Causal Inference in Natural Language Processing: Estimation, Prediction, Interpretation and Beyond

Amir Feder[*,1], Katherine A. Keith[2], Emaad Manzoor[3], Reid Pryzant[4], Dhanya Sridhar[5], Zach Wood-Doughty[6], Jacob Eisenstein[7], Justin Grimmer[4], Roi Reichart[1], Margaret E. Roberts[8], Brandon M. Stewart[9], Victor Veitch[7,10], and Diyi Yang[11]

[1]Technion - Israel Institute of Technology
[2]University of Massachusetts Amherst
[3]University of Wisconsin - Madison
[4]Stanford University
[5]Columbia University
[6]Johns Hopkins University
[7]Google Research
[8]University of California San Diego
[9]Princeton University
[10]University of Chicago
[11]Georgia Tech

arXiv:2109.00725v1 [cs.CL] 2 Sep 2021

Abstract

A fundamental goal of scientific research is to learn about causal relationships. However, despite its critical role in the life and social sciences, causality has not had the same importance in Natural Language Processing (NLP), which has traditionally placed more emphasis on predictive tasks. This distinction is beginning to fade, with an emerging area of interdisciplinary research at the convergence of causal inference and language processing. Still, research on causality in NLP remains scattered across domains without unified definitions, benchmark datasets and clear articulations of the remaining challenges. In this survey, we consolidate research across academic areas and situate it in the broader NLP landscape. We introduce the statistical challenge of estimating causal effects, encompassing settings where text is used as an outcome, treatment, or as a means to address confounding. In addition, we explore potential uses of causal inference to improve the performance, robustness, fairness, and interpretability of NLP models. We thus provide a unified overview of causal inference for the computational linguistics community. [1]

[*]All authors equally contributed to this paper. Author names are organized alphabetically in two clusters: First students and post-docs and then faculty members. The email address of the corresponding (first) author is: feder@campus.technion.ac.il.

[1]An online repository containing existing research on causal inference and language processing is available here: https://github.com/causaltext/causal-text-papers

1 Introduction

Many scientific fields are increasingly interested in incorporating text as data (e.g., Roberts et al., 2014; Pryzant et al., 2017; Zhang et al., 2020a). A key property of these fields that may be unfamiliar to natural language processing (NLP) researchers is the emphasis on causal inference, often to evaluate policy interventions. For example, before recommending a new drug therapy, clinicians want to know the causal effect of the drug on disease progression. Causal inference involves a question about a counterfactual world created by taking an intervention: what would a patient's disease progression have been if we had given them the drug? As we explain below, in observational data, the causal effect is not equivalent to the correlation between patients taking the drug and their observed disease progression. There is now a deep literature on techniques for making valid inferences using traditional (non-text) datasets (e.g., Morgan and Winship, 2015), but the application of these techniques to natural language data raises new and fundamental challenges.

Conversely, in classical NLP applications, the goal is simply to make accurate predictions: any statistical correlation is typically considered to be admissible, regardless of the underlying causal relationship. However, as NLP systems are increasingly deployed in challenging and high-stakes scenarios, we cannot rely on the usual assumption that training and test data are identically distributed,

and we may not be satisfied with uninterpretable black-box predictors. For both of these problems, causality offers a promising path forward: domain knowledge of the causal structure of the data generating process can suggest inductive biases that lead to more robust predictors, and a causal view of the predictor itself can offer new insights on its inner workings.

This core claim of this survey paper is that deepening the connection between causality and NLP has the potential to advance the goals of both social science and NLP researchers. We divide the intersection of causality and NLP into two distinct areas: estimating causal effects from text, and using causal formalisms to make NLP methods more reliable. We will illustrate this distinction with two examples.

Example 1. *An online forum has allowed its users to indicate their preferred gender in their profiles with an icon. They notice that users who label themselves with the "woman" icon tend to receive fewer "likes" on their posts. To better evaluate their policy of allowing gender information in profiles, they ask: does being perceived as a woman cause a decrease in popularity for a post?*

Example 1 is about the causal effect of being perceived as a woman (treatment) on the likes a post receives (outcome). The counterfactual question is: if we could manipulate the gender icon on a given post, how many likes would the post have received?

The observed correlation between being perceived as a woman and receiving fewer likes typically does not coincide with the causal effect. This is because the correlation arises for two possible reasons: the true causal effect, and a spurious correlation induced by confounders, variables that are correlated with both the treatment and outcome. In the example, the topic of a post is likely a confounder: posts labeled with the woman icon may be about a certain topic more often, and that topic may not receive as many likes. As we will see in Section 2, due to confounding, estimating a causal effect is not possible without assumptions.

Example 1 highlights the setting where the text encodes the relevant confounders of a causal effect. The text as a confounder setting is one of many causal inferences we can make with text data. The text data can also encode outcomes or treatments of interest. For example, we may wonder about how perceived gender affects the sentiment of the reply that a post receives (text as outcome), or about how a writing style affects the likes a post receives (text as treatment)?

NLP helps causality. Causal inference with text data involves several challenges that distinct from typical causal inference settings: text is high-dimensional, needs sophisticated modeling to measure semantically meaningful factors like topic, and demands careful thought to formalize the intervention that a causal question corresponds to. The developments in NLP around modeling language, from topic models to contextual embeddings, offer promising ways to extract the information we need from text to estimate causal effects. However, we need new assumptions to ensure that the use of NLP methods leads to valid causal inferences. We will discuss existing research on estimating causal effects from text and emphasize these challenges and opportunities in Section 3.

Example 2. *A medical research center wants to build a classifier to detect clinical diagnoses from the textual narratives of patient medical records. The records are aggregated across multiple hospital sites, which vary both in the frequency of the target clinical condition and the writing style of the narratives. When the classifier is applied to records from sites that were not in the training set, its accuracy decreases. Post-hoc analysis indicates that it puts significant weight on seemingly irrelevant features, such as formatting markers.*

Like Example 1, Example 2 also involves a counterfactual question: does the classifier's prediction change if we intervene to change the hospital site, while holding the true clinical status fixed? We want the classifier to rely on phrases that express clinical facts, and not writing style. However, in the training data, the clinical condition and the writing style are spuriously correlated, due to the site acting as a confounding variable: for example, a site might be more likely to encounter the target clinical condition due to its location or speciality, and that site might also employ distinctive textual features, such as boilerplate text at the beginning of each narrative. In the training set, these features will be predictive of the label, but they are unlikely to be useful in deployment scenarios at new sites. In this example, the hospital site acts like a confounder: it creates a spurious correlation between some features of the text and the prediction target.

