BALANCING THE ASSUMPTIONS OF CAUSAL INFERENCE AND NATURAL LANGUAGE PROCESSING

by

Zach Wood-Doughty

A dissertation submitted to Johns Hopkins University

in conformity with the requirements for the degree of

Doctor of Philosophy

Baltimore, Maryland

August 2021

Abstract

Drawing conclusions about real-world relationships of cause and effect from data collected without randomization requires making assumptions about the true processes that generate the data we observe. Causal inference typically considers low-dimensional data such as categorical or numerical fields in structured medical records. Yet a restriction to such data excludes natural language texts – including social media posts or clinical free-text notes – that can provide a powerful perspective into many aspects of our lives. This thesis explores whether the simplifying assumptions we make in order to model human language and behavior can support the causal conclusions that are necessary to inform decisions in healthcare or public policy. An analysis of millions of documents must rely on automated methods from machine learning and natural language processing, yet trust is essential in many clinical or policy applications. We need to develop causal methods that can reflect the uncertainty of imperfect predictive models to inform robust decision-making.

We explore several areas of research in pursuit of these goals. We propose a measurement error approach for incorporating text classifiers into causal analyses and demonstrate the assumption on which it relies. We introduce a framework for generating synthetic text datasets on which causal inference methods can be evaluated, and use it to demonstrate that many existing approaches make assumptions that are likely violated. We then propose a proxy model methodology that provides explanations for uninterpretable black-box models, and close by incorporating it into our measurement error approach to explore the assumptions necessary for an analysis of gender and toxicity on Twitter.

Readers: Mark Dredze, Ilya Shpitser, and David Broniatowski

Acknowledgements

This thesis is a culmination of five years of work. I owe thanks to many people who supported me throughout my Ph.D. and many more who made that journey possible in the first place. While I cannot hope to thank everyone who has helped me along the way, I will try my best.

I am immensely thankful to my adviser, Mark Dredze. Mark recruited me to Hopkins and has supported me throughout my time here. When happenstance led to my initial interest in causal inference, Mark gave me the freedom to pursue and eventually foreground causality in my research. A testament to Mark's quality is the group of students, postdocs, and collaborators that he has cultivated and mentored throughout the years.

Ilya Shpitser gradually became my unofficial co-adviser throughout my years at Hopkins. His causal inference course in Fall 2016 unequivocally changed the course of my Ph.D.; counterfactually speaking, I have no clue what my thesis would have been had I not "fallen in with the wrong crowd."

I'm thankful to David Broniatowski for serving on my committee and for mentoring me while co-leading the grant[1] that supported me for several years.

Research is a team sport, and I am indebted to my other co-authors: Aaron, Alicia, Amelia, Anju, Becky, Eric, Gideon, Isabel, Michael, Nick, Paiheng, Praateek, Sandra, Silvio, and Xiao. After these acknowledgements, this thesis will use plural

[1] Grant 5R01GM114771 from the National Institute of General Medical Sciences.

pronouns to reflect that this work is not mine alone.

CLSP and Hopkins CS are unique research communities where I have had the chance to learn from many amazing people. I'm particularly thankful for the other advisees of Mark and Ilya, who have provided crucial emotional support on Slack and in person over the years. Razieh and Rohit have been especially helpful while navigating the job market. Thank you to Adarsh, Becky, Hongyuan, Matthew, Pamela, Patrick, and Teodor for camaraderie over many years; to Rebecca and Huda for mentorship and commiseration; to the umami book club; to the Spirit Island support group; to the yellow book reading group; and to the otter and eagle fans in the 2nd-floor Hackerman lab. I was lucky enough to have a chance to learn from and mentor fellow students while at Hopkins; thank you to Praateek, Vibhu, Xiao, Chris, and Lili. A special thanks to Isabel and Paiheng – I look forward to seeing what you do next! Thank you as well to the students, TAs, and CAs who made our Fall 2020 Machine Learning class successful despite the odds. These opportunities, collaborations, and friendships were made possible by the unrelenting efforts of the CLSP and CS staffs; we are all indebted to Carl, Jaime, Jennifer, Joanne, Kim, Lauren, Steven, Yamese, Zack, and – most of all – Ruth.

Outside of Hopkins, there are many people who have helped me through the past five years. Thanks to the friends and teammates from West and Carleton that have kept in touch, including: the lunch group, the poets, the Penguins, and those from Fitness House and Watson. Anju and Gideon mentored me during and after my internship at Bloomberg LP, and I learned a great deal from Katie and other fellow interns. I must also thank the therapists who have helped me stay healthy through these challenging years.

As I look forward to teaching full-time, I want to thank the many teachers who laid the foundation for my educational journey so far. Each taught me more lessons than I can remember. Thank you to Melanie and each teacher at Tenney. From

Franklin, Randall and Hamilton: thank you to Connie, Christie, Barb, Sheryl, Ms. Spiegel, Ms. Mincberg, Ms. Bauch, Ms. Dickman, and Ms. Bauer. From West, thank you to Ms. Keys, Ms. Hager, Ms. Crouch, Sra. Schilz, Mr. Howe, and Mr. D. From Carleton, I am indebted to Jessica Leiman, Jeff Ondich, Sherri Goings, Silvia López, and Eric Egge. A special thanks to Bob Dobrow for introducing me to stochastic processes and Ueshiro Shorin-Ryu Karate, and to David Liben-Nowell who introduced me to both NLP and research, advised our "copy cats" comps group, and gave inspiration for and caution against my own grad school journey.

My most heartfelt thanks goes to my family, who were my first and most influential teachers. I am indebted to my grandparents for their immense passion for education, which my parents have passed on to me. From a very early introduction to handwriting pseudocode `for` loops and a well-timed gift of a beginner's Java textbook, to their insights into navigating life in and outside of academia, my parents have given me every possible opportunity. Their constant support throughout my educational journey has sustained me more than I can express. My brother has also been helpful at times. My final thanks goes to Milana, who has shared her love and support with me every day of the past five years.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Correlation is not causation, yet correlations in observational data are often the only way to measure relationships between variables when randomization is infeasible. The field of causal inference explores under what assumptions an observational dataset can provide evidence for causality. When we make policies based on the belief that smoking causes cancer, we are trusting in decades of research that relies on causal methods. Observational data is much more plentiful than data from randomized studies, so causal inference has the potential to produce insights that would otherwise be expensive, slow, or impossible.

What clinical insights into long-term care can we uncover by studying millions of physician's notes in aggregate? Can we study harassment and toxic language on social media to create new moderation policies that enable better online communities? Causal inference research, however, typically only considers structured, low-dimensional data. In many datasets, whether from Twitter or medical records, unstructured texts contains an enormous amount of information that cannot be trivially converted into structured variables. The vast quantities of text data we create provides both an incredible opportunity for research and a massive challenge for domain experts. For any rare disease or specific cultural phenomenon, the vast majority of clinical notes or social media posts are irrelevant; yet such texts may be the only way to study such issues.

When analyzing millions of natural language documents, automated methods are necessary. Machine learning (ML) and natural language processing (NLP) classifiers have the ability to predict a wide variety of structured variables from unstructured text data. With the large amount of available data and increasingly sophisticated methods, such classifiers have demonstrated impressive performance on many real-world tasks, exceeding human performance in some (highly-controlled) environments. Yet despite this success, these methods are inherently imperfect. While we may want to use ML or NLP to infer variables that are otherwise unobserved in our data, trust is essential in many clinical or policy applications. We need causal methods that reflect the uncertainty of imperfect predictive models to inform robust decision-making. This thesis will explore how to combine the assumptions and methods of causal inference and NLP to produce reliable estimates of causal effects from datasets containing text.

