mid do(T = t)) \qquad (2)$$

in which the *do*-notation represents intervening to set variable T to the value t and thereby removing all incoming arrows to the variable T.

Identification. In most cases, Equation 2 is *not* equal to the ordinary conditional distribution

[3] In this work, we only address binary treatments, but multi-value treatments are also possible (e.g., Imbens (2000)).

[4] Other estimands include the average treatment effect on the treated (ATT) and average treatment effect on the control (ATC) (Morgan and Winship, 2015)

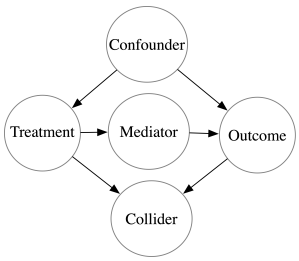

Figure 3: A causal diagram showing common causal relationships.

$P(Y \mid T = t)$ since the latter is simply filtering to the sub-population and the former is changing the underlying data distribution via intervention. Thus, for observational studies that lack intervention, one needs an *identification strategy* in order to represent $P(Y \mid do(T = t))$ in terms of distributions of observed variables. One such identification strategy (assumed by the applications throughout this review) is the *backdoor criterion* which applies to a set of variables, $\mathcal{S}$, if they (i) block every backdoor path between treatment and outcome, and (ii) no node in $\mathcal{S}$ is a descendant of treatment. Without positive identification, the causal effects cannot be estimated and measuring variables from text is a secondary concern.

Drawing the causal graph. Causal graphs help clarify which variables should and should not be conditioned on. The causal graphs in Figure 3 illustrate how the direction of the arrows differentiates confounder, collider, and mediator variables. Identifying the differences in these variables is crucial since, by *d-separation*, conditioning on a confounder will block the treatment-confounder-outcome path, removing bias. By contrast, conditioning on a collider can create dependence between treatment-collider-outcome[5] (Pearl, 2009a) potentially introducing more bias (Montgomery et al., 2018; Elwert and Winship, 2014). Mediator variables require a different set of adjustments than confounders to find the "natural direct effect" between treatment and outcome (VanderWeele, 2015; Pearl, 2014). A practitioner typically draws a causal graph by explicitly encoding theoretical and domain assumptions as well as the results of prior data analyses.[6]

Open Problems: When could text potentially encode confounders and colliders simultaneously? If so, is it possible to use text to adjust exclusively for confounders?

4 Measuring confounders via text

After drawing the causal graph, the next step is to use available text data to recover latent confounders. Some approaches *pre-specify* the confounders of interest and measure them from text, $P(z \mid x)$. Others learn confounders *inductively* and use a low-dimensional representation of text as the confounding variable z in subsequent causal adjustments.

Pre-specified confounders. When a practitioner can specify confounders they want to measure from text (e.g., extracting "occupation" from text in our smoking example), they can use either (1) *lexicons* or (2) trained *supervised classifiers* as the instrument of measurement. Lexicons are word lists that can either be hand-crafted by researchers or taken off-the-shelf. For example, Saha et al. (2019) use categories of the Linguistic Inquiry and Word Count (LIWC) lexicon (Pennebaker et al., 2001) such as tentativeness, inhibition, and negative affect, and use indicators of these categories in the text as confounders. Trained supervised classifiers use annotated training examples to predict confounders. For instance, Saha et al. (2019) also build machine learning classifiers for users' mental states (e.g., depression and anxiety) and apply these classifiers on Twitter posts that are temporally prior to treatment. If these classifiers *accurately* recover mental states and there are no additional latent confounders, then conditioning on the measured mental states renders treatment independent of potential outcomes.

Open problems: Since NLP methods are still far from perfectly accurate, how can one mitigate error that arises from *approximating* confounding variables? Closely related to this question is *effect restoration* which addresses error from using proxy variables (e.g., a father's occupation) in place of true confounders (e.g, socioeconomic status) (Kuroki and Pearl, 2014; Oktay et al., 2019). Wood-

[5] In Pearl et al. (2016)'s example of a collider, suppose scholarships at a college are only given to two types of students: those with unusual musical talents and high grade point averages. In the general population, musical and academic talent are independent. However, if one discovers a person is on a scholarship (conditioning on the collider) then knowing a person lacks musical talent tells us that they are extremely likely to have a high GPA.

