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

Anonymous TACL submission

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 challenges and opportunities in the application of causal inference to the textual domain, with its unique properties. 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 with text, 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 NLP community.

1 Introduction

The increasing effectiveness of NLP has created exciting new opportunities for interdisciplinary collaborations, bringing NLP techniques to a wide range of external research disciplines (e.g., Roberts et al., 2014; Zhang et al., 2020; Ophir et al., 2020) and incorporating new data and tasks into mainstream NLP (e.g., Thomas et al., 2006; Pryzant et al., 2018). In such interdisciplinary collaborations, many of the most important research questions relate to the inference of causal relationships. 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, with observational data, the causal effect is not equivalent to the correlation between whether the drug is taken and the observed disease progression. There is now a vast 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 fundamental challenges.

Conversely, in many classical NLP applications, the main goal is to make accurate predictions: Any statistical correlation is 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.

The 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 areas: Estimating causal effects from text, and using causal formalisms to make NLP methods more reliable. We next illustrate this distinction.

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

Ex. 1 addresses the causal effect of signaling female gender (treatment) on the likes a post receives (outcome) (see discussion on signaling at (Keith et al., 2020)). The counterfactual question is: If we could manipulate the gender icon of a post, how many likes would the post have received?

The observed correlation between the gender icons and the number of "likes" generally does not coincide with the causal effect: it might instead be a *spurious correlation*, induced by other variables, known as confounders, which are correlated with both the treatment and the outcome (see Gururangan et al., 2018, for an early discussion of spurious correlation in NLP). One possible confounder is the topic of each post: Posts written by users who have selected the female icon may be about certain topics (e.g., child birth or menstruation) more often, and those topics may not receive as many likes from the audience of the broader online platform. As we will see in § 2, due to confounding, estimating a causal effect requires 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 gender signal 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 causal inference. Causal inference with text data involves several challenges that are 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 (Blei et al., 2003) to contextual embeddings (e.g. (Devlin et al., 2019)), 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 discuss existing research on estimating causal effects from text and emphasize these challenges and opportunities in § 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 Ex. 1, Ex. 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).

Causal models 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 §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, separating the intersection of causality and NLP into two distinct lines of research on estimating causal effects in which text is at least one causal variable (§3) and using causal formalisms to improve robustness and interpretability in NLP methods (§4). 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; and 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 causal formalisms for robust and explainable prediction) 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 from §1.

Example 1 involves online forum posts and the number of likes Y that they receive. We use a binary variable T to indicate whether a post uses a "female icon" ($T = 1$) or a "male icon" ($T = 0$). We view the post icon T as the "treatment" in this example, but do not assume that the treatment is randomly assigned (it may be selected by the posts' authors). The *counterfactual* outcome $Y(1)$ represents the number of likes a post would have received had it used a female icon. The counterfactual outcome $Y(0)$ is defined analogously.

The *fundamental problem of causal inference* (Holland, 1986) is that we can never observe $Y(0)$ and $Y(1)$ simultaneously for any *unit of analysis*, the smallest unit about which one wants to make counterfactual inquiries (e.g. a post in Ex. 1). This problem is what makes causal inference harder than statistical inference and impossible without identification assumptions (see § 2.2).

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

2.1 Causal Estimands

An analyst begins by specifying target causal quantities of interest, called causal estimands, which typically involve counterfactuals. In Example 1, one possible causal estimand is the average treatment effect (ATE),

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

where the expectation is over the generative distribution of posts. The ATE can be interpreted as the change in the number of likes a post would have received, on average, had the post used a female icon instead of a male icon.

Another possible causal effect of interest is the conditional average treatment effect (CATE),

$$\text{CATE} = \mathbb{E}[Y(1) - Y(0) \mid G]. \quad (2)$$

where G is a predefined subgroup of the population. For example, G could be all posts on political topics. In this case, the CATE can be interpreted as the change in the number of likes a post on a political topic would have received, on average, had the post used a male icon instead of a female icon. CATEs are used to quantify the heterogeneity of causal effects in different population subgroups.

2.2 Identification Assumptions for Causal Inference

We will focus on Example 1 and the ATE in Equation (1) to explain the assumptions needed for causal inference. Although we focus on the ATE, related assumptions are needed in some form for

all causal estimands. Variables are the same as those defined previously in this section.