1.1 Motivating Examples: Social Media and Clinical Notes

While the methods introduced in this thesis can be applied to many domains, we introduce two representative examples to motivate the work throughout. We will return to these examples throughout to highlight how more abstract details of predictive model performance or causal assumptions might translate into meaningful differences in an application with real-world consequences.

1.1.1 Clinical Notes

What clinical insights into long-term care can we uncover by studying millions of physicians' notes in aggregate? Free text notes in medical records contain information about patients' histories, possible diagnoses, or patient-doctor relationships (Rajkomar et al., 2018; McVeigh et al., 2016). Importantly, such information often does not appear anywhere else in a patient's medical record, and thus is inaccessible to retrospective causal analyses that do not use free text data (Wu et al., 2013; Rosenbloom et al., 2011; Zheng et al., 2011).

Suppose we want to study whether maternal vitamin D deficiency is a risk factor for the pregnancy complication preeclampsia (Bodnar et al., 2014). Individuals of lower socioeconomic status (SES) are both more likely to have a vitamin D deficiency and are at higher baseline risk for preeclampsia (Silva et al., 2008). Because correlation does not imply causation, a valid causal analysis would need to consider SES when estimating the effect of vitamin D deficiency. It is reasonable to imagine that some combination of clinical practice, privacy concerns, or data logistics could result in a dataset of patient records that does not explicitly record SES.

In this hypothetical scenario, privacy laws and practical concerns might make it impossible for human annotators to go through thousands of patient records to annotate for SES. But if stylistic or semantic differences in how physicians write their notes are correlated with patients' SES, then a statistical model could provide reliable predictions to infer this missing variable, enabling a scientific analysis that would otherwise be impossible. Such a hypothetical application motivates the methodological contributions in Chapters 5, 6, and 7.

1.1.2 Social Media

What can we learn about human behavior and culture by studying how people describe themselves online? Can we understand the social factors that enable health misinformation to spread through a social network? Twitter and other platforms host millions of new posts per day, providing both an incredible opportunity for research and a massive challenge for domain experts.

Traditional survey methods provide a central tool for social scientists and

Chapter	Paper(s)	URL
4	Wood-Doughty, Mahajan, and Dredze (2018); Wood-Doughty et al. (2018); and Wood-Doughty et al. (2021)	bitbucket.org/mdredze/demographer
5	Wood-Doughty, Shpitser, and Dredze (2018)	github.com/zachwooddoughty/emnlp2018-causal
6	Wood-Doughty, Shpitser, and Dredze (2020)	github.com/zachwooddoughty/cdml20_sensitivity
7	Wood-Doughty, Shpitser, and Dredze (2021)	github.com/zachwooddoughty/causal_text_dgps
8	Wood-Doughty, Cachola, and Dredze (2021)	github.com/isabelcachola/mimic-proxy

Table 1-I. URLs for code and/or data for each applicable chapter.

policy-makers. Gold-standard methods like random digit dialing, however, are expensive and slow. If health policy experts want to repeat a national survey on a regular basis to understand public opinion around a vaccination campaign, social media may be the only possible option.

While Twitter is large enough that many people do post their opinions on public health policy, the vast majority of posts are irrelevant to such a specific topic. Furthermore, while Twitter users may publicly share some aspects of their identity (e.g. geolocation, first name), they typically do not report their demographics in the same format as expected by traditional social science methods. Automated methods that can predict demographic information from a user's profile or Twitter history provide a necessary approach to link such data back to the gold-standard survey data they are meant to supplement. We discuss such demographic classifiers in Chapter 4, and applications to large-scale studies of Twitter behavior are explored in Chapters 6 and 9.

1.2 Outline & Contributions

The main contributions of this thesis are methods designed to help enable analyses like those of our motivating examples. Such an analysis, if it is to provide a reliable lens on the real-world phenomena it seeks to measure, must stand upon a structure of assumptions that have been introduced and questioned over many years of research. In each chapter, we discuss the assumptions from prior work we build on, a method that can enable new analyses, or a framework for evaluating such methods. Together, they provide a path towards analyses that can robustly address our motivating examples. Table 1-I lists code and data releases that correspond to each chapter.

In Chapter 2, we introduce many of the fundamental assumptions of causal inference and highlight how the questions it consider differ from those of traditional supervised ML.

Chapter 3 switches focus to the fundamentals of ML and NLP, and discusses the implicit assumptions of predictive methods and how they do or do not comport with causal reasoning.

In Chapter 4, we discuss several papers on predicting the demographics of social media users. While this work is not the main focus of this thesis, it provides relevant background to our analyses in Chapter 6 and 9, as well as some examples of the intersection of assumptions between causal and predictive reasoning. This work draws from several of our publications (Wood-Doughty et al., 2017; Wood-Doughty, Mahajan, and Dredze, 2018; Wood-Doughty et al., 2018; Wood-Doughty et al., 2021).

Chapter 5 introduces a framework for combining text classifier predictions into causal analyses. It derives estimators that incorporate a classifier outputs into a provably unbiased estimate of a causal effect, and evaluates them on simple synthetic datasets. This work was published in Wood-Doughty, Shpitser, and

Dredze (2018).

Chapter 6 builds directly on the previous chapter by addressing concerns about the uncertainty of a causal estimate that relies on a noisy classifier with limited validation data. We introduce a sensitivity analysis that provides empirical bounds around our estimate of the causal effect and use our method to analyze a dataset of Twitter users who post opinions about vaccinations. This work was published in Wood-Doughty, Shpitser, and Dredze (2020).

In Chapter 7, we establish a general framework for evaluating methods for causal inference that use text data. To improve upon the simple toy datasets of our previous work, we use recent work in natural language generation to produce datasets in which we can carefully control the causal effects between low-dimensional variables and text. We use this framework to evaluate our approach from Chapter 5 and show that it is more robust than three other methods proposed in recent work. This work is available as a preprint at Wood-Doughty, Shpitser, and Dredze (2021).

Chapter 8 departs from a focus on causal inference and introduces a methodology for interpretability of black-box models. We demonstrate that our method can fulfill many of the desiderata of ML model explanations, and apply our approach to two distinct tasks. Half of this work is available as a preprint at Wood-Doughty, Cachola, and Dredze (2021).

Chapter 9 seeks to tie together several threads from throughout the thesis. We use demographic classifiers introduced in Chapter 4 and our proxy model methodology from Chapter 8 to explore the joint distribution of gender and toxic language in a new dataset of millions of tweets. We apply our measurement error framework from Chapter 5 and sensitivity analyses from Chapter 6 to formalize and study assumptions in our approach and highlight the uncertainty in our results.

Chapter 10 recaps our contributions and discusses future work.

Chapter 2

Assumptions in Causal Inference

2.1 Introduction

Causal understanding is the ultimate goal of much of scientific inquiry, as researchers seek to understand whether a clinical treatment improves patients' outcomes or whether a specific economic policy was responsible for an observed change in behavior. The field of causal inference studies whether and how it is possible to make causal claims (e.g. "smoking causes cancer") from a specific dataset. While the roots of such reasoning dates back hundreds of years, causal methods have become increasingly widespread in recent years following the greater availability of data (Grimmer, 2015) and the adoption of new statistical methods (Pearl, 2009; Morgan and Winship, 2015).