[6] See Morgan and Winship (2015) pgs. 33-34 on both the necessity and difficulty of specifying a causal graph for applied social research. *Time-ordering* can be particularly helpful when encoding causal relationships (for instance, there cannot be an arrow pointing from variable A to variable B if B preceded A in time).

Doughty et al. (2018) build upon effect restoration for causal inference with text classifiers, but there are still open problems in accounting for error arising from other text representations and issues of calibration (Nguyen and O'Connor, 2015) and prevalence estimation (Card and Smith, 2018; Keith and O'Connor, 2018) in conjunction with NLP. Ideas from the large literature on measurement error models may also be helpful (Fuller, 1987; Carroll et al., 2006; Buonaccorsi, 2010).

Inductively derived confounders. Other researchers *inductively* learn confounders in order to condition on *all* aspects of text, known and unknown. For example, some applications condition on the entirety of news (Johansson et al., 2016) or scientific articles (Veitch et al., 2019; Roberts et al., 2020). This approach typically summarizes textual information with text representations common in NLP. Ideally, this would encode all aspects of language (meaning, topic, style, affect, etc.), though this is an extremely difficult, open NLP problem. Typical approaches include the following. (1) *Bag-of-words* representations discard word order and use word counts as representations. (2) *Topic models* are generative probabilistic models that learn latent topics in document collections and represent documents as distributions over topics (Blei et al., 2003; Boyd-Graber et al., 2014; Roberts et al., 2014). (3) *Embeddings* are continuous, vector-based representations of text. To create vector representations of longer texts, off-the-shelf word embeddings such as *word2vec* (Mikolov et al., 2013) or *GloVe* (Pennington et al., 2014) or combined via variants of weighted averaging (Arora et al., 2017) or neural models (Iyyer et al., 2015; Bojanowski et al., 2017; Yang et al., 2016). (4) Recently, fine-tuned, large-scale neural language models such as BERT (Devlin et al., 2019) have achieved state-of-the-art performance on semantic benchmarks, and are now used as text representations. Each of these text representations is a real-valued vector that is used in place of the confounder, z, in a causal adjustment method (§5)

Open problems: Estimates of causal effects are contingent on the "garden of forking paths" of data analysis, meaning any "paths" an analyst did not take could have resulted in different conclusions (Gelman and Loken, 2013). For settings with causal confounders from text, the first fork is the choice of representation (e.g., topic models or embeddings) and the second fork is the pre-processing and hyperparameter decisions for the chosen representations.

We highlight that these decisions have been shown to alter results in predictive tasks. For instance, studies have shown that pre-processing decisions dramatically change topic models (Denny and Spirling, 2018; Schofield et al., 2017); embeddings are sensitive to hyperparameter tuning (Levy et al., 2015) and the construction of the training corpus (Antoniak and Mimno, 2018); and fine-tuned language model performance is sensitive to random restarts (Phang et al., 2018). Thus, reporting *sensitivity analysis* of the causal effects from these decisions seems crucial: how robust are the results to variations in modeling specifications?

5 Adjusting for confounding bias

Given a set of variables Z that satisfy the backdoor criterion (§3.2), one can use the *backdoor adjustment* to estimate the causal quantity of interest,

$$P(Y = y \mid do(T = t)) = \int P(Y = y \mid T = t, Z = z)\, P(Z = z)\, dz \quad (3)$$

Conditioning on all confounders is often impractical in high-dimensional settings such as those found in natural language. We provide an overview of methods used by applications in this review that approximate such conditioning, leading to unbiased estimates of treatment effect; however, we acknowledge this is not an exhaustive list of methods and direct readers to more extensive guides (Morgan and Winship, 2015; Athey et al., 2017).

Open problems: Causal studies typically make an assumption of *overlap*, also known as *common support* or *positivity*, meaning that any individual has a non-zero probability of assignment to each treatment condition for all possible values of the covariates: $\forall z,\ 0 < P(T = 1 \mid Z = z) < 1$. D'Amour et al. (2017) show that as the dimensionality of covariates grows, strict overlap converges to zero. What are the implications of these results for high-dimensional text data?