Example 2 shows how the lack of robustness

can make NLP methods less trustworthy. A related problem is that NLP systems are often black boxes, making it hard to understand how human-interpretable features of the text lead to the observed predictions. In this setting, we want to know if some part of the text (e.g., some sequence of tokens) causes the output of an NLP method (e.g., classification prediction).

Causality can help NLP. To address the robustness and interpretability challenges posed by NLP methods, we need new criteria to learn models that go beyond exploiting correlations. For example, we want predictors that are invariant to certain changes that we make to text, such as changing the format while holding fixed the ground truth label. There is considerable promise in using causality to develop new criteria in service of building robust and interpretable NLP methods. In contrast to the well-studied area of causal inference with text, this area of causality and NLP research is less well-understood, though well-motivated by recent empirical successes. In Section 4, we cover the existing research and review the challenges and opportunities around using causality to improve NLP.

This position paper follows a small body of surveys that review the role of text data within causal inference (Egami et al., 2018; Keith et al., 2020). We take a broader view in this paper, separating the intersection of causality and NLP into two distinct lines of research on estimating causal effects and causally motivated NLP methods. After reading this paper, we envision that the reader will have a broad understanding of:

- different types of causal queries and the challenges they present;

- the statistical and causal challenges that are unique to working with text data and NLP methods;

- open problems in estimating effects from text and applying causality to improve NLP methods.

2 Background

Both focal problems of this survey (causal effect estimation and causally motivated NLP) involve causal inference. The key ingredient to causal inference is defining counterfactuals based on an intervention of interest. We will illustrate this idea with the motivating examples in Section 1.

Example 1. Example 1 involves online forum posts W and the number of likes Y that they receive. For ease of explanation, we use a binary variable T to indicate if the author of a post uses the "woman icon" (1) or not (0). We would ideally like to intervene upon the variable T, referred to as the treatment. The counterfactual outcome $Y(1)$ represents the number of likes a post would have received had we intervened to use the woman icon. The counterfactual outcome $Y(0)$ is defined analogously.

The *fundamental problem of causal inference* is that we can never observe both counterfactual outcomes for a post; it is impossible to make both interventions to the treatment and observe the resulting outcomes. This problem is what makes causal inference harder than statistical inference and impossible without assumptions.

Example 2. Example 2 involves a trained classifier $\hat{Y}(W)$ that takes a clinical narrative W as input and outputs a diagnosis prediction. The text W is written based on the physician's diagnosis Y, and is also influenced by the writing style employed at the hospital Z. We want to intervene upon the hospital Z while holding fixed the label Y. The counterfactual narrative $W(z)$ is the text we would have had we set the hospital to the value z while holding the diagnosis fixed. The counterfactual prediction $\hat{Y}(W(z))$ is the output the trained classifier would have produced had we given the counterfactual review $W(z)$ as input.

2.1 Counterfactual queries

Based on the defined counterfactuals, an analyst must then specify a quantity of interest that involves the distribution of counterfactuals. In Example 1, the causal effect of interest is the average treatment effect (ATE),

$$\text{ATE} = \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)]. \tag{1}$$

It is the average difference in counterfactual likes received by posts if we could intervene to change the gender with which the post author is perceived. Example 2 involves more counterfactual questions; they will be explained in Section 4.

2.2 Causal graphical models

Causal inference requires assumptions to identify counterfactual queries, i.e., express them as a function of the observed outcomes. Thus, we need to encode our model of the world to verify whether the model satisfies the assumptions
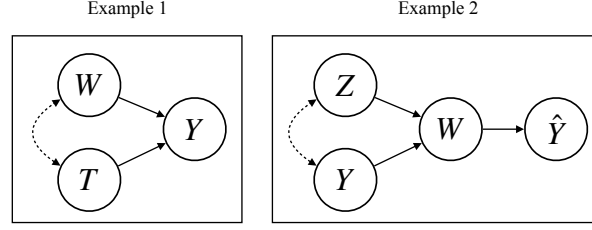
Figure 1: Causal models for the motivating examples. (**Left**) In Example 1, the perceived gender of a post's author (T) is correlated with attributes of the post (W), and both variables affect the number of likes a post receives (Y). (**Right**) In Example 2, the label (Y, i.e., diagnosis) and hospital site (Z) are correlated, and both affect the clinical narrative (W). Predictions $\hat{Y}$ from the trained classifier depend on W.

needed to make causal inferences or not. We can articulate our models of the world using causal directed acyclic graphs (causal DAGs). In a causal DAG, edges from a variable X to a variable Y means that changing the value of X *may* change the distribution of Y. We use bi-directed dotted arrows between variables to indicate that they are correlated.

Figure 1 illustrates the assumed causal DAGs for Example 1 and Example 2. In Example 1, the gender-signaling icon chosen by a post's author (T) is correlated with attributes of the post (W), and both variables affect the number of likes a post receives (Y). In Example 2, the diagnosis (Y) and hospital (Z) are correlated, and both affect the text of the clinical narrative (W). The trained classifier then makes predictions $\hat{Y}$ based on the text W.

Causal DAGs entail all statistical dependencies between variables. For example, in the right DAG in Figure 1, the prediction $\hat{Y}$ is not independent of the hospital Z for each narrative W. We can read off such independence statements using the d-separation algorithm (Pearl, 1994). As we will see below, these statements allow us to check whether a causal DAG satisfies some of the assumptions needed for causal inference.

2.3 Assumptions for causal inference

We will focus on Example 1 to explain the assumptions needed for causal inference. Specifically, we will review the assumptions that make it possible to identify the ATE in Equation (1). Although we focus on the ATE, related assumptions are needed in some form for all causal inferences.

Ignorability is the most important and difficult to justify assumption. It requires that treatment assignment be independent of the realized coun-

terfactual outcomes,

$$(Y(1), Y(0)) \perp\!\!\!\perp T. \tag{2}$$

Consider Example 1: suppose that users who use the woman icon write about topics that receive fewer likes. Then, observing $T = 1$ for a post—it is perceived as being written by a woman—gives us information about the counterfactual likes $Y(1)$ and $Y(0)$, violating ignorability.