The primary goal of causal inference is to understand when observational (non-randomized) datasets can be used to estimate the causal relationships between real-world variables. Randomized controlled trials (RCTs), in which a treatment is assigned completely at random, are the gold standard of determining causal effects of clinical treatments on outcomes because they ensure that no variable can cause both a patient's outcome and the patient's assigned treatment. However, RCTs can be expensive or impossible in many settings and thus most datasets are collected without randomization. Many studies simply report correlations from observational data; causal inference examines what assumptions and analyses make it possible to identify causal effects.

We formalize a causal statement like “smoking causes cancer” as “if we were to conduct a RCT and assign smoking as a treatment, we would see a higher incidence of cancer among those assigned smoking than among the control group.” Building on foundational frameworks of Neyman (1923), Rubin (1976) and Pearl (1995), we can write our causal question as a counterfactual: what *would have* been the cancer incidence among smokers if smoking *had been* randomized? Specifically, we consider a causal effect as the counterfactual outcome of a hypothetical intervention on some treatment variable. If we denote smoking as our treatment variable A and cancer as our outcome variable Y, then we are interested in functions of the counterfactual distribution. We will interpret a counterfactual distribution $p(Y(a))$ as “the distribution over Y had A been set, possibly contrary to fact, to value a.” For a binary treatment A, the causal effect of A on Y is denoted $E[Y(a=1)] - E[Y(a=0)]$; the average difference between if you had received the treatment and if you had not.

Causal directed acyclic graphs (DAGs) provide a helpful way to reason about the assumptions underlying many causal analyses, where nodes represent variables and edges represent direct causal effects between variables. Figure 2-1 shows an example of simple confounding. This is the simplest DAG in which counterfactual distribution $p(Y(a))$ is not simply $p\left(Y|A\right)$, as C influences both the treatment A and the outcome Y. To recover the counterfactual distribution $p(Y(a))$ that would follow an intervention upon A, we “adjust” for C, applying the so-called “back-door criterion” (Pearl, 1995).

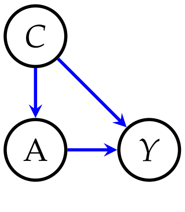


Figure 2-1. A simple causal DAG.

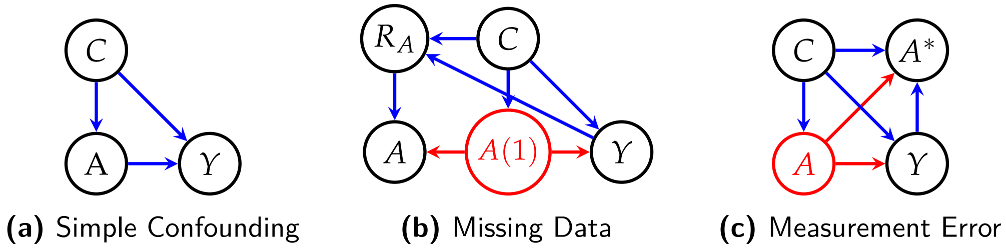
	Surgery 1	Surgery 2
Young	93%	89%
Old	71%	68%
Total	78%	83%

Table 2-I. Simpson's Paradox.

$$\begin{aligned} p(Y(a)) &= \sum_C p(Y(a) \mid C)p(C) && (2.1)\\ &= \sum_C p(Y(a) \mid A, C)p(C) && (2.2)\\ &= \sum_C p(Y \mid A, C)p(C) && (2.3) \end{aligned}$$

This derivation relies primarily on two assumptions. (2.2) relies on the conditional independence $Y(a) \perp A \mid C$ that is encoded in the DAG. In causal inference, an independence between a counterfactual $Y(a)$ and the treatment A is referred to as (conditional) ignorability. If there were confounders other than C, we would also need to condition on them; if there were unobserved confounders, we would need additional assumptions. (2.3) follows from consistency; in our smoking example, this can be described as the assumption that the act of smoking affects your probability of cancer in the same way regardless of whether you chose to smoke or were assigned to, or whether you smoke in the afternoon or evening (VanderWeele, 2009; Rehkopf, Glymour, and Osypuk, 2016). Consistency might be violated if *how* smoking causes cancer depends on *how much* you smoke, but our analysis erroneously assumes that smoking can be treated as a binary variable. Additionally, given the cultural context, we might expect that a real-world experiment that *forces* you to smoke cigarettes would be traumatic and otherwise influence your health outcomes.

When a given set of assumptions suffices to connect the desired counterfactual distribution to the observed data distribution, we say that our causal question is



(a) Simple Confounding **(b)** Missing Data **(c)** Measurement Error

Figure 2-2. DAGs for causal inference. Red variables are unobserved. A is a treatment, Y is an outcome, and C is a confounder.

identified. Identification is a primary concern of causal inference and the subject of a wide literature (Shpitser and Pearl, 2008). Simpson's paradox, shown in Table 2-I, highlights the challenge of confounding bias (Simpson, 1951; Blyth, 1972). In this example, our treatment A is one of two surgeries, age is a confounder C, and the cells of the table show the recovery probability: $p\,(Y|A,C)$. If we compare surgeries within age groups, Surgery 1 looks superior; but if we aggregate across all patients, Surgery 2 looks better. To correctly conclude that Surgery 1 is best, we need to know that a patient's age may influence surgery, but surgery cannot change one's age. This critical distinction is represented in the DAG in Figure 2-1 by the edge from C to A. For a real-world problem, if we don't know the causal structure, we may not be able to learn anything from causal methods. Similarly, if any of the variables that are relevant to our analysis are not recorded in our dataset, our desired counterfactual may be unidentified and our causal question may be unanswerable.

2.2 Measurement Error and Missing Data

Real-world observational data is messy and often imperfectly collected. While unobserved variables can render causal questions unanswerable, there are many approaches that can recover from data recorded with missing values or systematic mismeasurement.

A	C	Y
1	1	0
0	1	1
0	0	1
1	0	1

(a) Simple Confounding

R_A	A	C	Y
1	1	1	0
0	?	1	1
1	0	0	1
0	?	0	1

(b) Missing Data

A^*	C	Y
0	1	0
0	1	1
0	0	1
1	0	1

(c) Measurement Error

A^*	A
1	1
0	1
0	0
1	1

(d) Mismeasurement

Figure 2-3. Example data rows for causal inference without text data.

2.2.1 Missing Data

Our dataset has "missing data" if it contains individuals (instances) for which some variables are sometimes, but not always, missing from the dataset. This may occur if some survey respondents choose not to answer certain questions, or if certain variables are difficult to collect and thus infrequently recorded. Missing data is closely related to causal inference – both are interested in hypothetical distributions that we cannot directly observe (Robins, Rotnitzky, and Scharfstein, 2000; Shpitser, Mohan, and Pearl, 2015).

Consider a causal model where the treatment A is sometimes missing (Figure 2-2b). The variable R_A is a binary indicator for whether A is observed ($R_A = 1$) or missing. The variable $A(R_A = 1)$, written as $A(1)$, represents the counterfactual value of A were it never missing. Finally, A is the observed proxy for $A(1)$: it has the same value as $A(1)$ if $R_A = 1$, and the special value "?" if $R_A = 0$.