Ignorability requires that the treatment assignment be statistically independent of the counterfactual outcomes,

$$T \perp\!\!\!\perp Y(a) \quad \forall a \in \{0,1\}. \tag{3}$$

Note that this assumption is not equivalent to independence between the treatment assignment and the *observed* outcome Y. For example, if ignorability holds, $Y \perp\!\!\!\perp T$ would *additionally* imply that the treatment has no effect.

Randomized treatment assignment guarantees ignorability by design. For example, we can guarantee ignorability in Example 1 by flipping a coin to select the icon for each post, and disallowing post authors from changing it.

Without randomized treatment assignment, ignorability could be violated by *confounders*, variables that influence both the treatment status and potential outcomes. In Example 1, suppose that: (i) the default post icon is male, (ii) only experienced users change the icon for their posts based on their gender, (iii) experienced users write posts that receive relatively more likes. In this scenario, the experience of post authors is a confounder: posts having female icons are more likely to be written by experienced users and thus, receive more likes. In the presence of confounders, causal inference is only possible if we assume *conditional ignorability*,

$$T \perp\!\!\!\perp Y(a) \mid X \quad \forall a \in \{0,1\} \tag{4}$$

where X is a set of observed variables, conditioning on which ensures independence between the treatment assignment and the potential outcomes. In other words, we can assume that all confounders are observed.

Positivity requires that the probability of receiving treatment is bounded away from 0 and 1 for all values of the confounders X:

$$0 < \Pr(T = 1 \mid X = x) < 1, \forall x \tag{5}$$

Intuitively, positivity requires that each unit under study has the possibility of being treated and has the possibility of being untreated. Randomized treatment assignment can also guarantee positivity by design.

Consistency requires that the outcome observed for each unit under study at treatment level $a \in \{0,1\}$ is identical to the outcome we would have observed had that unit been assigned to treatment level a,

$$T = a \Leftrightarrow Y(a) = Y \quad \forall a \in \{0,1\} \tag{6}$$

Consistency ensures that the potential outcomes for each unit under study take on a single value at each treatment level. Consistency will be violated if different unobservable "versions" of the treatment lead to different potential outcomes. For example, if red and blue female icons had different effects on the number of likes received, but icon color was not recorded. Consistency will also be violated if the treatment assignment of one unit affects the potential outcomes of another; a phenomenon called *interference* (Rosenbaum, 2007). Randomized treatment assignment does not guarantee consistency by design. e.g. if different icon colors affect the number of likes but are not considered by the model, then a randomized experiment will not solve the problem. As Hernán (2016) discusses, consistency assumptions are a "matter of expert agreement" and, while subjective, these assumptions are at least made more transparent by causal formalisms.

These three assumptions enable identifying the ATE defined in Equation (1), as formalized in the following identification proof:

$$\begin{aligned}\mathbb{E}[Y(a)] &\overset{(i)}{=} \mathbb{E}_X[\mathbb{E}[Y(a) \mid X]] \\ &\overset{(ii)}{=} \mathbb{E}_X[\mathbb{E}[Y(a) \mid X, T = a]] \\ &\overset{(iii)}{=} \mathbb{E}_X[\mathbb{E}[Y \mid X, T = a]], \forall a \in \{0,1\}\end{aligned}$$

where equality (i) is due to iterated expectation, equality (ii) follows from conditional ignorability, and equality (iii) follows from consistency and positivity, which ensures that the conditional expectation $\mathbb{E}[Y \mid X,\ T = a]$ is well defined. The final expression can be computed from observable quantities alone.

We refer to other background material to discuss how to identify and estimate causal effects with these assumptions in hand (Rubin, 2005; Pearl, 2009; Imbens and Rubin, 2015; Egami et al., 2018; Keith et al., 2020).

2.3 Causal graphical models

Finding a set of variables X that ensure conditional ignorability is challenging, and requires making several carefully assessed assumptions

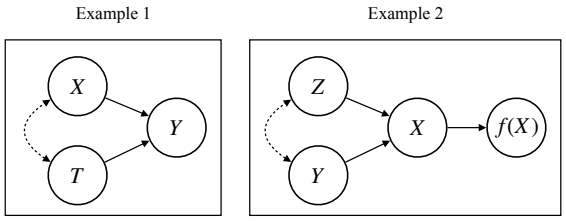

Figure 1: Causal graphs for the motivating examples. (**Left**) In Example 1, the post icon (T) is correlated with attributes of the post (X), and both variables affect the number of likes a post receives (Y). (**Right**) In Example 2, the label (Y, i.e., diagnosis) and hospital site (Z) are correlated, and both affect the clinical narrative (X). Predictions $f(X)$ from the trained classifier depend on X.

about the causal relationships in the domain under study. Causal directed-acyclic graphs (DAGs) (Pearl, 2009) enable formally encoding these assumptions and deriving the set of variables X after conditioning on which ignorability is satisfied.