5.1 Propensity scores

A *propensity score* estimates the conditional probability of treatment given a set of possible confounders (Rosenbaum and Rubin, 1984, 1983; Caliendo and Kopeinig, 2008). The true model of treatment assignment is typically unknown so one must estimate the propensity score from data (e.g., from a logistic regression model),

$$\pi \equiv P(T = 1 \mid Z). \quad (4)$$

Inverse Probability of Treatment Weighting (IPTW) assigns a weight to each unit based on the propensity score (Lunceford and Davidian, 2004),

$$w_i = t_i/\hat{\pi}_i + (1 - t_i)/(1 - \hat{\pi}_i), \quad (5)$$

thus emphasizing, for example, treated units that were originally unlikely to be treated ($t_i = 1$, low π_i). The ATE is calculated with weighted averages between the treatment and control groups,[7]

$$\hat{\tau}_{\text{IPTW}} = \frac{1}{n_1} \sum_{i:t_i=1} w_i y_i - \frac{1}{n_0} \sum_{j:t_j=0} w_j y_j \quad (6)$$

5.2 Matching and stratification

Matching aims to create treatment and control groups with similar confounder assignments; for example, grouping units by observed variables (e.g., age, gender, occupation), then estimating effect size within each stratum (Stuart, 2010). *Exact matching* on confounders is ideal but nearly impossible to obtain with high-dimensional confounders, including those from text. A framework for matching with text data is described by Mozer et al. (2020) and requires choosing: (1) a text representation (§4); (2) a distance metric (cosine, Euclean, absolute difference in propensity score etc.); and (3) a matching algorithm. As Stuart (2010) describes, the matching algorithm involves additional decisions about (a) greedy vs. optimal matching; (b) number of control items per treatment item; (c) using calipers (thresholds of maximum distance); and (d) matching with or without replacement. *Coarsened exact matching (CEM)* matches on discretized raw values of the observed confounders (Iacus et al., 2012).

Instead of directly matching on observed variables, *stratified propensity-score matching* partitions propensity scores into intervals (strata) and then all units are compared within a single strata (Caliendo and Kopeinig, 2008). *Stratification* is also known as interval matching, blocking, and subclassification.

Once the matching algorithm is implemented, counterfactuals (estimated potential outcomes) are obtained from the matches $\mathcal{M}_i$ for each unit i:

$$\hat{y}_i(k) = \begin{cases} y_i & \text{if } t_i = k \\ \frac{1}{|\mathcal{M}_i|} \sum_{j \in \mathcal{M}_i} y_j & \text{if } t_i \neq k \end{cases} \quad (7)$$

which is plugged into the matching estimator,[8]

$$\hat{\tau}_{\text{match}} = \frac{1}{n} \sum_{i}^{n} \Big(\hat{y}_i(1) - \hat{y}_i(0) \Big). \quad (8)$$

Open problems: Ho et al. (2007) describe matching as a method to reduce model dependence because, unlike regression, it does not rely on a parameteric form. Yet, estimated causal effects may still be sensitive to other matching method decisions such as the number of bins in coarsened exact matching, the number of controls to match with each treatment in the matching algorithm, or the choice of caliper. Are causal estimates made using textual covariates particularly sensitive or robust to such choices?

5.3 Regression adjustment

Regression adjustment fits a supervised model from observed data about the expected conditional outcomes

$$q(t, z) \equiv \mathbb{E}(Y \mid T = t, Z = z) \quad (9)$$

Then the learned conditional outcome, $\hat{q}$, is used to predict counterfactual outcomes for each observation under treatment and control regimes,

$$\hat{\tau}_{\text{reg}} = \frac{1}{n} \sum_{i}^{n} (\hat{q}(1, z_i) - \hat{q}(0, z_i)) \quad (10)$$

5.4 Doubly-robust methods

Unlike methods that model only treatment (IPTW) or only outcome (regression adjustment), doubly robust methods model both treatment and outcome, and have the desirable property that if either the treatment or outcome models are unbiased then the effect estimate will be unbiased as well. These methods often perform very well in practice (Dorie et al., 2019). *Adjusted inverse probability of treatment weighting (A-IPTW)* combines estimated propensity scores (Eqn. 4) and conditional outcomes (Eqn. 9), while the more general *targeted maximum likelihood estimator (TMLE)* updates the conditional outcome estimate with a regression on the propensity weights (Eqn. 5) and $\hat{q}$ (Van der Laan and Rose, 2011).

5.5 Causal-driven representation learning

Several research efforts design representations of text specifically for causal inference goals. These

[7] Lunceford and Davidian (2004) note there are two versions of IPTW, where both the weighted sum and the raw count have been used for the n_0 and n_1 denominators.