Solving missingness can seen as intervening to set R_A to 1. Given $p(A, R_A, C, Y)$, we want to recover $p(A(1), C, Y)$. We may need to make a "Missing at Random" (MAR) assumption, which says that the missingness process is independent of the true missing values, conditional on observed values. Figure 2-2b reflects the MAR assumption; R_A is independent of $A(1)$ given fully-observed C and Y. If an edge existed from $A(1)$ to R_A, we have "Missing Not at Random" (MNAR) and would not be identified except in special cases (Shpitser, Mohan, and Pearl, 2015).

To derive the causal effect despite missing data, we can first identify the causal effect in terms of the true $A(1)$ using the same derivation as above.

$$\begin{aligned} p(Y(A(1)=a)) &= \sum_C p(Y(A(1)=a) \mid C)p(C) && (2.4) \\ &= \sum_C p(Y(A(1)=a) \mid A(1), C)p(C) && (2.5) \\ &= \sum_C p(Y \mid A(1), C)p(C) && (2.6) \end{aligned}$$

Where 2.4 holds by chain rule, 2.5 holds by $A(1) \perp Y(a) \mid C$, and 2.6 by consistency. Now, we identify $A(1)$ in terms of observed data.

$$\begin{aligned} & p(A(1), C, Y) \\ &\quad = p(A(1) \mid C, Y)p(C, Y) && (2.7) \\ &\quad = p(A(1) \mid C, Y, R_A = 1)p(C, Y) && (2.8) \\ &\quad = p(A \mid C, Y, R_A = 1)p(C, Y) && (2.9) \end{aligned}$$

Where 2.7 holds by chain rule, 2.8 by $A(1) \perp R_A \mid C, Y$, and 2.9 by consistency. Now, use Eq 2.9 to identify $p(Y \mid A(1), C)$ from Eq 2.6 in terms of observed data.

$$\begin{aligned} & p(Y \mid A(1), C) \\ &\quad = \frac{p(Y, A(1), C)}{p(A(1), C)} && (2.10) \\ &\quad = \frac{p(Y, A(1), C)}{\sum_Y p(Y, A(1), C)} && (2.11) \\ &\quad = \frac{p(A \mid C, Y, R_A = 1)p(C, Y)}{\sum_Y p(A \mid C, Y, R_A = 1)p(C, Y)} && (2.12) \\ &\quad = \frac{p(A \mid C, Y, R_A = 1)p(Y \mid C)}{\sum_Y p(A \mid C, Y, R_A = 1)p(Y \mid C)} && (2.13) \end{aligned}$$

$$\tau_S = \sum_C \left(p(Y=1 \mid A=1, C) - p(Y=1 \mid A=0, C)\right) p(C) \tag{2.15}$$

$$\tau_{\text{MD}} = \sum_C \Bigg(\frac{p(A=1 \mid C, Y=1, R_A=1)}{\sum_y' p(A=1 \mid C, y', R_A=1) p(Y=y' \mid C)} - \frac{p(A=0 \mid C, Y=1, R_A=1)}{\sum_y' p(A=0 \mid C, Y=y', R_A=1) p(Y=y' \mid C)} \Bigg) p(Y=1, C) \tag{2.16}$$

$$\tau_{\text{ME}} = \sum_C \left(\frac{\frac{-\delta_{c,y=1} q_{c,y=1}(0) + (1-\delta_{c,y=1}) q_{c,y=1}(1)}{(1-\epsilon_{c,y=1}-\delta_{c,y=1})}}{\sum_{y'} \frac{-\delta_{c,y'} q_{c,y'}(0) + (1-\delta_{c,y'}) q_{c,y'}(1)}{(1-\epsilon_{c,y'}-\delta_{c,y'})}} - \frac{\frac{(1-\epsilon_{c,y=1}) q_{c,y=1}(0) - \epsilon_{c,y=1} q_{c,y=1}(1)}{(1-\epsilon_{c,y=1}-\delta_{c,y=1})}}{\sum_{y'} \frac{(1-\epsilon_{c,y'}) q_{c,y'}(0) - \epsilon_{c,y'} q_{c,y'}(1)}{(1-\epsilon_{c,y'}-\delta_{c,y'})}} \right) p(C) \tag{2.17}$$

Define $\epsilon_{c,y} = p(A=0 \mid A^*=1, C=c, Y=y)$, $\delta c, y = p(A=1 \mid A^*=0, C=c, Y=y)$,
$q_{c,y}(0) = p(C=c, Y=y, A^*=0)$, and $q_{c,y}(1) = p(C=c, Y=y, A^*=1)$.

Figure 2-4. Functionals for the causal effects for simple confounding (τ_{SC}), Missing Data (τ_{MD}) and Measurement Error (τ_{ME}).

Where 2.10 holds by definition, 2.11 holds by marginalization, 2.12 holds by an application of 2.9 twice, and 2.13 holds by canceling out p(C). Finally, combine Eq 2.6 and Eq 2.13 to get:

$$p(Y(A(1)=a)) = \sum_C \frac{p(A \mid C, Y, R_A=1) p(Y \mid C)}{\sum_Y p(A \mid C, Y, R_A=1) p(Y \mid C)} p(C) \tag{2.14}$$

Plugging this distribution into $\tau_{\text{MD}} = E[Y(1)] - E[Y(0)]$ gives us the causal effect presented in Figure 2-4, Eq 2.16.

2.2.2 Measurement Error

Sometimes a necessary variable is never observed, but is instead proxied by a variable which differs from the truth by some error. A canonical example of measurement error is the use of body mass index (BMI) as a proxy for obesity in a clinical study (Michels, Greenland, and Rosner, 1998). Obesity can be risk factor for many health outcomes, but has a complex clinical definition and is nontrivial to measure. BMI is a simple deterministic function of height and weight. To conduct a causal analysis of obesity on cancer when only BMI and cancer are measured, we can proceed as if we had measured obesity and then correct our analysis for the known error that comes from using BMI as a proxy for obesity (Hernán and Cole, 2009). Questions of measurement error have been widely studied in diverse fields, from epidemiology (Willett, 1989; Cessie et al., 2012) to statistics (Stefanski and Carroll, 1985; Wang and Wang, 2015) to management science (Eliashberg and Hauser, 1985).

To generalize this concept, we can replace obesity with our ground truth variable A and replace BMI with a noisy proxy A^*. Figure 2-2c gives the DAG for this model. Unlike missing data problems, there is no hypothetical intervention which recovers the true data distribution $p(A, C, Y)$. Instead, we manipulate the observed distribution $p(A^*, C, Y)$ with the known relationship $p(A^*, A)$ to recover the desired $p(A, C, Y)$.

Unlike missing data, measurement error conceptualization can be used even when we never observe A (e.g. the table in Figure 2-3c) as long as we have knowledge about the error mechanism $p(A^*, A)$. Using this knowledge, we can correct for the error using 'matrix adjustment' (Pearl, 2010). In practice we might learn $p(A^*, A)$ from data such as that found in Figure 2-3d.