Randomizing treatment assignment is one way to satisfy ignorability. Randomization ensures that in expectation there are no systematic pre-treatment differences between treated and untreated samples. For example, the online forum could run an A/B test, a randomized trial where some readers are randomly assigned posts labeled with a woman icon and some are randomly assigned posts with other icons.

Randomized assignment may not always be feasible. In this case, we may need to rely on *conditional* ignorability,

$$((Y(1), Y(0)) \perp\!\!\!\perp T) \,|\, X.$$

where X is a set of variables such that treatment assignment and the potential outcomes is unconfounded within levels of X. We can use causal DAGs to read off all necessary confounders based on the backdoor criteria (Pearl, 2009), an algorithm derived from d-separation. As an example, Figure 1 tells us that for Example 1, considering the post W itself as a confounder satisfies conditional ignorability. Conditional ignorability may seem like a free lunch but it requires that there are *no unobserved confounders*; this is a strong assumption that analyst must carefully assess.

Positivity is the assumption that the probability of receiving treatment is bounded between 0 and

1. Under conditional ignorability this must hold for all x such that $\Pr(X = x) \neq 0$,

$$0 < \Pr(T = 1 | X = x) < 1. \tag{3}$$

Intuitively this requires that in order to learn about causal effects, it needs to be possible to observe units under both types of treatment statuses. It also implies that the treatment status cannot be perfectly predicted given the variables X. Randomized treatment assignment also ensures positivity.

Consistency is the last assumption and requires that the observed outcome at a given treatment status for a given unit, is the same as we would observe if that unit was assigned to the treatment,

$$\forall i \text{ s.t. } T_i = t, Y_i(t) = Y_i. \tag{4}$$

Consistency relates counterfactual outcomes to observed ones and captures two main assertions. First, there is **no interference**: the outcome for a sample i is affected only by sample i's treatment status, and not the treatment status of other samples. The second is that there is only **one version of treatment**. In Example 1, this means that the perceived gender is what matters for the counterfactual likes received by a post; the manner by which we intervene to change the perceived gender – whether we manipulate an icon or change the username – does not change the counterfactual outcomes.

We refer to other background material to discuss how to identify and estimate causal effects with these assumptions in hand (Pearl, 2009; Rubin, 2005; Keith et al., 2020; Egami et al., 2018).

3   Estimating Causal Effects with Text

In this section, we outline the possibilities and challenges of using NLP to identify causal effects in each case – text as outcome, text as treatment, and text as confounder, and how assumptions and estimation are complicated by the addition of text data.

3.1   Text as Outcome

Text can serve naturally as the outcome of a randomized experiment where subjects produce an open-ended response (Roberts et al., 2014). The core challenge is to distill the high-dimensional text data into some low-dimensional quality of interest. Consider a researcher studying the effect of a particular educational intervention on the readability of student essays. Suppose that the researcher can run a randomized experiment, randomly assigning students to two groups. The treatment group receives the educational intervention and the control group does not. This is a case of text as outcome, where the researcher is interested in comparing an aspect of the writing – readability – between the treated and control groups (Egami et al., 2018). Because "readability" is a latent aspect of language, the measure of the outcome is unknown at the start of the experiment which complicates our usual assumptions for causal inference.

As we discussed in Section 2, randomizing the treatment assignment ensures positivity and ignorability. To satisfy consistency, however, we require more careful assessment even with randomization. First, we need to ensure that the students in the experiment did not affect each other during the study (i.e., no interference). Second, the researcher needs to create the measure of the outcome, e.g., readability, before conducting the study.

Suppose that after students are randomly assigned treated and control groups, the researcher uses manual coding, supervised classification, or unsupervised learning to develop a measure of readability based on students' essays. Developing the measure within the study has many advantages because a more relevant measure may be discovered and the measure can be validated on the data of interest. However, by developing the measure with reference to the data, the treatment status of one observation could affect the outcome of another through the measurement model. This approach would not only be a violation of consistency, but also could lead to multiple testing concerns if the researcher were exploring multiple different measures on the same data (Dror et al., 2017). Split samples or sequential experiments can be used to effectively address these issues by developing the measure on a sample of the data and estimating the effect in the held-out data only after the measure has been finalized (Egami et al., 2018).

3.2   Text as Confounder

Consider a dataset of scholarly articles and the number of citations they receive. Each article also contains the names of its authors. We might be

interested in the impact of the first author being perceived as female on the number of citations the article receives (Maliniak et al., 2013). Here, we cannot randomly assign author names to papers – such an experiment would be not only unethical but also impractical. Without random assignment, estimating the correlation between author gender and citations might be confounded by other variables. One confounder might be the topic of the article: Some subject areas have a higher proportion of authors that can be perceived as female. The subject area also affects the citation count of a paper. The text of the article provides evidence of the subject area – this is a text as confounding problem (Roberts et al., 2020).

The challenge of text as confounder is to condition on the text in a way that blocks confounding using NLP methods. One approach is by applying unsupervised dimensionality reduction methods that reduce high-dimensional text data to a low-dimensional set of variables. Such methods include latent variable models such as topic models, embedding methods, and auto-encoders. Roberts et al. (2020) and Sridhar and Getoor (2019) have applied topic models to extract confounding patterns from text data, and performed an adjustment for these inferred variables. Mozer et al. (2020) match directly based on distance metrics on the bag of words representation.

A second approach to adjusting for confounders from text with NLP methods is with supervised models. Recently, Veitch et al. (2020) adapted pre-trained language models and supervised topic models to predict the treatment and outcome. By learning low-dimensional variables that predict treatment and outcome well, they showed that confounding properties could be found within text data. Roberts et al. (2020) combines these strategies with the topic model approach in a text matching framework.

Because text as confounder is applied to settings where only observational data is available, using NLP to adjust for confounding requires particularly strong assumptions around ignorability – all aspects of confounding must be measured by the model. This is particularly challenging with high-dimensional data because as the number of variables we need to condition on grows it becomes increasingly difficult to satisfy positivity (D'Amour et al., 2020) and we risk accidentally inducing a subtle problem by conditioning on collider variables that open a backdoor path that would otherwise be closed (Pearl, 2009). Keith et al. (2020) provides a recent overview of this fast growing area of research as well as a deeper look at many of these threats to inference.

3.3 Text as Treatment

In text as treatment questions, we are interested in the causal relationship that language has on downstream decisions, behaviors, and other outcomes. Often, researchers are interested in how specific *aspects* of the text cause these outcomes. For example, does the writing style of a "get out the vote" ad cause readers to vote?