In a causal DAG, an edge $X \rightarrow Y$ implies that X *may or may not* cause Y. The absence of an edge between X and Y implies that X *does not* cause Y. Bi-directed dotted arrows between variables indicate that they are correlated potentially through some unobserved variable.

Figure 1 illustrates the causal DAGs we assume for Example 1 and Example 2. Given a causal DAG, causal dependencies between any pair of variables can be derived using the d-separation algorithm (Pearl, 1994). These dependencies can then be used to assess whether conditional ignorability holds for a given treatment, outcome, and set of conditioning variables X. For example, in the left DAG in Figure 1, the post icon T is not independent of the number of likes Y unless we condition on X. In the right DAG, the prediction $f(X)$ is not independent of the hospital Z even after conditioning on the narrative X.

3 Estimating Causal Effects with Text

In §2, we described assumptions for causal inference when the treatment, outcome and confounders were directly measured. In this section, we contribute a novel discussion about how causal assumptions are complicated when variables necessary for a causal analysis are extracted automatically from text. Addressing these open challenges will require collaborations between the NLP and causal estimation communities to understand how the requisite assumptions to draw valid causal conclusions. We highlight prior approaches and future challenges in settings where the text is a confounder, the outcome, or the treatment – but this discussion applies broadly to many text-based causal problems.

To make these challenges clear, we will expand upon Example 1 by supposing a hypothetical online forum wants to understand and reduce harassment on its platform. Many such questions are causal: Do gendered icons influence the harassment users receive? Do longer suspensions make users less likely to harass others? How can a post be rewritten to avoid offending others? In each case, using NLP to measure aspects of language is integral to any causal analysis.

3.1 Text as Confounder

Returning to Example 1, suppose the platform worries that users with female icons are more likely to receive harassment from other users. Such a finding might significantly influence plans for a new moderation strategy (Jhaver et al., 2018; Rubin et al., 2020). We may be unable or unwilling to randomize our treatment (the gender signal of the author's icon), so the causal effect of gender signal on harassment received might be confounded by other variables. The topic of the post may be an important confounder: some subject areas may be discussed by a larger proportion of users with female icons, and more controversial subjects may attract more harassment. The text of the post provides evidence of the topic and thus acts as a confounder (Roberts et al., 2020).

Previous approaches. The main idea in this setting is to use NLP methods to extract confounding aspects from text and then adjust for those aspects in an estimation approach such as propensity score matching. However, how and when these methods violate causal assumptions are still open questions. Keith et al. (2020) provides a recent overview of several such methods and many potential threats to inference.

One set of methods apply 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) matches texts using distance metrics on the bag-of-words representation.

A second set of methods adjust for confounders from text with supervised NLP methods. Recently, Veitch et al. (2020) adapted pre-trained language models and supervised topic models with multiple classification heads for binary treatment and counterfactual outcomes. By learning a "sufficient" embedding that obtained low classification loss on the treatment and counterfactual outcomes, they show 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.

Challenges for causal assumptions with text. In settings without randomized treatments, NLP methods that adjust for text confounding require a particularly strong statement of conditional ignorability (Equation 4): all aspects of confounding must be measured by the model. Because we cannot test this assumption, we should seek domain expertise to justify it or understand the theoretical and empirical consequences if it is violated.

When the text is a confounder, its high-dimensionality makes positivity unlikely to hold (D'Amour et al., 2020). Even for approaches that extract a low-dimensional representation of the confounder from text, positivity is a concern. For example, in Example 1, posts might contain phrases that near-perfectly encode the chosen gender-icon of the author. If the learned representation captures this information alongside other confounding aspects, it would be nearly impossible to imagine changing the gender icon while holding the gendered text fixed.

3.2 Text as Outcome

Suppose platform moderators can choose to suspend users who violate community guidelines for either one day or one week, and we want to know which option has the greatest effect at decreasing the toxicity of the suspended user. If we could collect them for each user's post, ground-truth human annotations of toxicity would be our ideal outcome variable. We would then use those outcomes to calculate the ATE, following the discussion in § 2. Our analysis of suspensions is complicated if, instead of ground-truth labels for our toxicity outcome, we rely on NLP methods to extract the outcome from the text. A core challenge is to distill the high-dimensional text into a low-dimensional measure of toxicity.