To derive the estimand for the causal effect of A on Y in Figure 2-2c, we first define the following terms for convenience:

$$\epsilon_{c,y} = p(A = 0 \mid A^* = 1, C = c, Y = y) \tag{2.18}$$
$$\delta c, y = p(A = 1 \mid A^* = 0, C = c, Y = y) \tag{2.19}$$
$$q_{c,y}(0) = p(C = c, Y = y, A^* = 0) \tag{2.20}$$
$$q_{c,y}(1) = p(C = c, Y = y, A^* = 1) \tag{2.21}$$

Eq (5) and (7) from Pearl (2010) gives us:

$$\begin{aligned} & p(A=1, C=c, Y=y) \\ & \quad = \frac{-\delta_{c,y} q_{c,y}(0) + (1-\delta_{c,y}) q_{c,y}(1)}{(1-\epsilon_{c,y}-\delta_{c,y})} \end{aligned} \tag{2.22}$$

$$\begin{aligned} & p(A=0, C=c, Y=y) \\ & \quad = \frac{(1-\epsilon_{c,y}) q_{c,y}(0) - \epsilon_{c,y} q_{c,y}(1)}{(1-\epsilon_{c,y}-\delta_{c,y})} \end{aligned} \tag{2.23}$$

Now,

$$\begin{aligned} & p(Y \mid A=1, C) \\ & \quad = \frac{p(Y, A=1, C)}{p(A=1, C)} && (2.24) \\ & \quad = \frac{p(Y, A=1, C)}{\sum_Y p(Y, A=1, C)} && (2.25) \\ & \quad = \frac{\frac{-\delta_{c,y} q_{c,y}(0) + (1-\delta_{c,y}) q_{c,y}(1)}{(1-\epsilon_{c,y}-\delta_{c,y})}}{\sum_{y'} \frac{-\delta_{c,y'} q_{c,y'}(0) + (1-\delta_{c,y'}) q_{c,y'}(1)}{(1-\epsilon_{c,y'}-\delta_{c,y'})}} && (2.26) \end{aligned}$$

and then,

$$\begin{aligned} & p(Y \mid A=0, C) \\ & \quad = \frac{p(Y, A=0, C)}{p(A=0, C)} && (2.27) \\ & \quad = \frac{p(Y, A=0, C)}{\sum_Y p(Y, A=0, C)} && (2.28) \\ & \quad = \frac{\frac{(1-\epsilon_{c,y}) q_{c,y}(0) - \epsilon_{c,y} q_{c,y}(1)}{(1-\epsilon_{c,y}-\delta_{c,y})}}{\sum_{y'} \frac{(1-\epsilon_{c,y'}) q_{c,y'}(0) - \epsilon_{c,y'} q_{c,y'}(1)}{(1-\epsilon_{c,y'}-\delta_{c,y'})}} && (2.29) \end{aligned}$$

Plugging this distribution into $\tau_{\mathrm{ME}} = E[Y(1)] - E[Y(0)]$ gives us the causal effect presented in Figure 2-4, Eq 2.17.

The measurement error approach we introduce here will be used throughout the thesis, particularly in Chapters 5, 6, and 9. We note that this is not the only way to handle measurement error in causal analyses. A growing line of work has explored the conditions under which identification is possible when we do not have any data on $p(A^*, A)$ but multiple independent proxies for the unobserved variables are available (Kuroki and Pearl, 2014; Miao, Geng, and Tchetgen, 2018; Shi et al., 2020). These methods, and the method we consider, both allow us to point-identify our causal parameter of interest. A separate approach, following a large body of work on bounding causal parameters in measurement error settings and in graphical models more generally (Manski, 1990; Balke and Pearl, 1997; Drton, Sturmfels, and Sullivant, 2009; Evans, 2016). In particular, recent work has introduced a method for bounding a causal parameter in the presence of measurement error, with different combinations of assumptions (Finkelstein et al., 2020). While we will not use these partial identification methods in this thesis, we will discuss them as an alternative to the measurement error sensitivity analyses we introduce in Chapter 6.

2.3 Synthetic Data

Evaluation of causal estimation methods is made much harder by the fact that we never observe counterfactuals – for each patient in a hospital, we can only observe one of many hypothetical worlds in which they receive different treatments. RCTs provide an unbiased *estimate* of the true causal effect, but are still subject to inherent randomness. If we use a method to estimate a causal effect from observational data and find that our estimate differs from that of a RCT, we cannot necessarily tell whether that difference is due to a flawed method, random chance, or some other reason (Hannan, 2008).

Because of such difficulties, causal inference methods are very often evaluated on synthetic datasets for which the causal effect is known by design. Researchers need complete knowledge of a DGP to test the assumptions of a causal method, but such knowledge is often impossible for real-world datasets. Thus while synthetic data has its limitations (Jensen et al., 2019; Gentzel, Garant, and Jensen, 2019), it plays a crucial role in understanding how a causal method performs when its assumptions are met or violated. Recently, causal inference evaluations have tested proposed methods against held-out synthetic DGPs (Hahn, Dorie, and Murray, 2019; Dorie et al., 2019; Shimoni et al., 2018). These synthetic datasets are designed to test different empirical properties of the methods, such as the coverage of confidence intervals or the finite-sample behavior variance of an estimator.

This reliance on synthetic data draws a distinction from most predictive methods in ML and NLP research, where enormous quantities of text and image data have been curated to produce widely-used datasets (Deng et al., 2009; Brown et al., 2020). On such datasets, held-out test error is a sufficient metric to compare one predictive model against another. However, synthetic datasets have been used to explore how ML or NLP models handle edge cases or low-resource settings (Elman, 1990; Patki, Wedge, and Veeramachaneni, 2016; Khayrallah and Koehn, 2018; Wang and Eisner, 2018; Kim and O'Neill-Brown, 2019; Winata et al., 2019). This is especially true in domains where data is not as widely available, such as clinical settings (Boag, Naumann, and Szolovits, 2016; Belinkov and Bisk, 2018; Melamud and Shivade, 2019).

In Chapters 5 and 6, we will use simple synthetic data distributions to highlight the performance of different causal methods. In Chapter 7 we introduce a framework for synthetic data that incorporates text generation and allows for more robust analyses of causal methods that incorporate text.

2.4 Connections to the Campbell and Stanley Framework

While causal DAGs provide a useful approach for reasoning about the assumptions necessary for causal inference, there are other approaches. In particular, the work of Campbell and Stanley (1963) has provided a foundation for many decades of social science research. This line of research, which has been expanded upon by the original authors and the broader scientific community, categorizes *threats* to the causal validity of an empirical study. For example, Cook, Campbell, and Shadish (2002) defines 'Ambiguous Temporal Precedence' as "Lack of clarity about which variable occurred first may yield confusion about which variable is the cause and which is the effect." In total, there are 37 established threats within the typology, grouped as issues of internal validity, external validity, construct validity, or statistical conclusion validity.

Matthay and Glymour (2020) provides an excellent bridge between this typology and the language of DAGs by providing graphical representations of these threats to validity. For example, the threat of Ambiguous Temporal Precedence can be thought of in graphical terms as not knowing the direction of an edge. This question of temporal precedence is particularly relevant to the subfield of causal structure learning, which considers algorithms for learning the causal structure of a DGP from observational data (Drton and Maathuis, 2017). The widely-used PC algorithm learns an *equivalence class* of DAGs, indicating which edges cannot be directed – or, in terms of validity, where temporal precedence remains ambiguous (Kalisch and Bühlman, 2007). In many theoretical contributions to causal inference research (and in this thesis), the methods assume the true causal DAG has been given to the researcher by a domain expert. Whereas much causal research using DAGs focuses on provable guarantees for the identification of a given causal parameter under an exhaustive list of assumptions, the Campbell and Stanley framework provides general principles that can better incorporate domain knowledge about the underlying DGP.