One approach to studying the effects of text involves treatment *discovery*: producing interpretable features of the text – e.g. topics and other latent dimensions of influence (Fong and Grimmer, 2016) or lexical features like n-grams (Pryzant et al., 2018) – that can be causally linked to outcomes. For example, Fong and Grimmer (2016) discovered features of candidate biographies that drove voter evaluations, Pryzant et al. (2017) discovered writing styles in marketing materials that are influential in increasing sales figures, and Zhang et al. (2020b) discovered conversational tendencies that lead to positive mental health counseling sessions.

Another approach is to estimate the causal effects of specific properties extracted from text (Pryzant et al., 2020; Wood-Doughty et al., 2018). For example, Gerber et al. (2008) studied the effect of appealing to civic duty on voter turnout. In this setting, factors are latent properties of the text for which we need a measurement model.

The text as treatment setting presents several unique challenges for causal inference. First, since people choose the text they read for reasons related to the outcome of interest, ignorability is typically violated. To overcome this, Fong and Grimmer (2016, 2021) randomly assign texts to readers. Pryzant et al. (2020) estimate effects of text properties when randomized assignment is not possible, but require the strong assumption that the text contains all confounders. If there are unobserved reasons that affect which texts people read and their outcome (e.g., their political affiliation), the assumption is violated.

The second challenge unique to text as treatment involves positivity. Suppose we want to estimate the causal effect of politeness on email re-

sponse times. Even if the text contains all confounders (e.g., topic, tone, writing style), it may be impossible to imagine a polite email that contains a certain writing style (e.g., profane).

The last challenge posed by text as treatment relates to the same consistency concerns as we discussed in the text as outcome case: to make valid inferences, we need to develop our measure of the treatment in different data than we use to estimate the causal effect.

3.4 Opportunities

There are several opportunities for NLP researchers to facilitate causal inference from text. We highlight key open challenges in the area of causal inference from text data.

Heterogeneous effects. Texts are read and interpreted differently by different people. This creates several problems for interpreting causal effects and communicating research findings. For example, when text is the treatment, the same text might have a different effect on different people in part because each person's interpretation of the text is different. Formulating and identifying heterogeneous causal effects with text will require new counterfactual questions, assumptions and methods.

Benchmarks. Benchmark datasets have propelled machine learning forward by creating shared metrics by which predictive models can be evaluated. While benchmarks are growing in the field of causal explanations (see §4.3), such NLP datasets are absent in the estimation literature (Feder et al., 2021). This is because the potential outcomes that are counter to fact are by definition unobservable — any benchmark dataset that assumes knowledge of counterfactuals is necessarily making a strong assumption about the data generating process that is unverifiable. While we might be able to compare aggregate results to experiments, we can never observe a true causal effect to benchmark against.

4 Robust and Explainable Predictions from Causality

Thus far we have focused on using NLP tools for estimating causal effects in the presence of text data. We now consider the converse relationship: using causal reasoning to help solve more traditional NLP tasks like understanding, manipulating, and generating natural language.

At first glance, NLP may appear to have little need for causal ideas. The remarkable progress that the field has achieved in the past few years has been derived from the use of increasingly high-capacity neural architectures to extract correlations from large-scale datasets (Peters et al., 2018; Devlin et al., 2019; Liu et al., 2019). These architectures make no distinction between causes, effects, and confounders, and they make no attempt to identify causal relationships: a feature may be a powerful predictor even if it has no direct causal relationship with the desired output.

Yet despite the march of state-of-the-art results, correlational predictive models can be untrustworthy (Jacovi et al., 2021): they may latch onto spurious correlations ("shortcuts"), leading to errors in out-of-distribution (OOD) settings (e.g., McCoy et al., 2019); they may exhibit unacceptable performance differences across groups of users (e.g., Zhao et al., 2017); and their behavior may be too inscrutable to incorporate into high-stakes decisions (Guidotti et al., 2018). Each of these shortcomings can potentially be addressed by the causal perspective: knowledge of the causal relationship between observations and labels can be used to formalize spurious correlations and to mitigate predictor reliance on them (Bühlmann, 2020; Veitch et al., 2021); causality also provides a language for specifying and reasoning about fairness conditions (Kilbertus et al., 2017); the task of explaining predictions may be naturally formulated in terms of counterfactuals (Feder et al., 2021).

For these reasons, a growing line of research has attempted to reorient machine learning around causal foundations (e.g., Schölkopf, 2019; Schölkopf et al., 2021). Thus far, applications have largely focused on other domains, such as biomedical data, where trustworthiness is particularly critical (e.g., Muandet et al., 2013), and computer vision, where there is a strong interest in cross-domain learning (e.g., Ghifary et al., 2015) and where it is relatively easy to construct datasets with artificial spurious correlations.[2]

In this section, we review work applying causal ideas to natural language processing, focusing on the unique problems and opportunities that arise. We begin in §4.1 with a discussion of the threats to generalization posed by spurious correlations,

---

[2]A prominent example is "colored MNIST", in which the color channel is utilized to create challenging distributional shifts (Kim et al., 2019).

distributional shift, and causally-motivated approaches to achieving greater robustness. We then briefly review attempts to develop a causal formalization of fairness and bias problems (§4.2), and conclude with a discussion of causal approaches to interpreting and explaining the behavior of complex machine learning models (§4.3).

4.1 Learning Robust Predictors

The NLP field has grown increasingly concerned with *spurious correlations* (Geirhos et al., 2020; Wang and Culotta, 2020; Gardner et al., 2021). Spurious correlations arise when two conditions are met. First, there must be some factor(s) Z that are informative (in the training data) about both the features X and label Y. Second, Y and Z must be dependent in the training data in a way that is not guaranteed to hold in general. The issue is that a predictor $f : X \rightarrow Y$ will learn to use parts of X that carry information about Z (because Z is informative about Y), but the learned relationship between these parts of X and Y need not hold when the model is deployed.

As an example, consider the task of predicting a medical condition from the text of patient records, using a training set from multiple hospitals. Suppose that there is a hospital in which (a) patients are more likely to be diagnosed with the condition and (b) doctors tend to use unique textual features, e.g. due to hospital practices or local dialect. A predictor trained on such data will use textual features that carry information about the hospital, even when they are useless at predicting the diagnosis *within* any individual hospital. This can lead to poor out-of-distribution performance, as the spurious correlation between writing style and medical condition may not hold for hospitals that are not part of the training data: for example, a hospital chain might open a new facility, which uses the same writing style but experiences a different distribution over medical conditions.