[8] For alternative matching estimators see Abadie et al. (2004). This estimator is techinally the *sample* average treatment effect (SATE), not the population-level ATE, since we have pruned treatment and control pairs that do not have matches (Morgan and Winship, 2015).

approaches still initialize their models with representations of text described in Section 4, but then the representations are updated with machine learning architectures that incorporate the observed treatment assignment and other causal information. Johansson et al. (2016) design a network with a multi-task objective that aims for low prediction error for the conditional outcome estimates, q, and minimizes the discrepancy distance between $q(1, z_i)$ and $q(0, z_i)$ in order achieve balance in the confounders. Roberts et al. (2020) combine structural topic models (STM; Roberts et al. (2014)), propensity scores, and matching. They use the observed treatment assignment as the content covariate in the STM, append an estimated propensity score to the topic-proportion vector for each document, and then perform coarsened exact matching on that vector. Veitch et al. (2019) fine-tune a pre-trained BERT network with a multi-task loss objective that estimates (a) the original masked language-modeling objective of BERT, (b) propensity scores, and (c) conditional outcomes for both treatment and control. They use the predicted conditional outcomes and propensity scores in regression adjustment and the TMLE formulas.

Open problems: These methods have yet to be compared to one another on the same benchmark evaluation datasets. Also, when are the causal effects sensitive to hyperparameter and network architecture choices and what should researchers do in these settings?

6 Human evaluation of intermediate steps

Text data has the advantage of being *interpretable*—matched pairs and some low-dimensional representations of text can be read by humans to evaluate their quality. When possible, we suggest practitioners use (1) interpretable balance metrics and/or (2) human judgements of treatment propensity to evaluate intermediate steps of the causal estimation pipeline.

6.1 Interpretable balance metrics

For matching and propensity score methods, the confounder balance should be assessed, since ideally $P(Z \mid T = 1) = P(Z \mid T = 0)$ in a matched sample (Stuart, 2010). A standard numerical balance diagnostic is the *standardized difference in means (SDM)*,

$$SDM(j) = \frac{\frac{1}{n_1}\sum_{i:t_i=1} z_{ij} - \frac{1}{n_0}\sum_{i:t_i=0} z_{ij}}{\sigma_j^{t=1}}$$

where z_{ij} is a single confounder j for a single unit i and $\sigma_j^{t=1}$ is the standard deviation of z_{ij} for all i such that $t_i = 1$. SDM can also be used to evaluate the propensity score, in which case there would only be a single j (Rubin, 2001).

For causal text applications, Roberts et al. (2020) and Sridhar and Getoor (2019) estimate the difference in means for each topic in a topic-model representation of confounders and Sridhar et al. (2018) estimate the difference in means across structured covariates but not the text itself. As an alternative to SDM, Roberts et al. (2020) use string kernels to perform similarity checks. Others use domain-specific, known structured confounders to evaluate the balance between treatment and control groups. For instance, De Choudhury and Kiciman (2017) sample treatment-control pairs across all propensity score strata and label the sampled text based on known confounders (in their case, from a previously-validated codebook of suicidal ideation risk markers).

Open problems: For embeddings and causally-driven representations, each dimension in the confounder vector z is not necessarily meaningful. How can balance metrics be used in this setting?

6.2 Judgements of treatment propensity

When possible, one can also improve validation by evaluating matched items (posts, sentences, documents etc.) to humans for evaluation. Humans can either (a) use a scale (e.g., a 1-5 Likert scale) to rate items individually on their propensity for treatment, or (b) assess similarity of paired items after matching. A simple first step is for analysts to do "in-house" evaluation on a small sample (e.g., Roberts et al. (2020)), but a larger-sample experiments on crowd-working platforms can also increase the validity of these methods (e.g., Mozer et al. (2020)).

Open problems: How can these human judgement experiments be improved and standardized? Future work could draw from a rich history in NLP of evaluating representations of topic models and embeddings (Wallach et al., 2009; Bojanowski et al., 2017; Schnabel et al., 2015) and evaluating *semantic similarity* (Cer et al., 2017; Bojanowski et al., 2017; Reimers and Gurevych, 2019).

7 Evaluation of causal methods

Because the true causal effects in real-world causal inference are typically unknown, causal *evaluation* is a difficult and open research question. As

algorithmic complexity grows, the expected performance of causal methods can be difficult to estimate theoretically (Jensen, 2019). Other causal evaluations involve *synthetic data*. However, as Gentzel et al. (2019) discuss, synthetic data has no "unknown unknowns" and many researcher degrees of freedom, which limits their effectiveness. Thus, we encourage researchers to evaluate with *constructed observational studies* or *semi-synthetic datasets*, although measuring latent confounders from text increases the difficulty of creating realistic datasets that can be used for empirical evaluation of causal methods.