Of particular relevance to the work in this thesis are the threats to internal validity and to construct validity. Following Cook, Campbell, and Shadish (2002), internal validity asks whether an observed correlation reflects a causal relationship between the measured variables. From the previous example of temporal precedence, a study would lack internal validity if it incorrectly assumed the direction of a causal effect between two variables. Construct validity asks whether inferences about the measured variables can be extended to draw conclusions about higher-order concepts. For example, a measured clinical outcome for a cancer drug might be 'all-cause mortality after one year,' but the higher-order concept might be cancer progression itself (Cook, Campbell, and Shadish, 2002).

The graphical causal framework can provide provable guarantees of internal validity, but only when a (long) list of assumptions are met. For example, a practical analysis using must assume that all relevant variables (measured or not) can be written down in the graph, before methods such as the `ID` algorithm can be applied (Shpitser and Pearl, 2008). Writing down such a graph requires substantial domain knowledge, and incorrect assumptions regarding the relevant variables, conditional independences, or temporal precedences can render subsequent analyses invalid.

Graphical causal analyses often ignore questions of construct validity, assuming that the variable in the DAG refers precisely to the construct of interest. A notable exception to this is when considering measurement error. The approaches we discuss in §2.2.2 explicitly assume that the measured proxy (e.g. A^* in Figure 2-2c) is not equivalent to the underlying variable (e.g. A). This approach relies on additional assumptions to achieve internal validity, such as knowledge of the error rate $p\left(U^*|U\right)$ (Kuroki and Pearl, 2014) or multiple proxies Shi et al., 2020. This approach does not solve questions of construct validity more generally; in the example of Figure 2-2c, even if we handle the measurement error between A and A^*, we need domain knowledge to justify that A itself corresponds to a higher-order construct of interest.

Questions of construct validity are particularly challenging for our motivating examples in §1.1. In clinical settings, the underlying physiological status of a patient is extremely complex, and doctors may have access to only a limited number of test results or subjective assessments. Determining whether these tests and assessments have construct validity is extremely difficult but crucial to improving clinical care (Eyres et al., 2005; Maerlender et al., 2010). On social media, questions of validity are made difficult by the noisy nature of the data and the relative novelty of the domain (Broniatowski and Tucker, 2017; Olteanu et al., 2019). For example, assessing construct validity may require grappling with questions of agreement among expert or crowdworker annotations (Chancellor and De Choudhury, 2020; Smith, 2020). In an analysis of 75 papers predicting aspects of mental health from social media, Chancellor and De Choudhury (2020) found that when papers purported to study concepts such as anxiety or depression, there was little discussion of how the constructs that could be measured from social media posts corresponded to these complex and subjective higher-order concepts. Broniatowski and Tucker (2017) connects the successes and failures of ML methods trained on social media data to these underlying questions of validity.

While the methodology of this thesis does not draw primarily from the typology of validity proposed by Cook, Campbell, and Shadish (2002), we will discuss trade-offs between construct validity and internal validity in the context of the assumptions made in our analyses. In particular, we will connect our measurement error framework to questions of construct validity in Chapters 5 and 9.

Chapter 3

Assumptions in ML and NLP

3.1 Introduction

As our motivating examples from §1.1 involve combining predictive models into causal analyses, we now discuss how the causal assumptions from the previous chapter differ from and comport with the traditional assumptions of ML and NLP. These distinctions are important to the methods we propose throughout the thesis; while prior work on predictive modeling has rarely considered how those predictions could be used in a causal analysis, past work can still inform our methods. At the same time, it may be helpful to delineate the working assumptions that are used to demonstrate the efficacy of a predictive method, especially when those assumptions do not suffice for an evaluation of a causal method.

To elucidate the difference in assumptions between causal and predictive models, we first introduce the (cautionary) example of Google Flu Trends (GFT), a model designed to predict the incidence of influenza cases from counts of internet searches aggregated over millions of users in the United States. While initially heralded as a showcase of the potential for big data to transform public health and related fields, it began to make wildly inaccurate predictions, demonstrating that even with tremendous resources, ML methods can fail when not carefully designed for their target domain (Cook et al., 2011; Butler, 2013; Lazer et al., 2014). Rather than focus on the specifics of this model, we will describe it in general terms and leave the details to the original paper (Ginsberg et al., 2009).

We can denote the GFT dataset as $\{X_i, Y_i\}_{i=1}^N$, where X_i is a vector of counts for 50 million search terms and Y_i is the number of hospitalizations due to influenza-like illness. Traditional supervised learning assumes that $\{X_i, Y_i\}_{i=1}^N$ is sampled from a distribution $\mathcal{D}_{\mathcal{X},\mathcal{Y}}$ and the goal of "learning" is to find a function $f : X \to Y$ such that for some loss function L – such as mean squared error for regression – we minimize the expected distribution-level loss $E_{\mathcal{D}}[L(f(X), Y)]$ (Shalev-Shwartz and Ben-David, 2014). By randomly holding-out some of the dataset as a test set, we can evaluate loss on the test set, which provides an unbiased estimate of the distribution-level loss. For this approach to be effective, it requires assumptions. For example, we must assume the training and test set are in fact representative of the larger distribution, and we must make assumptions about the complexity of the function class of f if we want any theoretical guarantees of learning an effective predictor (Mohri, Rostamizadeh, and Talwalkar, 2018).

The traditional supervised learning assumptions do not provide for any causal interpretation of a model's predictions. In the case of GFT, it is clear that we cannot prevent hospitalizations or deaths due to influenza by preventing people from searching for certain queries; an individual's search queries do not *cause* them to become sick. Nor can we necessarily interpret the correlations relevant to a predictive model as relevant to any real-world phenomena. In the case of GFT, its creators initially found that several search terms related to basketball were highly predictive of flu rates, simply because both flu and basketball co-occur in the winter (Lazer et al., 2014). In fact, as we explore in Chapter 8, we cannot always reliably use the parameters of a trained model even to understand that model's own predictive behavior! Many of these challenges come back to what assumptions are made and whether those assumptions are met.

3.2 ML Assumptions: The Example of Naive Bayes

Consider Naive Bayes, a simple supervised ML method used widely in NLP applications (Kim et al., 2006). Naive Bayes makes the assumption that each feature is conditionally independent of the others given the label. We can write this assumption as $X_j \perp \{X_k\}_{j \neq k} \mid Y$. This assumption is almost certainly violated in any real-world dataset, and yet Naive Bayes is commonly used because this assumption simplifies the training algorithm to a process of counting (Raschka, 2014). In particular, if we have m feature dimensions, the Naive Bayes classifier looks like:

$$p(Y = y \mid X = x) = \prod_{j=1}^{m} p(Y = y \mid X_j = x_j) \tag{3.1}$$

$$\propto p(Y = y) \prod_{j=1}^{m} p(X_j = x_j \mid Y = y) \tag{3.2}$$

Learning $p(Y)$ is just a matter of counting how often the label y appears in the data, and $p(X_j = x_j \mid Y = y)$ is just the conditional probability of 'among the examples labeled as y, in how many did feature X_j have value x_j?' The Naive Bayes independence assumption lets us multiply these two-variable conditional probabilities rather than modeling a 2^m-dimensional joint probability distribution and the relationships between features.

Unlike in the case of causal inference, where if we make an incorrect assumption we may get arbitrarily bad results and have no way to check, in supervised learning we can use our held-out test data to evaluate whether our prediction function $f = p(Y \mid X)$ is "good enough." Suppose we have a test dataset with T examples, we can approximate the distribution-level loss:

$$E_{(X,Y)\sim\mathcal{D}}[L(f(X), Y)] \approx \sum_{i=1}^{T} L(f(X_i), Y_i) \tag{3.3}$$

Under the assumptions of supervised learning, if we achieve low error on a held-out test dataset, we have theoretical justification to expect our prediction function f

will perform well on new data. Unlike in causal inference, we may be okay with a simple model that make erroneous assumptions about the true data-generating process as long as that model has good empirical performance. Because we can never construct a test set of counterfactual data (unless we use synthetic data; see §2.3), such an approach does not work in causal settings.