To find examples of spurious correlations, one need not look so far as multi-hospital deployments. Spurious correlations also appear in widely-used benchmarks such as natural language inference, where negation words are correlated with semantic contradictions (Gururangan et al., 2018; Poliak et al., 2018). This correlation is spurious because it arises from the tendency of annotators to use negation words when asked to produce a contradictory sentence in a crowdsourced

elicitation, and it does not hold for text that is produced under more natural conditions.

Such observations have led to several proposals for novel evaluation methodologies (Naik et al., 2018; Ribeiro et al., 2020; Gardner et al., 2020) to ensure that predictors are not "right for the wrong reasons". These evaluations generally take two forms: *invariance tests,* which assess whether predictions are affected by perturbations that are causally unrelated to the label, and *sensitivity tests,* which apply perturbations that should in some sense be the minimal change necessary to flip the true label. Invariance tests can be motivated by a causal intuition: the purpose is to test whether the predictor behaves differently on counterfactual inputs $X(Z = \tilde{z})$, where Z labels a cause of the text that an analyst believes should be causally irrelevant to Y. A model whose predictions are invariant across such counterfactuals can in some cases be expected to perform better on test distributions with a different relationship between Y and Z (Veitch et al., 2021). Similarly, sensitivity tests can be viewed as evaluations of counterfactuals $X(Y = \tilde{y})$, in which the label is changed but all other causal influences on X are held constant (Kaushik et al., 2020).

A number of approaches have been proposed for learning predictors that pass tests of sensitivity and invariance. Many of these approaches are either explicitly or implicitly motivated by a causal perspective; they can be viewed as ways to incorporate domain knowledge of the causal structure of the data into the learning objective. We now survey these approaches, which fall into two main groups: counterfactual data augmentation and causally-motivated distributional criteria.

4.1.1 Data augmentation

To learn predictors that pass tests of invariance and sensitivity, a popular and straightforward approach is *data augmentation*: elicit or construct counterfactual instances, and incorporate them into the training data. In the case of invariance tests, additional focus can be provided by adding a term to the learning objective to explicitly penalize disagreements in the predictions for counterfactual pairs, e.g., $|f(X(Z = z)) - f(X(Z = \tilde{z}))|$, with f indicating the prediction function (Garg et al., 2019). In the case of interventions on the label Y, training on label counterfactuals $X(Y = \tilde{y})$ can improve out-of-domain generalization and reduce sensitivity to noise (Kaushik et al., 2019, 2020).

Counterfactual examples can be generated in several ways: (1) manual post-editing (e.g., Kaushik et al., 2019; Gardner et al., 2020), (2) heuristic replacement of keywords (e.g., Shekhar et al., 2017; Garg et al., 2019), and (3) automated text rewriting (e.g., Zmigrod et al., 2019; Riley et al., 2020; Wu et al., 2021). Manual editing is typically fluent and accurate but relatively expensive. Keyword-based approaches are appropriate in some cases — for example, when counterfactuals can be obtained by making local substitutions of closed-class words like pronouns — but they cannot guarantee fluency or coverage of all labels and covariates of interest (Antoniak and Mimno, 2021), and are difficult to generalize across languages. Fully generative approaches could potentially combine the fluency and coverage of manual editing with the ease of lexical heuristics, but these methods are still relatively immature.

Counterfactual examples are a powerful resource because they directly address the missing data issues that are inherent to causal inference, as described in §2. However, in many cases it is difficult for even a fluent human to produce meaningful counterfactuals: imagine the task of converting a book review into an restaurant review while somehow leaving "everything else" constant. A related concern is that the generation of counterfactuals will introduce new spurious correlations. For example, when asked to rewrite NLI counterfactuals without using negation, annotators (or automated text rewriters) may simply find another shortcut, introducing a new spurious correlation. Similarly, keyword substitution approaches may introduce new spurious correlations if the keyword lexicons are incomplete (Joshi and He, 2021).

4.1.2 Distributional Criteria

The challenges inherent to counterfactual data augmentation motivate alternative approaches that operate directly on the observed data. In the case of invariance tests, one strategy is to derive distributional properties of invariant predictors, and then ensure that these properties are satisfied by the trained model.

Veitch et al. (2021) show that counterfactually-invariant predictors will satisfy independence criteria that can be derived from the causal structure of the data generating process. For example, consider medical record example above, with spurious feature (dialect) Z and label (diagnosis) Y. The desideratum that changing the dialect should not

change the diagnosis can be formalized as *counterfactual invariance* to Z: the predictor f should satisfy $f(X(z)) = f(X(z'))$ for all z, z'. In this case, both Z and Y are causes of the text features X.[3] Using this observation, it can be shown that any counterfactually invariant predictor will satisfy $f(X) \perp\!\!\!\perp Z \mid Y$; i.e., the prediction $f(X)$ is independent of the covariate Z conditioned on the true label Y. In other cases, the label is an effect of the text.

For example, consider data where moderators have flagged forum comments X as toxic (or not) Y, and suppose that toxic comments tend (in training) to have strong negative sentiment Z. In this case, it can be shown that a counterfactually-invariant predictor will satisfy $f(X) \perp\!\!\!\perp Z$ (without conditioning on Y). In this fashion, knowledge of the true causal structure of the problem can be used to derive observed-data signatures of the counterfactual invariance. Such signatures can be incorporated as regularization terms in the training objective (e.g., using kernel-based measures of statistical dependence). These criteria do not guarantee counterfactual invariance — the implication works in the other direction — but in practice they increase counterfactual invariance and improve performance in out-of-distribution settings without requiring the analyst to produce counterfactual examples.