7.1 Constructed observational studies

Constructed observational studies collect data from both randomized and non-randomized experiments with similar participants and settings. Evaluations of this kind include job training programs in economics (LaLonde, 1986; Glynn and Kashin, 2013), advertisement marketing campaigns (Gordon et al., 2019), and education (Shadish et al., 2008). For instance, Shadish et al. (2008) randomly assign participants to a randomized treatment (math or vocabulary training) and non-randomized treatment (participants choose their own training). They compare causal effect estimates from the randomized study with observational estimates that condition on confounders from participant surveys (e.g., sex, age, marital status, like of mathematics, extroversion, etc.).

Open problems: To extend *constructed observational studies* to text data, one could build upon Shadish et al. (2008) and additionally (a) ask participants to write free-form essays of their past educational and childhood experiences and/or (b) obtain participants' public social media posts. Then causal estimates that condition on these textual representation of confounders could be compared to both those with surveys and the randomized settings. Alternatively, one could find observational studies with both real covariates and text and (1) randomize treatment conditional on the propensity score model (constructed from the covariates but not the text) and (2) estimate causal effect given only text (not the covariates). Then any estimated non-zero treatment effect is only bias.

7.2 Semi-synthetic datasets

Semi-synthetic datasets use real covariates and synthetically generate treatment and outcome, as in the 2016 Atlantic Causal Inference Competition (Dorie et al., 2019). Several applications in this review use real metadata or latent aspects of text to simulate treatment and outcome: Johansson et al. (2016) simulate treatment and outcome from two centroids in topic model space from newswire text; Veitch et al. (2019) use indicators of an article's "buzzy" keywords; Roberts et al. (2020) use "quantitative methodology" categories of articles that were hand-coded by other researchers.

Open problems: Semi-synthetic datasets that use real covariates of text seem to be a better evaluation strategy than purely synthetic datasets. However, with semi-synthetic datasets, researchers could be inadvertently biased to choose metadata that they know their method will recover. A promising future direction is a competition-style evaluation like Dorie et al. (2019) in which one group of researchers generates a causal dataset with text as a confounder and other groups of researchers evaluate their causal methods without access to the data-generating process.

8 Discussion and Conclusion

Computational social science is an exciting, rapidly expanding discipline. With greater availability of text data, alongside improved natural language processing models, there is enormous opportunity to conduct new and more accurate causal observational studies by controlling for latent confounders in text. While text data ought to be as useful for measurement and inference as "traditional" low-dimensional social-scientific variables, combining NLP with causal inference methods requires tackling major open research questions. Unlike predictive applications, causal applications have no ground truth and so it is difficult distinguish modeling errors and forking paths from the true causal effects. In particular, we caution against using all available text in causal adjustment methods *without* any human validation or supervision, since one cannot diagnose any potential errors. Solving these open problems, along with the others presented in this paper, would be a major advance for NLP as a social science methodology.

Acknowledgments

The authors thank Sam Witty, Jacob Eisenstein, Brandon Stewart, Zach Wood-Doughty, Andrew Halterman, Laura Balzer, and members of the University of Massachusetts Amherst NLP reading group for helpful feedback, as well as the anonymous referees for detailed peer reviews.

References

Alberto Abadie, David Drukker, Jane Leber Herr, and Guido W Imbens. 2004. Implementing matching estimators for average treatment effects in stata. *The Stata Journal*, 4(3):290–311.

Maria Antoniak and David Mimno. 2018. Evaluating the stability of embedding-based word similarities. *Transactions of the Association for Computational Linguistics*, 6:107–119.

Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2017. A simple but tough-to-beat baseline for sentence embeddings. In *ICLR*.

Susan Athey, Guido Imbens, Thai Pham, and Stefan Wager. 2017. Estimating average treatment effects: Supplementary analyses and remaining challenges. *American Economic Review*, 107(5):278–81.

Isabelle Augenstein, Kris Cao, He He, Felix Hill, Spandana Gella, Jamie Kiros, Hongyuan Mei, and Dipendra Misra. 2018. Proceedings of the Third Workshop on Representation Learning for NLP. In *Proceedings of The Third Workshop on Representation Learning for NLP*.