3.3 Violations of ML Assumptions

While the actual reasons for why GFT struggled when deployed for real-world predictions are complicated and still a matter of discussion (Kandula and Shaman, 2019), it performed much better on the training and test data it was initially built with than it did with new data from the underlying distribution of influenza and search data. A common but theoretically uninteresting reason for this occurrence in the ML literature is simply poor experimental design; if a researcher evaluates too many models on the test set, they may unintentionally tune their model to the test set and lose any guarantees about behavior on future data from the distribution (Lipton and Steinhardt, 2019). Despite the importance of these issues, they are less specific to the concerns of this thesis. Instead, we will mention how violations of implicit assumptions in supervised learning may complicate our goals of combining predictive models into causal analyses.

3.3.1 Confounding and Selection Bias

While Chapter 2 discussed how confounding can introduce bias into an estimate of a causal effect, confounding can also complicate predictive modeling. In the case of Naive Bayes, our model simply counts occurrences of a feature conditional on the label, $p(X_j \mid Y)$. A simple form of confounding could be a setting where another variable C acts as a confounder so that the feature X_j strongly predicts label $Y = 1$ when $C = 1$, but predicts label $Y = 0$ when $C = 0$. If C is not available as a feature and is only evaluated on a dataset where $C = 1$, it may achieve low test set error but fail badly on new data where $C = 0$. This issue has been studied in text classification settings by Landeiro and Culotta (2016) and subsequent work, considering examples such as how gender may confound predictions of Twitter user's locations from text. Similar work in this area has fallen under the header of *dataset shift* (Quinonero-Candela et al., 2009), with proposed solutions such as transfer learning (Zhuang et al., 2020) or invariant feature learning (Muandet, Balduzzi, and Schölkopf, 2013). More recent approaches to dataset shift have incorporated causal assumptions to make more robust predictions (Meinshausen, 2018; Subbaswamy, Schulam, and Saria, 2019).

Another way in which our training dataset might differ from future data from our target distribution is the presence of selection bias. In many ML and NLP applications, this is framed in terms of *domain adaptation*, where a model is trained on (widely available) data from one domain, but the goal is to have good performance on (expensive) data from a second domain (Blitzer, Dredze, and Pereira, 2007; Glorot, Bordes, and Bengio, 2011). If it is difficult or impossible to collect more data in the expensive domain, it may instead be possible to model the differences in data distribution between the two domains (Jiang and Zhai, 2007; Khayrallah et al., 2018). More generally, if our training data is drawn from a distribution $p(X, Y \mid S = 1)$ but we want to apply it to a test dataset drawn from $p(X, Y \mid S = 2)$, it may be necessary to model the *selection mechanism* S (Bareinboim and Pearl, 2012). While domain adaptation is rarely explicitly framed as a problem of selection bias, the causal approaches to selection bias may be an effective tool in developing better predictive methods (Bareinboim and Tian, 2015).

3.4 Combining Assumptions

Both selection and confounding biases are major challenges to training predictive models that perform well in a diverse settings. The assumptions necessary to correct for such biases are essential to understanding how and whether a trained model can be reasonably used in downstream applications. If the classifier's predictions have human consequences, it becomes essential to understand the classifier's biases or systematic errors.

Some recent research has investigated the possible real-world consequences of deployed NLP models. A widely-known example is that of gender bias in word embeddings (Bolukbasi et al., 2016). Such biases can occur because embeddings are learned from real-world text datasets that reflect underlying data-generating processes (Garg et al., 2018). There have been a wide variety of methods proposed for 'de-biasing' these learned representations (Bordia and Bowman, 2019), some of which have focused on causal methods (Vig et al., 2020). While this thesis does not directly engage with the growing literature on fairness in ML and NLP (Selbst et al., 2019), such work also studies how the assumptions of a classifier may be violated by a real-world application and is often explicit in its use of causal inference methods (Nabi and Shpitser, 2018)

In Chapters 5, 6, 7, and 9 we will explore how we may incorporate imperfect ML classifiers into robust causal analyses by understanding which assumptions we need to focus on.

Chapter 4

Predicting Social Media User Demographics

4.1 Introduction

We now take a slight detour into our past work on predictive models for social media user demographics. This provides more concrete examples of the NLP methods developed in the previous chapter and introduces methods that we use in Chapters 6 and 9.

Contextualization of population studies with demographics forms a central analysis method within the social sciences. In domains such as political science or public health, standard demographic panels in telephone surveys enable better analyses of opinions and trends. Demographics such as age, gender, race, and location are often proxies for important socio-cultural groups. As the social sciences increasingly rely on computational analyses of online text data, the unavailability of demographic attributes hinders comparison of these studies to traditional methods (Al Baghal et al., 2020; Amir, Dredze, and Ayers, 2019; Jiang and Vosoughi, 2020).

Computational social science increasingly utilizes methods for the automatic inference of demographic attributes from social media, such as Twitter (Burger et al., 2011; Chen et al., 2015; Ardehaly and Culotta, 2017; Jung et al., 2018; Huang and Paul, 2019). Demographics factor into social media studies across domains such as health, politics, and linguistics (O'Connor et al., 2010; Eisenstein et al., 2014).

Gender, race, and ethnicity are sociocultural categories with competing definitions and measurement approaches (Comstock, Castillo, and Lindsay, 2004; Vargas and Stainback, 2016; Culley, 2006; Andrus et al., 2021). Despite this complexity, understanding race and ethnicity is crucial for public health research (Coldman, Braun, and Gallagher, 1988; Dressler, Oths, and Gravlee, 2005; Fiscella and Fremont, 2006; Elliott et al., 2008; Elliott et al., 2009). Analyses that explore mental health on Twitter (Loveys et al., 2018) should consider racial disparities in healthcare (Satcher, 2001; Amir, Dredze, and Ayers, 2019) or online interactions (Delisle et al., 2018; Burnap and Williams, 2016).

We address several challenges within this context, by developing methods that can make predictions with only a single tweet per user, by developing better datasets on Twitter users' race and ethnicity self-descriptions, and applying these methods to widespread analyses of Twitter usage.

4.2 Ethical Considerations

Complexities of predicting facets of users' identities raise ethical considerations, requiring discussion of the benefits and harms of this work (Benton, Coppersmith, and Dredze, 2017). The benefits are clear in settings such as public health; many studies use social media data to research health behaviors or support health-based interventions (Paul and Dredze, 2011; Sinnenberg et al., 2017). These methods have transformed areas of public health which otherwise lack accessible data (Ayers, Althouse, and Dredze, 2014). Aligning social media analyses with traditional data sources requires demographic information.

The concerns and potential harms of these methods are more complex. Ongoing discussions in the literature concern the need for informed consent from social media users (Fiesler and Proferes, 2018; Marwick and boyd, 2011; Olteanu et al., 2019). Twitter's privacy policy states that the company "make[s] public data on Twitter available to the world," but many users may not be aware of the scope or nature of research conducted using their data (Mikal, Hurst, and Conway, 2016). Participant consent must be *informed*, and we should study users' comprehension of terms of service when conducting sensitive research. IRBs have applied established human subjects research regulations in ruling that passive monitoring of social media data falls under public data exemptions.