An alternative set of distributional criteria can be derived by viewing the training data as arising from a finite set of *environments*, in which each environment is endowed a unique distribution over causes, but the causal relationship between X and Y is invariant across environments. This view motivates a set of environmental invariance criteria: the predictor should include a representation function that is invariant across environments (Muandet et al., 2013; Peters et al., 2016); we should induce a representation such that the same predictor is optimal in every environment (Arjovsky et al., 2019); the predictor should be equally well calibrated across environments (Wald et al., 2021). Multi-environment training is conceptually similar to domain adaptation (Ben-David et al., 2010), but here the goal is not to learn a predictor for any specific target domain, but rather to learn a predictor that works well across a set of causally-compatible

---

[3]This is sometimes called the *anticausal* setting, because the predictor $f : X \rightarrow \hat{Y}$ must reverse the causal direction of the data generating process (Schölkopf et al., 2012).

domains, known as *domain generalization* (Ghifary et al., 2015; Gulrajani and Lopez-Paz, 2020).

A third alternative to control for confounding between Y and Z is to estimate $P(Y \mid X, Z)$ and then compute a predictive distribution by summing over values of Z, $\tilde{P}(Y \mid X) = \sum_z P(Y \mid X, Z = z) \Pr(Z = z)$, weighting by the marginal $P(Z)$ instead of $P(Z|X)$ (Landeiro and Culotta, 2018). When the labels Y are caused by the text X and Z is a confound, this factorization is closely related to the "backdoor adjustment" for converting observational probabilities into interventional probabilities (Pearl, 2000). Landeiro and Culotta (2018) apply this idea to text classification, demonstrating improved robustness to domain shifts induced by synthetic spurious correlations.

In general these approaches require richer training data than in the typical supervised learning setup: either explicit labels Z for the causes of the text that should not influence the prediction, or access to data gathered from multiple labeled environments. Whether obtaining such data is easier than creating counterfactual instances depends on the situation. Furthermore, the distributional approaches have thus far been applied only to classification problems, while data augmentation can be applied to more structured problems such as machine translation. Future work on distributional approaches must therefore address structured prediction, as well as the setting in which the supplementary annotations are noisy or approximate.

Future work should also consider the consequences of using incomplete causal models, e.g. due to unobserved confounding. Unobserved confounding is challenging for causal inference in general, but it is likely to be ubiquitous in language applications, in which the text arises from the author's intention to express a structured arrangement of semantic concepts, and the label corresponds to a query, either directly on the intended semantics or on the semantics that are likely to be understood by the reader.

4.2 Fairness and bias

NLP systems inherit and sometimes amplify undesirable biases that are encoded in text training data (Barocas et al., 2019; Blodgett et al., 2020). Here too, causality can provide a language for specifying desired fairness conditions across demographic attributes like race and gender. Indeed,

fairness and bias in predictive models have close connections to causality: Hardt et al. (2016) show that a causal analysis is required to determine whether an observed distribution of data and predictions raises fairness concerns; Kilbertus et al. (2017) show that fairness metrics can be motivated by causal interpretations of the data generating process; Kusner et al. (2017) study "counterfactually fair" predictors where, for each individual, predictions are the same for that individual and for a counterfactual version of them created by changing a protected attribute. However, there are important questions about the legitimacy of treating attributes like race as variables subject to intervention or counterfactual reasoning (e.g., Kohler-Hausmann, 2018; Hanna et al., 2020), and Kilbertus et al. (2017) propose to focus instead on invariance to observable proxies such as names.

The fundamental connections between causality and unfair bias have been explored mainly in the context of relatively low-dimensional tabular data rather than text. However, there are several applications of the counterfactual data augmentation strategies from §4.1.1 in this setting: for example, Garg et al. (2019) construct counterfactuals by swapping lists of "identity terms", with the goal of reducing bias in text classification, and Zhao et al. (2018) swap gender markers such as pronouns and names for coreference resolution. Counterfactual data augmentation has also been applied to reduce bias in pre-trained contextualized word embedding models (e.g., Huang et al., 2019; Maudslay et al., 2019) but the extent to which biases in pre-trained models propagate to downstream applications remains unclear (Goldfarb-Tarrant et al., 2021). Fairness applications of the distributional criteria discussed in §4.1.2 are relatively rare, but Adragna et al. (2020) show that invariant risk minimization (Arjovsky et al., 2019) — which attempts to learn an invariant predictor across multiple "environments" — can reduce the use of spurious correlations with race on the Civil Comments dataset (Borkan et al., 2019) for toxicity detection.

4.3 Causal Model Interpretations

NLP models are notoriously difficult to explain, but it is nonetheless crucial to diagnose errors and establish trust with decision makers (Guidotti et al., 2018). One prominent approach to generate explanations is to exploit network artifacts, such as attention weights (Bahdanau et al., 2014), which

are computed on the path to generating a prediction (e.g., Xu et al., 2015; Wang et al., 2016). Alternatively, there have been attempts to estimate simpler and more interpretable models by using perturbations of test examples or their hidden representations (Ribeiro et al., 2016; Lundberg and Lee, 2017; Kim et al., 2018). However, both attention and perturbation-based methods have important limitations. Attention-based explanations can be misleading (Jain and Wallace, 2019), and are generally possible only for individual tokens; they cannot explain predictions in terms of more abstract linguistic concepts. Existing perturbation-based methods often generate implausible counterfactuals and also do not allow for estimating the effect of sentence-level concepts.

Viewed as a causal inference problem, a natural approach to explanation is to generate counterfactual examples (see §4.1.1) and then compare the prediction for each example and its counterfactual. Such a controlled setting is similar to the randomized experiment described in §2, where it is possible to compute the difference between an actual observed text, and what the text would have been had a specific concept not existed in it. Indeed, in cases where counterfactual texts can be generated, we can often estimate causal effects on text-based models (Ribeiro et al., 2020; Ross et al., 2021; Gardner et al., 2020). However, generating natural language counterfactuals is often too hard to do automatically, and too costly to do manually, particularly for abstract concepts such as linguistic style, topic or sentiment.

To overcome the counterfactual generation problem, another class of approaches proposes to manipulate the representation of the text and not the text itself (Feder et al., 2021; Elazar et al., 2021; Ravfogel et al., 2021). These solutions share similarities with methods from the robustness literature (e.g., Muandet et al., 2013), but are focused with identifying invariances in a given trained model, and not with enforcing them during training. Feder et al. (2021) compute the counterfactual representation by pre-training an additional instance of the language representation model employed by the classifier, with an adversarial component designed to "forget" the concept of choice, while controlling for confounding concepts. Ravfogel et al. (2020) offered a method for removing information from neural representations by iteratively training linear classifiers and projecting the representations on their null-spaces.