Challenges for causal assumptions with text. We saw in § 2 that randomizing the treatment assignment can ensure ignorability and positivity; but even with randomization, we require more careful assessment to satisfy consistency. Suppose we randomly assign suspension lengths to users and then once those users return and continue to post, we use a clustering method to discover toxic and non-toxic groupings among the formerly-suspended users. To estimate the causal effect of suspension length, we rely on the trained clustering model to infer our outcome variable. Assuming that the suspension policy does in truth have a causal effect on posting behavior, then because our clustering model depends on all posts in its training data, it also depends on the treatment assignments that influenced each post. Thus, when we use the model to infer outcomes, each user's outcome depends on all other users' treatments. This violates the assumption of consistency – that potential outcomes do not depend on the treatment status of other units. This undermines the theoretical basis for our causal estimate, and in practice, implies that different randomized treatment assignments could lead to different treatment effect estimates. These issues can be addressed by developing the measure on only a sample of the data and then estimating the effect on a separate, held-out data sample (Egami et al., 2018).

3.3 Text as Treatment

As a third example, suppose we want to understand what makes a post offensive. This might allow the platform to provide automated suggestions that encourage users to rephrase their post. Here, we are interested in the causal effect of the text itself on whether a reader reports it as offensive. Theoretically, the counterfactual $Y(t)$ is defined for any t, but could be limited to an exploration of specific *aspects* of the text. For example, do second-person pronouns make a post more likely to be reported?

Previous approaches. One approach to studying the effects of text involves treatment *discovery*: producing interpretable features of the text—such as latent topics 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 biogra-

phies 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. (2020) discovered conversational tendencies that lead to positive mental health counseling sessions.

Another approach is to estimate the causal effects of specific latent properties that are intervened on during an experiment or extracted from text for observational studies (Pryzant et al., 2021; 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.

Challenges for causal assumptions with text. Ensuring positivity and consistency remains a challenge in this setting, but assessing conditional ignorability is particularly tricky. Suppose the treatment is the use of second-person pronouns, but the relationship between this treatment and the outcome is confounded by other properties of the text (e.g., politeness). For conditional ignorability to hold, we would need to extract from the text and condition on all such confounders, which requires assuming that we can disentangle the treatment from many other aspects of the text (Pryzant et al., 2021). Such concerns could be avoided by randomly assigning texts to readers (Fong and Grimmer, 2016, 2021), but that may be impractical. Even if we could randomize the assignment of texts, we still have to assume that there is no confounding due to latent properties of the reader, such as their political ideology or their tastes.

3.4 Future Work

We next highlight key challenges and opportunities for NLP researchers to facilitate causal inference from text.

Heterogeneous effects. Texts are read and interpreted differently by different people; NLP researchers have studied this problem in the context of heterogeneous perceptions of annotators (Paun et al., 2018; Pavlick and Kwiatkowski, 2019). In the field of causal inference, the idea that different subgroups experience different causal effects is formalized by a heterogeneous treatment effect, and is studied using conditional average treatment effects (Equation (2)) for different subgroups. It may also be of interest to discover subgroups where the treatment has a strong effect on an outcome of interest. For example, we may want to identify text features that characterize when a treatment such as a content moderation policy is effective. Wager and Athey (2018) proposed a flexible approach to estimating heterogeneous effects based on random forests. However, such approaches, which are developed with tabular data in mind, may be computationally infeasible for high-dimensional text data. There is an opportunity to extend NLP methods to discover text features that capture subgroups where the causal effect varies.

Representation learning. Causal inference from text requires extracting low-dimensional features from text. Depending on the setting, the low-dimensional features are tasked with extracting confounding information, outcomes or treatments. The need to measure latent aspects from text connects to the field of text representation learning (Le and Mikolov, 2014; Liu et al., 2015; Liu and Lapata, 2018). The usual objective of text representation learning approaches is to model language. Adapting representation learning for causal inference offers open challenges; for example, we might augment the objective function to ensure that (i) positivity is satisfied, (ii) confounding information is not discarded, or (iii) noisily-measured outcomes or treatments enable accurate causal effect estimates.