Isabelle Augenstein, Spandana Gella, Sebastian Ruder, Katharina Kann, Burcu Can, Johannes Welbl, Alexis Conneau, Xiang Ren, and Marek Rei. 2019. Proceedings of the 4th Workshop on Representation Learning for NLP. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*.

Ananth Balashankar, Sunandan Chakraborty, Samuel Fraiberger, and Lakshminarayanan Subramanian. 2019. Identifying predictive causal factors from news streams. In *Empirical Methods in Natural Langugage Processing*.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022.

Su Lin Blodgett and Brendan O'Connor. 2017. Racial disparity in natural language processing: A case study of social media african-american english. In *Fairness, Accountability, and Transparency in Machine Learning (FAT/ML) Workshop, KDD*.

Phil Blunsom, Antoine Bordes, Kyunghyun Cho, Shay Cohen, Chris Dyer, Edward Grefenstette, Karl Moritz Hermann, Laura Rimell, Jason Weston, and Scott Yih. 2017. Proceedings of the 2nd Workshop on Representation Learning for NLP. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*.

Phil Blunsom, Kyunghyun Cho, Shay Cohen, Edward Grefenstette, Karl Moritz Hermann, Laura Rimell, Jason Weston, and Scott Wen-tau Yih. 2016. Proceedings of the 1st Workshop on Representation Learning for NLP. In *Proceedings of the 1st Workshop on Representation Learning for NLP*.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Jordan Boyd-Graber, David Mimno, and David Newman. 2014. Care and feeding of topic models: Problems, diagnostics, and improvements. *Handbook of Mixed Membership Models and Their Applications*, 225255.

John P Buonaccorsi. 2010. *Measurement Error: Models, Methods, and Applications*. CRC Press.

Marco Caliendo and Sabine Kopeinig. 2008. Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys*, 22(1):31–72.

Dallas Card and Noah A Smith. 2018. The importance of calibration for estimating proportions from annotations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics*.

Raymond J Carroll, David Ruppert, Leonard A Stefanski, and Ciprian M Crainiceanu. 2006. *Measurement Error in Nonlinear Models: a Modern Perspective*. CRC Press.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, Vancouver, Canada. Association for Computational Linguistics.

Stevie Chancellor, Eric PS Baumer, and Munmun De Choudhury. 2019. Who is the human in human-centered machine learning: The case of predicting mental health from social media. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):147.

Alexander D'Amour, Peng Ding, Avi Feller, Lihua Lei, and Jasjeet Sekhon. 2017. Overlap in observational studies with high-dimensional covariates. *arXiv preprint arXiv:1711.02582*.

Munmun De Choudhury and Emre Kiciman. 2017. The language of social support in social media and its effect on suicidal ideation risk. In *International AAAI Conference on Web and Social Media (ICWSM)*.

Munmun De Choudhury, Emre Kiciman, Mark Dredze, Glen Coppersmith, and Mrinal Kumar. 2016. Discovering shifts to suicidal ideation from mental health content in social media. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pages 2098–2110. ACM.

Matthew J Denny and Arthur Spirling. 2018. Text preprocessing for unsupervised learning: Why it matters, when it misleads, and what to do about it. *Political Analysis*, 26(2):168–189.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *North American Association of Computational Linguistics*.

Vincent Dorie, Jennifer Hill, Uri Shalit, Marc Scott, and Daniel Cervone. 2019. Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. *Statistical Science*, 34(1):43–68.

Long Duong, Trevor Cohn, Steven Bird, and Paul Cook. 2015. Low resource dependency parsing: Cross-lingual parameter sharing in a neural network parser. In *Association for Computational Linguistics*.

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

Felix Elwert and Christopher Winship. 2014. Endogenous selection bias: The problem of conditioning on a collider variable. *Annual Review of Sociology*, 40:31–53.

Seyed Amin Mirlohi Falavarjani, Hawre Hosseini, Zeinab Noorian, and Ebrahim Bagheri. 2017. Estimating the effect of exercising on users' online behavior. In *Eleventh International AAAI Conference on Web and Social Media*.

Christian Fong and Justin Grimmer. 2016. Discovery of treatments from text corpora. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1600–1609.

Wayne A Fuller. 1987. *Measurement Error Models*. John Wiley & Sons.

Andrew Gelman and Eric Loken. 2013. The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time. *Department of Statistics, Columbia University*.