Rather than using counterfactuals to identify invariances, a complementary approach is to *generate* counterfactuals with minimal changes that obtain a different model prediction (Wachter et al., 2017; Karimi et al., 2021; Mothilal et al., 2020). Such examples serve as explanations as they allow us to observe the changes required to change a model's prediction.

Finally, a causal perspective on the attention-based explanations mentioned earlier is to view internal nodes as *mediators* on the causal effect from the input on the output (Vig et al., 2020; Finlayson et al., 2021). By querying models using manually-crafted counterfactuals, we can observe how information flows through different model components and identify where in the model is it encoded.

5   Conclusion

Our main goal in this survey was to collect the various touchpoints of causality and NLP into one space, which we then subdivided into the problems of statistical causal inference and more traditional NLP tasks. These branches of scientific inquiry share common goals, intuitions, and are beginning to show methodological synergies. In Section 3 we showed how recent advances in NLP modeling can help researchers make causal conclusions with text data, and in Section 4 we showed how ideas from causal inference can be used to make NLP models more trustworthy and transparent. Both of these spaces remain nascent with a large number of open challenges which we have detailed throughout this paper.

The causal methodology forces practitioners to explicate their assumptions. To improve scientific standards, we believe that the computational linguistics community should be clearer about these assumptions and analyze their data using causal reasoning. Pushing our language processing methodologies in this direction could lead to a better understanding of language and the models we build to process it.

References

Robert Adragna, Elliot Creager, David Madras, and Richard Zemel. 2020. Fairness and robustness in invariant learning: A case study in toxicity classification. *arXiv preprint arXiv:2011.06485*.

Maria Antoniak and David Mimno. 2021. Bad seeds: Evaluating lexical methods for bias measurement. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1889–1904, Online. Association for Computational Linguistics.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. 2019. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2019. *Fairness and Machine Learning*. fairmlbook.org. http://www.fairmlbook.org.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. 2010. A theory of learning from different domains. *Machine learning*, 79(1):151–175.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in nlp. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019. Nuanced metrics for measuring unintended bias with real data for text classification. *CoRR*, abs/1903.04561.

Peter Bühlmann. 2020. Invariance, causality and robustness. *Statistical Science*, 35(3):404–426.

Alexander D'Amour, Peng Ding, Avi Feller, Lihua Lei, and Jasjeet Sekhon. 2020. Overlap in observational studies with high-dimensional covariates. *Journal of Econometrics*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Rotem Dror, Gili Baumer, Marina Bogomolov, and Roi Reichart. 2017. Replicability analysis for natural language processing: Testing significance with multiple datasets. *Transactions of the Association for Computational Linguistics*, 5:471–486.

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

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. 2021. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175.

Amir Feder, Nadav Oved, Uri Shalit, and Roi Reichart. 2021. Causalm: Causal model explanation through counterfactual language models. *Computational Linguistics*, 47(2):333–386.

Matthew Finlayson, Aaron Mueller, Sebastian Gehrmann, Stuart Shieber, Tal Linzen, and Yonatan Belinkov. 2021. Causal analysis of syntactic agreement mechanisms in neural language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1828–1843, Online. Association for Computational Linguistics.

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)*, pages 1600–1609.

Christian Fong and Justin Grimmer. 2021. Causal inference with latent treatments. *American Journal of Political Science*. Forthcoming.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen,

Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. 2020. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1307–1323, Online. Association for Computational Linguistics.

Matt Gardner, William Merrill, Jesse Dodge, Matthew E Peters, Alexis Ross, Sameer Singh, and Noah Smith. 2021. Competency problems: On finding and removing artifacts in language data. *arXiv preprint arXiv:2104.08646*.

Sahaj Garg, Vincent Perot, Nicole Limtiaco, Ankur Taly, Ed H Chi, and Alex Beutel. 2019. Counterfactual fairness in text classification through robustness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 219–226.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. 2020. Shortcut learning in deep neural networks. *arXiv preprint arXiv:2004.07780*.

Alan S Gerber, Donald P Green, and Christopher W Larimer. 2008. Social pressure and voter turnout: Evidence from a large-scale field experiment. *American political Science review*, 102(1):33–48.

Muhammad Ghifary, W Bastiaan Kleijn, Mengjie Zhang, and David Balduzzi. 2015. Domain generalization for object recognition with multi-task autoencoders. In *Proceedings of the IEEE international conference on computer vision*, pages 2551–2559.

Seraphina Goldfarb-Tarrant, Rebecca Marchant, Ricardo Muñoz Sánchez, Mugdha Pandya, and Adam Lopez. 2021. Intrinsic bias metrics do not correlate with application bias. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1926–1940, Online. Association for Computational Linguistics.

Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):1–42.

Ishaan Gulrajani and David Lopez-Paz. 2020. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R Bowman, and Noah A Smith. 2018. Annotation artifacts in natural language inference data. *arXiv preprint arXiv:1803.02324*.

Alex Hanna, Emily Denton, Andrew Smart, and Jamila Smith-Loud. 2020. Towards a critical race methodology in algorithmic fairness. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 501–512.

Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29:3315–3323.

Po-Sen Huang, Huan Zhang, Ray Jiang, Robert Stanforth, Johannes Welbl, Jack Rae, Vishal Maini, Dani Yogatama, and Pushmeet Kohli. 2019. Reducing sentiment bias in language models via counterfactual evaluation. *arXiv preprint arXiv:1911.03064*.

Alon Jacovi, Ana Marasović, Tim Miller, and Yoav Goldberg. 2021. Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in ai. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 624–635.

Sarthak Jain and Byron C Wallace. 2019. Attention is not explanation. *arXiv preprint arXiv:1902.10186*.

Nitish Joshi and He He. 2021. An investigation of the (in) effectiveness of counterfactually augmented data. *arXiv preprint arXiv:2107.00753*.

Amir-Hossein Karimi, Bernhard Schölkopf, and Isabel Valera. 2021. Algorithmic recourse:

from counterfactual explanations to interventions. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 353–362.

Divyansh Kaushik, Eduard Hovy, and Zachary C Lipton. 2019. Learning the difference that makes a difference with counterfactually-augmented data. *arXiv preprint arXiv:1909.12434*.

Divyansh Kaushik, Amrith Setlur, Eduard Hovy, and Zachary C Lipton. 2020. Explaining the efficacy of counterfactually-augmented data. *arXiv preprint arXiv:2010.02114*.