Benchmarks. Benchmark datasets have propelled machine learning forward by creating shared metrics by which predictive models can be evaluated. There are currently no real-world text-based causal estimation benchmarks due to the *fundamental problem of causal inference* that we can never obtain counterfactuals on an individual and observe the true causal effects. However, as Keith et al. (2020) discuss, there has been some progress in evaluating text-based estimation methods on semi-synthetic datasets in which real covariates are used to generate treatment and outcomes, e.g. Veitch et al. (2020); Roberts et al. (2020); Pryzant et al. (2021); Feder et al. (2021); Weld et al. (2022). Wood-Doughty et al. (2021) employed large-scale language models for controlled synthetic generation of text on which causal methods can be evaluated. An open problem is the degree to which methods that perform well on synthetic data generalize to real-world data.

Controllable Text Generation. When running a randomized experiment or generating syn-

thetic data, researchers make decisions using the empirical distribution of the data. If we are studying whether a drug prevents headaches, it would make sense to randomly assign a 'reasonable' dose – one that is large enough to plausibly be effective but not so large as to be toxic. But when the causal question involves natural language, domain knowledge might not provide a small set of 'reasonable' texts. Instead, we might turn to controllable text generation to sample texts that fulfill some requirements (Kiddon et al., 2016). Such methods have a long history in NLP; for example, a conversational agent should be able to answer a user's question while being perceived as polite (Niu and Bansal, 2018). In our text as treatment example where we want to understand which textual aspects make a text offensive, such methods could enable an experiment allowing us to randomly assign texts that differ on only a specific latent aspect. For example, we could change the style of a text while holding its content fixed (Logeswaran et al., 2018). Recent work has explored text generation from a causal perspective (Hu and Li, 2021), but future work could develop these methods for causal estimation.

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. In this section, we consider using causal reasoning to help solve traditional NLP tasks such as understanding, manipulating, and generating natural language.

At a first glance, NLP may appear to have little need for causal ideas. The field has achieved remarkable progress 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 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 mitigate their impact (§ 4.1); causality also provides a language for specifying and reasoning about fairness conditions (§ 4.2); and the task of explaining predictions may be naturally formulated in terms of counterfactuals (§ 4.3). The application of causality to these problems is still an active area of research, which we attempt to facilitate by highlighting previously implicit connections among a diverse body of prior work.

4.1 Learning Robust Predictors

The NLP field has grown increasingly concerned with *spurious correlations* (Gururangan et al., 2018; McCoy et al., 2019, inter alia). From a causal perspective, 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. A predictor $f : X \rightarrow Y$ will learn to use parts of X that carry information about Z (because Z is informative about Y), which can lead to errors if the relationship between Y and Z changes when the predictor is deployed.[1]

This issue is illustrated by Example 2, where the task is to predict a medical condition from the text of patient records. The training set is drawn from multiple hospitals which vary both in the frequency of the target clinical condition (Y) and the writing style of the narratives (represented in X). A predictor trained on such data will use textual features that carry information about the hospital (Z), even when they are useless at predicting the diagnosis *within* any individual hospital. Spurious correlations also appear as *artifacts* in benchmarks

[1] From the perspective of earlier work on domain adaptation (Søgaard, 2013), spurious correlations can be viewed as a special case of a more general phenomenon in which feature-label relationships change across domains. For example, the lexical feature *boring* might have a stronger negative weight in reviews about books than about kitchen appliances, but this is not a spurious correlation because there is a direct causal relationship between this feature and the label. Spurious correlations are a particularly important form of distributional shift in practice because they can lead to inconsistent predictions on pairs of examples that humans view as identical.

for tasks such as natural language inference, where negation words are correlated with semantic contradictions in crowdsourced training data but not in text that is produced under more natural conditions (Gururangan et al., 2018; Poliak et al., 2018).

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. Both types of test can be motivated by a causal perspective. The purpose of an invariance test is to determine whether the predictor behaves differently on counterfactual inputs $X(Z = \tilde{z})$, where Z indicates a property 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 Y is changed but all other causal influences on X are held constant (Kaushik et al., 2020). Features that are spuriously correlated with Y will be identical in the factual X and the counterfactual $X(Y = \tilde{y})$. A predictor that relies solely on such spurious correlations will be unable to correctly label both factual and counterfactual instances.

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 knowledge of the causal structure of the data into the learning objective.