Amanda Gentzel, Dan Garant, and David Jensen. 2019. The case for evaluating causal models using interventional measures and empirical data. In *Advances in Neural Information Processing Systems*.

Adam Glynn and Konstantin Kashin. 2013. Front-door versus back-door adjustment with unmeasured confounding: Bias formulas for front-door and hybrid adjustments. In *71st Annual Conference of the Midwest Political Science Association*, volume 3.

Brett R Gordon, Florian Zettelmeyer, Neha Bhargava, and Dan Chapsky. 2019. A comparison of approaches to advertising measurement: Evidence from big field experiments at facebook. *Marketing Science*, 38(2):193–225.

MA Hernán and JM Robins. 2020. *Causal Inference: What If*. Boca Raton: Chapman Hall/CRC.

Daniel E Ho, Kosuke Imai, Gary King, and Elizabeth A Stuart. 2007. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis*, 15(3):199–236.

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

Stefano M Iacus, Gary King, and Giuseppe Porro. 2012. Causal inference without balance checking: Coarsened exact matching. *Political Analysis*.

Guido W Imbens. 2000. The role of the propensity score in estimating dose-response functions. *Biometrika*, 87(3):706–710.

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

Mohit Iyyer, Varun Manjunatha, Jordan Boyd-Graber, and Hal Daumé III. 2015. Deep unordered composition rivals syntactic methods for text classification. In *Association for Computational Linguistics*.

David Jensen. 2019. Comment: Strengthening empirical evaluation of causal inference methods. *Statistical Science*, 34(1):77–81.

Fredrik Johansson, Uri Shalit, and David Sontag. 2016. Learning representations for counterfactual inference. In *ICML*.

Katherine Keith and Brendan O'Connor. 2018. Uncertainty-aware generative models for inferring document class prevalence. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Emre Kiciman, Scott Counts, and Melissa Gasser. 2018. Using longitudinal social media analysis to understand the effects of early college alcohol use. In *Twelfth International AAAI Conference on Web and Social Media*.

Manabu Kuroki and Judea Pearl. 2014. Measurement bias and effect restoration in causal inference. *Biometrika*, 101(2):423–437.

Mark J Van der Laan and Sherri Rose. 2011. *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer Science & Business Media.

Robert J LaLonde. 1986. Evaluating the econometric evaluations of training programs with experimental data. *The American Economic Review*, pages 604–620.

Omer Levy, Yoav Goldberg, and Ido Dagan. 2015. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225.

Sheng Li, Nikos Vlassis, Jaya Kawale, and Yun Fu. 2016. Matching via dimensionality reduction for estimation of treatment effects in digital marketing campaigns. In *IJCAI*.

Christos Louizos, Uri Shalit, Joris M Mooij, David Sontag, Richard Zemel, and Max Welling. 2017. Causal effect inference with deep latent-variable models. In *Advances in Neural Information Processing Systems*.

Jared K Lunceford and Marie Davidian. 2004. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in Medicine*, 23(19):2937–2960.

Subramani Mani and Gregory F Cooper. 2000. Causal discovery from medical textual data. In *Proceedings of the AMIA Symposium*, page 542. American Medical Informatics Association.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*.

Jacob M Montgomery, Brendan Nyhan, and Michelle Torres. 2018. How conditioning on posttreatment variables can ruin your experiment and what to do about it. *American Journal of Political Science*, 62(3):760–775.

Stephen L Morgan and Christopher Winship. 2015. *Counterfactuals and Causal Inference*. Cambridge University Press.

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*.

Khanh Nguyen and Brendan O'Connor. 2015. Posterior calibration and exploratory analysis for natural language processing models. In *Empirical Methods in Natural Langugage Processing*.

Hüseyin Oktay, Akanksha Atrey, and David Jensen. 2019. Identifying when effect restoration will improve estimates of causal effect. In *Proceedings of the 2019 SIAM International Conference on Data Mining*, pages 190–198. SIAM.

Alexandra Olteanu, Onur Varol, and Emre Kiciman. 2017. Distilling the outcomes of personal experiences: A propensity-scored analysis of social media. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pages 370–386. ACM.

Judea Pearl. 2000. *Causality: Models, Reasoning and Inference*. Springer.

Judea Pearl. 2009a. Causal inference in statistics: An overview. *Statistics Surveys*, 3:96–146.