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

Niki Kilbertus, Mateo Rojas-Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. 2017. Avoiding discrimination through causal reasoning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 656–666.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. 2018. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International Conference on Machine Learning*, pages 2668–2677.

Byungju Kim, Hyunwoo Kim, Kyungsu Kim, Sungjin Kim, and Junmo Kim. 2019. Learning not to learn: Training deep neural networks with biased data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9012–9020.

Issa Kohler-Hausmann. 2018. Eddie murphy and the dangers of counterfactual causal thinking about detecting racial discrimination. *Nw. UL Rev.*, 113:1163.

Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. In *Advances in neural information processing systems*, pages 4066–4076.

Virgile Landeiro and Aron Culotta. 2018. Robust text classification under confounding shift.

Journal of Artificial Intelligence Research, 63:391–419.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in neural information processing systems*, pages 4765–4774.

Daniel Maliniak, Ryan Powers, and Barbara F Walter. 2013. The gender citation gap in international relations. *International Organization*, 67(4):889–922.

Rowan Hall Maudslay, Hila Gonen, Ryan Cotterell, and Simone Teufel. 2019. It's all in the name: Mitigating gender bias with name-based counterfactual data substitution. *arXiv preprint arXiv:1909.00871*.

R Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. *arXiv preprint arXiv:1902.01007*.

Stephen L Morgan and Christopher Winship. 2015. *Counterfactuals and causal inference*. Cambridge University Press.

Ramaravind K Mothilal, Amit Sharma, and Chenhao Tan. 2020. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 607–617.

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

Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. 2013. Domain generalization via invariant feature representation. In *International Conference on Machine Learning*, pages 10–18.

Aakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig. 2018. Stress test evaluation for natural language inference. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2340–2353, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Judea Pearl. 1994. A probabilistic calculus of actions. In *Uncertainty Proceedings 1994*, pages 454–462. Elsevier.

Judea Pearl. 2000. Causality: Models, reasoning and inference. *Cambridge, MA, USA,*, 9:10–11.

Judea Pearl. 2009. *Causality*. Cambridge university press.

J Peters, P Bühlmann, and N Meinshausen. 2016. Causal inference using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society-Statistical Methodology-Series B*, 78(5):947–1012.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 2227–2237. Association for Computational Linguistics.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. *arXiv preprint arXiv:1805.01042*.

Reid Pryzant, Dallas Card, Dan Jurafsky, Victor Veitch, and Dhanya Sridhar. 2020. Causal effects of linguistic properties. *arXiv preprint arXiv:2010.12919*.

Reid Pryzant, Youngjoo Chung, and Dan Jurafsky. 2017. Predicting sales from the language of product descriptions. In *eCOM@ SIGIR*.

Reid Pryzant, Kelly Shen, Dan Jurafsky, and Stefan Wagner. 2018. Deconfounded lexicon induction for interpretable social science. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1615–1625.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. Null it out: Guarding protected attributes by iterative nullspace projection. *arXiv preprint arXiv:2004.07667*.

Shauli Ravfogel, Grusha Prasad, Tal Linzen, and Yoav Goldberg. 2021. Counterfactual interventions reveal the causal effect of relative clause representations on agreement prediction. *arXiv preprint arXiv:2105.06965*.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144. ACM.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Parker Riley, Noah Constant, Mandy Guo, Girish Kumar, David Uthus, and Zarana Parekh. 2020. Textsettr: Label-free text style extraction and tunable targeted restyling. *arXiv preprint arXiv:2010.03802*.

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

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.

Alexis Ross, Tongshuang Wu, Hao Peng, Matthew E Peters, and Matt Gardner. 2021. Tailor: Generating and perturbing text with semantic controls. *arXiv preprint arXiv:2107.07150*.

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

B Schölkopf, D Janzing, J Peters, E Sgouritsa, K Zhang, and J Mooij. 2012. On causal and anticausal learning. In *29th International Conference on Machine Learning (ICML 2012)*, pages 1255–1262. International Machine Learning Society.

Bernhard Schölkopf. 2019. Causality for machine learning. *arXiv preprint arXiv:1911.10500*.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. 2021. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634.

Ravi Shekhar, Sandro Pezzelle, Yauhen Klimovich, Aurélie Herbelot, Moin Nabi, Enver Sangineto, and Raffaella Bernardi. 2017. FOIL it! find one mismatch between image and language caption. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 255–265, Vancouver, Canada. Association for Computational Linguistics.

Dhanya Sridhar and Lise Getoor. 2019. Estimating causal effects of tone in online debates. In *International Joint Conference on Artificial Intelligence*.

Victor Veitch, Alexander D'Amour, Steve Yadlowsky, and Jacob Eisenstein. 2021. Counterfactual invariance to spurious correlations: Why and how to pass stress tests. *arXiv preprint arXiv:2106.00545*.

Victor Veitch, Dhanya Sridhar, and David M Blei. 2020. Adapting text embeddings for causal inference. In *UAI*.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart M. Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2017. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech.*, 31:841.

Yoav Wald, Amir Feder, Daniel Greenfeld, and Uri Shalit. 2021. On calibration and out-of-domain generalization. *arXiv preprint arXiv:2102.10395*.

Yequan Wang, Minlie Huang, Xiaoyan Zhu, and Li Zhao. 2016. Attention-based LSTM for aspect-level sentiment classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 606–615, Austin, Texas. Association for Computational Linguistics.

Zhao Wang and Aron Culotta. 2020. Identifying spurious correlations for robust text classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3431–3440, Online. Association for Computational Linguistics.

Zach Wood-Doughty, Ilya Shpitser, and Mark Dredze. 2018. Challenges of using text classifiers for causal inference. In *EMNLP*.

Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel S Weld. 2021. Polyjuice: Automated, general-purpose counterfactual generation. *arXiv preprint arXiv:2101.00288*.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR.

Justine Zhang, Sendhil Mullainathan, and Cristian Danescu-Niculescu-Mizil. 2020a. Quantifying the causal effects of conversational tendencies. In *CSCW*.

Justine Zhang, Sendhil Mullainathan, and Cristian Danescu-Niculescu-Mizil. 2020b. Quantifying the causal effects of conversational tendencies. In *Proceedings of CSCW*.

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*, pages 2979–2989, Copenhagen, Denmark. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics.

Ran Zmigrod, Sabrina J. Mielke, Hanna Wallach, and Ryan Cotterell. 2019. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1651–1661, Florence, Italy. Association for Computational Linguistics.