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. When the counterfactuals involve perturbations to confounding factors Z, it can help to add a term to the learning objective to explicitly penalize disagreements in the predictions for counterfactual pairs, e.g., $|f(X(Z = z)) - f(X(Z = \tilde{z}))|$, when f is the prediction function (Garg et al., 2019). When perturbations are applied to the label Y, training on label counterfactuals $X(Y = \tilde{y})$ can improve OOD generalization and reduce noise sensitivity (Kaushik et al., 2019, 2020; Jha et al., 2020).[2]

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; Feder et al., 2021), and (3) automated text rewriting (e.g., Zmigrod et al., 2019; Riley et al., 2020; Wu et al., 2021; Calderon et al., 2022). 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.

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 a restaurant review while somehow leaving "everything else" constant (as in Calderon et al. (2022)). A related concern is lack of precision in specifying the desired impact of the counterfactual. To revise a text from, say, U.S. to U.K. English, it is unambiguous that "colors" should be replaced with "colours", but should terms like "congress" be replaced with analogous concepts like "parliament"? This depends on whether we view the semantics of the text as a causal descendent of the locale. If such decisions are left to the annotators' intuitions, it is difficult to ascertain what robustness guarantees

[2] More broadly, there is a long history of methods that elicit or construct new examples and labels with the goal of improving generalization, e.g. self-training (McClosky et al., 2006; Reichart and Rappoport, 2007), co-training (Steedman et al., 2003), and adversarial perturbations (Ebrahimi et al., 2018). The connection of such methods to causal issues such as spurious correlations has not been explored until recently (Chen et al., 2020; Jin et al., 2021).

we can get from counterfactual data augmentation. Finally, there is the possibility that counterfactuals will introduce new spurious correlations. For example, when asked to rewrite NLI examples without using negation, annotators (or automated text rewriters) may simply find another shortcut, introducing a new spurious correlation. Keyword substitution approaches may also introduce new spurious correlations if the keyword lexicons are incomplete (Joshi and He, 2021). Automated methods for conditional text rewriting are generally not based on a formal counterfactual analysis of the data generating process (cf. Pearl, 2009), which would require modeling the relationships between various causes and consequences of the text. The resulting counterfactual instances may therefore fail to fully account for spurious correlations and may introduce new spurious correlations.

4.1.2 Distributional Criteria

An alternative to data augmentation is to design new learning algorithms 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.

Given observations of the potential confounder at training time, the counterfactually-invariant predictor will satisfy an independence criterion that can be derived from the causal structure of the data generating process (Veitch et al., 2021). Returning to Example 2, the desideratum is that the predicted diagnosis $f(X)$ should not be affected by the aspects of the writing style that are associated with the hospital Z. This 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, such as content moderation, the label is an effect of the text, rather than a cause — for a detailed discussion of this distinction, see Jin et al. (2021). In such cases, 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 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 domains, known as *domain generalization* (Ghifary et al., 2015; Gulrajani and Lopez-Paz, 2020). However, it may be necessary to observe data from a very large number of environments to disentangle the true causal structure (Rosenfeld et al., 2021).

Both general approaches require richer training data than in typical supervised learning: Either explicit labels Z for the factors to disentangle from the predictions or access to data gathered from multiple labeled environments. Obtaining such data may be rather challenging, even compared to creating counterfactual instances. Furthermore, the distributional approaches have thus far been applied only to classification problems, while data augmentation can easily be applied to structured outputs such as machine translation.

4.2 Fairness and bias

NLP systems inherit and sometimes amplify undesirable biases encoded in text training data (Barocas et al., 2019; Blodgett et al., 2020). Causality

[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).

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) argue that a causal analysis is required to determine the fairness properties of an observed distribution of data and predictions; 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 (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.

Fairness with text. 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 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) can reduce the use of spurious correlations with race for toxicity detection.

4.3 Causal Model Interpretations

Explanations of model predictions can be crucial to help diagnose errors and establish trust with decision makers (Guidotti et al., 2018; Jacovi and Goldberg, 2020). 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, explanation can be performed by comparing predictions for each example and its generated counterfactual. While it is usually not possible to observe counterfactual predictions, here the causal system is the predictor itself. In those cases it may be possible to compute counterfactuals, e.g. by manipulating the activations inside the network (Vig et al., 2020; Geiger et al., 2021). Treatment effects can then be computed by comparing the predictions under the factual and counterfactual conditions. Such a controlled setting is similar to the randomized experiment described in § 2, where it is possible to compute the difference between an actual 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; Gardner et al., 2020; Rosenberg et al., 2021; Ross et al., 2021; Meng et al., 2022; Zhang et al., 2022). However, generating such counterfactuals is challenging (see § 4.1.1).