Judea Pearl. 2009b. *Causality: Models, Reasoning and Inference*, Second edition. Cambridge University Press.

Judea Pearl. 2014. Interpretation and identification of causal mediation. *Psychological Methods*, 19(4):459.

Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. 2016. *Causal Inference in Statistics: A Primer*. John Wiley & Sons.

James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Langugage Processing*.

Thai T Pham and Yuanyuan Shen. 2017. A deep causal inference approach to measuring the effects of forming group loans in online non-profit microfinance platform. *arXiv preprint arXiv:1706.02795*.

Jason Phang, Thibault Févry, and Samuel R Bowman. 2018. Sentence encoders on stilts: Supplementary training on intermediate labeled-data tasks. *arXiv preprint arXiv:1811.01088*.

Fermin Moscoso del Prado Martin and Christian Brendel. 2016. Case and cause in icelandic: Reconstructing causal networks of cascaded language changes. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2421–2430.

Jeremy A Rassen, Robert J Glynn, M Alan Brookhart, and Sebastian Schneeweiss. 2011. Covariate selection in high-dimensional propensity score analyses of treatment effects in small samples. *American Journal of Epidemiology*, 173(12):1404–1413.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Empirical Methods in Natural Langugage Processing*.

Thomas S Richardson and James M Robins. 2013. Single world intervention graphs (SWIGs): A unification of the counterfactual and graphical approaches to causality. *Center for the Statistics and the Social Sciences, University of Washington Series. Working Paper*, (128).

Margaret E Roberts, Brandon M Stewart, and Richard A Nielsen. 2020. Adjusting for confounding with text matching. *American Journal of Political Science (forthcoming)*.

Margaret E Roberts, Brandon M Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson, and David G Rand. 2014. Structural topic models for

open-ended survey responses. *American Journal of Political Science*, 58(4):1064–1082.

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.

Paul R Rosenbaum and Donald B Rubin. 1984. Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, 79(387):516–524.

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

Donald B Rubin. 2001. Using propensity scores to help design observational studies: application to the tobacco litigation. *Health Services and Outcomes Research Methodology*, 2(3-4):169–188.

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

Koustuv Saha, Benjamin Sugar, John Torous, Bruno Abrahao, Emre Kıcıman, and Munmun De Choudhury. 2019. A social media study on the effects of psychiatric medication use. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, pages 440–451.

Matthew Salganik. 2017. *Bit By Bit: Social Research in the Digital Age*. Princeton University Press.

Tobias Schnabel, Igor Labutov, David Mimno, and Thorsten Joachims. 2015. Evaluation methods for unsupervised word embeddings. In *Empirical Methods in Natural Langugage Processing*.

Alexandra Schofield, Måns Magnusson, and David Mimno. 2017. Pulling out the stops: Rethinking stopword removal for topic models. In *Association for Computational Linguistics*.

William R Shadish, Margaret H Clark, and Peter M Steiner. 2008. Can nonrandomized experiments yield accurate answers? A randomized experiment comparing random and nonrandom assignments. *Journal of the American Statistical Association*, 103(484):1334–1344.

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

Dhanya Sridhar, Aaron Springer, Victoria Hollis, Steve Whittaker, and Lise Getoor. 2018. Estimating causal effects of exercise from mood logging data. In *IJCAI/ICML Workshop on CausalML*.

Elizabeth A Stuart. 2010. Matching methods for causal inference: A review and a look forward. *Statistical Science*, 25(1):1.

Narges Tabari, Piyusha Biswas, Bhanu Praneeth, Armin Seyeditabari, Mirsad Hadzikadic, and Wlodek Zadrozny. 2018. Causality analysis of twitter sentiments and stock market returns. In *Proceedings of the First Workshop on Economics and Natural Language Processing*. Association for Computational Linguistics.

Chenhao Tan, Lillian Lee, and Bo Pang. 2014. The effect of wording on message propagation: Topic- and author-controlled natural experiments on twitter. In *Association for Computational Linguistics*.

Tyler VanderWeele. 2015. *Explanation in Causal Inference: Methods for Mediation and Interaction*. Oxford University Press.

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

Hanna M Wallach, Iain Murray, Ruslan Salakhutdinov, and David Mimno. 2009. Evaluation methods for topic models. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1105–1112. ACM.

Zach Wood-Doughty, Ilya Shpitser, and Mark Dredze. 2018. Challenges of using text classifiers for causal inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4586–4598.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*.