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). 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) offer a method for removing information from representations by iteratively training linear classifiers and

projecting the representations on their null-spaces, but do not account for confounding concepts.

A complementary approach is to *generate* counterfactuals with minimal changes that obtain a different model prediction (Wachter et al., 2017; Mothilal et al., 2020). Such examples allow us to observe the changes required to change a model's prediction. Causal modeling can facilitate this by making it possible to reason about the causal relationships between observed features, thus identifying minimal actions which might have downstream effects on several features, ultimately resulting in a new prediction (Karimi et al., 2021).

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

4.4 Future work

In general we cannot expect to have full causal models of text, so a critical question for future work is how to safely use *partial* causal models, which omit some causal variables and do not completely specify the causal relationships within the text itself. A particular concern is unobserved confounding between the variables that are explicitly specified in the causal model. 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 those understood by the reader.

Partial causal models of text can be "top down", in the sense of representing causal relationships between the text and high-level document metadata such as authorship, or "bottom up", in the sense of representing local linguistic invariance properties, such as the intuition that a multiword expression like 'San Francisco' has a single cause. The methods described here are almost exclusively based on top-down models, but approaches such as perturbing entity spans (e.g., Longpre et al., 2021) can be justified by implicit bottom-up causal models. Making these connections more explicit may yield new insights. Future work may also explore hybrid models that connect high-level document metadata with medium-scale spans of text such as sentences or paragraphs.

A related issue is when the true variable of interest is unobserved but we do receive some noisy or coarsened proxy variable. For example, we may wish to enforce invariance to dialect but have access only to geographical information, with which dialect is only approximately correlated. This is an emerging area within the statistical literature (Tchetgen et al., 2020), and despite the clear applicability to NLP, we are aware of no relevant prior work.

Finally, applications of causality to NLP have focused primarily on classification, so it is natural to ask how these approaches might be extended to structured output prediction. This is particularly challenging for distributional criteria like $f(X) \perp\!\!\!\perp Z \mid Y$, because $f(X)$ and Y may now represent sequences of vectors or tokens. In such cases it may be preferable to focus on invariance criteria that apply to the loss distribution or calibration.

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 estimating the magnitude of causal effects. and more traditional NLP tasks. These branches of scientific inquiry share common goals, intuitions, and are beginning to show methodological synergies. In § 3 we showed how recent advances in NLP modeling can help researchers make causal conclusions with text data and the challenges of this process. In § 4, we showed how ideas from causal inference can be used to make NLP models more robust, trustworthy and transparent. We also gather approaches that are implicitly causal and explicitly show their relationship to causal inference. Both of these spaces remain nascent with a large number of open challenges which we have detailed throughout this paper.

A particular advantage of causal methodology is that it forces practitioners to explicate their assumptions. To improve scientific standards, we believe that the NLP community should be clearer about these assumptions and analyze their data using causal reasoning. This 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.

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, 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.

Nitay Calderon, Eyal Ben-David, Amir Feder, and Roi Reichart. 2022. Docogen: Domain counterfactual generation for low resource domain adaptation. In *Proceedings of the 60th Annual Meeting of the Association of Computational Linguistics (ACL)*.

Yining Chen, Colin Wei, Ananya Kumar, and Tengyu Ma. 2020. Self-training avoids using spurious features under domain shift. *Advances in Neural Information Processing Systems*, 33:21061–21071.

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.

Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. 2018. Hotflip: White-box adversarial examples for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 31–36.

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.

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.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. 2021. Causal abstractions of neural networks. *Advances in Neural Information Processing Systems*, 34.

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. *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*.

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.

Miguel A Hernán. 2016. Does water kill? a call for less casual causal inferences. *Annals of epidemiology*, 26(10):674–680.

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

Zhiting Hu and Li Erran Li. 2021. A causal lens for controllable text generation. *Advances in Neural Information Processing Systems*, 34.

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*.

Guido W Imbens and Donald B Rubin. 2015. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205.

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*.

Rohan Jha, Charles Lovering, and Ellie Pavlick. 2020. Does data augmentation improve generalization in nlp? *arXiv preprint arXiv:2004.15012*.

Shagun Jhaver, Sucheta Ghoshal, Amy Bruckman, and Eric Gilbert. 2018. Online harassment and content moderation: The case of blocklists. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 25(2):1–33.

Zhijing Jin, Julius von Kügelgen, Jingwei Ni, Tejas Vaidhya, Ayush Kaushal, Mrinmaya Sachan, and Bernhard Schoelkopf. 2021. Causal direction of data collection matters: Implications of causal and anticausal learning for nlp. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9499–9513.

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*.

Chloé Kiddon, Luke Zettlemoyer, and Yejin Choi. 2016. Globally coherent text generation with neural checklist models. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 329–339.

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.

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.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196. PMLR.

Xiaodong Liu, Jianfeng Gao, Xiaodong He, Li Deng, Kevin Duh, and Ye-Yi Wang. 2015. Representation learning using multi-task deep neural networks for semantic classification and information retrieval. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 912–921.

Yang Liu and Mirella Lapata. 2018. Learning structured text representations. *Transactions of the Association for Computational Linguistics*, 6:63–75.

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*.

Lajanugen Logeswaran, Honglak Lee, and Samy Bengio. 2018. Content preserving text generation with attribute controls. *Advances in Neural Information Processing Systems*, 31.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2021. Entity-based knowledge conflicts in question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7052–7063.

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.

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*.

David McClosky, Eugene Charniak, and Mark Johnson. 2006. Effective self-training for parsing. In *Proceedings of the main conference on human language technology conference of the North American Chapter of the Association of Computational Linguistics*, pages 152–159. Citeseer.

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*.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual knowledge in gpt. *arXiv preprint arXiv:2202.05262*.

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.

Tong Niu and Mohit Bansal. 2018. Polite dialogue generation without parallel data. *Transactions of the Association for Computational Linguistics*, 6:373–389.

Yaakov Ophir, Refael Tikochinski, Christa SC Asterhan, Itay Sisso, and Roi Reichart. 2020. Deep neural networks detect suicide risk from textual facebook posts. *Scientific reports*, 10(1):1–10.

Silviu Paun, Bob Carpenter, Jon Chamberlain, Dirk Hovy, Udo Kruschwitz, and Massimo Poesio. 2018. Comparing bayesian models of annotation. *Transactions of the Association for Computational Linguistics*, 6:571–585.

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.

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

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. 2021. Causal effects of linguistic properties. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4095–4109.

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*.

Roi Reichart and Ari Rappoport. 2007. Self-training for enhancement and domain adaptation of statistical parsers trained on small datasets. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 616–623.

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.

Paul R Rosenbaum. 2007. Interference between units in randomized experiments. *Journal of the american statistical association*, 102(477):191–200.

Daniel Rosenberg, Itai Gat, Amir Feder, and Roi Reichart. 2021. Are vqa systems rad? measuring robustness to augmented data with focused interventions. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 61–70.

Elan Rosenfeld, Pradeep Ravikumar, and Andrej Risteski. 2021. The risks of invariant risk minimization. In *International Conference on Learning Representations*, volume 9.

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.

Jennifer D Rubin, Lindsay Blackwell, and Terri D Conley. 2020. Fragile masculinity: Men, gender, and online harassment. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–14.

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.

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.

Anders Søgaard. 2013. Semi-supervised learning and domain adaptation in natural language processing. *Synthesis Lectures on Human Language Technologies*, 6(2):1–103.

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

Mark Steedman, Miles Osborne, Anoop Sarkar, Stephen Clark, Rebecca Hwa, Julia Hockenmaier, Paul Ruhlen, Steven Baker, and Jeremiah Crim. 2003. Bootstrapping statistical parsers from small datasets. In *10th conference of the European chapter of the association for computational linguistics*.

Eric J Tchetgen Tchetgen, Andrew Ying, Yifan Cui, Xu Shi, and Wang Miao. 2020. An introduction to proximal causal learning. *arXiv preprint arXiv:2009.10982*.

Matt Thomas, Bo Pang, and Lillian Lee. 2006. Get out the vote: Determining support or opposition from congressional floor-debate transcripts. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 327–335, Sydney, Australia. Association for Computational Linguistics.

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.

Stefan Wager and Susan Athey. 2018. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242.

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.

Galen Weld, Peter West, Maria Glenski, David Arbour, Ryan Rossi, and Tim Althoff. 2022. Adjusting for confounders with text: Challenges and an empirical evaluation framework for causal inference. *ICWSM*.

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

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

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. 2020. Quantifying the causal effects of conversational tendencies. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):1–24.

Yi-Fan Zhang, Hanlin Zhang, Zachary C Lipton, Li Erran Li, and Eric P Xing. 2022. Can transformers be strong treatment effect estimators? *arXiv preprint arXiv:2202.01336*.

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.