While data usage agreements can prohibit such behavior, malicious actors may attempt to use predicted user demographics to track or harass minority groups. Despite the severity of such a worst-case scenario, there are two arguments why the benefits may outweigh the harms. First, if open-source methods and models were used for such malicious behavior, platform moderators could simply incorporate those tools into combatting any automated harassment. Second, harassment against historically disenfranchised groups is already extremely widespread. Open-source tools would provide more good than harm in the hands of researchers or platform moderators (Jiang and Vosoughi, 2020). Recent work has show that women on Twitter, especially journalists and politicians, receive disproportionate amounts of abuse (Delisle et al., 2018). On Facebook, advertisers have used the platform's knowledge of users' racial identities to illegally discriminate when posting job or housing ads (Benner, Thrush, and Isaac, 2019; Angwin and Parris Jr, 2016).

Another concern of any predictive model for sensitive traits is that a descriptive model could be interpreted as a prescriptive assessment (Ho, Roberts, and Gelman, 2015; Crawford, 2017). Individual language usage may also differ from population-level demographics patterns (Bamman, Eisenstein, and Schnoebelen, 2014). Predictive models should not be used to profile individuals, and individuals' self-reported demographics should replace predictions whenever possible. Many predictive models of demographics are limited by available training data to a small subset of possible identities. For example, race and ethnicity classifiers are often limited to at most the four most common such categories in the United States. All publicly available gender classifiers make the restrictive assumption to treat gender as binary (Keyes, 2018; Cao and Daumé III, 2020; Keyes, May, and Carrell, 2021).

While our work shares many of these limitations, we strive to differentiate between biased models and biased applications. All predictive models are necessarily imperfect; analyses on which they rely must account for this uncertainty. If a noisy predictive model is used for a small-sample analysis, differences that appear significant may be an artifact of misclassifications. On the whole, we believe demographic tools provide significant benefits that justify the potential risks in their development.

4.3 Related Work

Numerous existing systems automatically infer missing demographics, such as gender, ethnicity, age and location (Mislove et al., 2011; Burger et al., 2011; Culotta, Kumar, and Cutler, 2015; Pennacchiotti and Popescu, 2011; Rao et al., 2010; Jurgens et al., 2015; Dredze et al., 2013; Rout et al., 2013). Most methods rely on content authored by the user, where words or phrases are strongly associated with specific demographic traits (Al Zamal, Liu, and Ruths, 2012). Friendship and follower relationships in social networks can also be informative (Chen et al., 2015; Volkova, Coppersmith, and Van Durme, 2014; Bergsma et al., 2013); people tend to be friends with people who live in the same geographic area (Jurgens, 2013) or tend to follow users with similar political orientations (Conover et al., 2011). Culotta, Kumar, and Cutler (2015) leveraged web traffic data to predict a user's gender and 4-class ethnicity (Caucasian, African-American, Hispanic/Latino, Asian) based on which, if any, of the Twitter accounts they follow. For example, because EPSN.com is popular with men, the method assumes that the `@ESPN` Twitter account is mostly followed by men.

The principal drawback of many such methods is their need for significant data per user, which is often time consuming or expensive to gather. When working with enormous datasets, researchers often avoid demographic analysis altogether, or use limited approaches. For example, a large-scale analysis by Mislove et al. (2011) inferred gender by simply string-matching common names, which failed to label 35.8% of the users studied. Paul and Dredze (2011) tracked flu and allergy symptoms in a dataset of 1.6 million tweets, in which 71% of users had only a single tweet and 97% had 5 or fewer. In a dataset with millions of users, obtaining sufficient content or network data for each user may require prohibitively many Twitter API calls. In production environments, a system may need to make rapid decisions based on a single message, rather than waiting until additional data can be gathered. For these reasons, methods have been proposed for inferring demographics based on the user's name and profile, such as for geolocation, gender, or social roles (Dredze et al., 2013; Osborne et al., 2014; Dredze, Osborne, and Kambadur, 2016; Knowles, Carroll, and Dredze, 2016; Burger et al., 2011; Beller et al., 2014).

Past work on predicting race and ethnicity of social media users has largely struggled to the lack of available resources. Crowdsourced annotation assumes that racial identity can be accurately perceived by others, an assumption that has serious flaws for gender and age (Flekova et al., 2016; Preoţiuc-Pietro, Chandra Guntuku, and Ungar, 2017). Rule-based or statistical systems for data collection may be effective (Burger et al., 2011; Chang et al., 2010), but raise concerns about selection bias: if we only label users who take a certain action, a model trained on those users may not generalize to users who do not take that action (Wood-Doughty et al., 2017). Gold-standard labels for sensitive traits requires individual survey responses, but this yields small or skewed datasets due to the expense (Preoţiuc-Pietro and Ungar, 2018). Publicly available datasets for this task have been limited to datasets of under 10,000 users, enabling relatively weak performance (see §4.4.1).

Several studies have examined the accuracy of demographic inference and the large-scale patterns it reveals. Chen et al. (2015) and Volkova, Coppersmith, and Van Durme (2014) examined the effect of different types of information on the accuracy of demographic predictions. Mislove et al. (2011) examined how inferred demographics compare to known demographics outside of Twitter in the United States and measured in what ways the user-base of Twitter is biased compared to the population as a whole. Sloan et al. (2013) performed a similar analysis of gender and language among Twitter users in the United Kingdom.

4.4 Demographic Classifiers

4.4.1 Predicting Demographics from Names Alone

Most demographic prediction methods on Twitter require either many tweets per user or substantial information about the user's friends or followers (Volkova and Bachrach, 2015; Culotta, Kumar, and Cutler, 2016). In many settings, gathering sufficient data for accurate predictions is expensive or impossible. In Chapter 9 we conduct an analysis that relies on analyzing the demographics of the authors of 13.3M tweets; existing demographic classifiers would require at least one API call per user, totalling weeks of data collection with a single Twitter API key.

Motivated by these concerns, in Wood-Doughty et al. (2018), we propose character-level models that learn a low-dimensional representation of a Twitter user's name and screen name, enabling demographic prediction from only a single tweet. Names are a reliable source of demographic information; the name `Sarah` or username

Model	Gender Acc	Gender F1	Ethnicity (3-way) Acc	Ethnicity (3-way) F1	Ethnicity (2-way) Acc	Ethnicity (2-way) F1
SVM	82.3	82.4	56.5	43.9	66.0	62.7
CNN	83.1	83.1	62.0	42.5	73.2	71.7
RNN	84.3	84.3	60.8	40.9	71.9	69.3
Content	86.2	86.1	81.0	71.6	88.9	88.1

Table 4-I. Accuracy and F1 on Twitter test data. Except for 3-way ethnicity predictions, our single-tweet models are competitive with models requiring 200 tweets per user.

therealjohn indicate gender, and names like Carlos and Wei may suggest ethnicity or race. These models expand upon prior work on that uses exact first-name matching, which only work when users use known names (Mislove et al., 2011; Liu and Ruths, 2013; Karimi et al., 2016). Neural models provide the flexibility to learn patterns in character sub-sequences, especially for Twitter names, which are irregular and can contain emojis or special characters. We consider both recurrent and convolutional models for either the Twitter name only or a combination of name and screen name.

We train our gender prediction models on a combination of 58k Twitter users (Burger et al., 2011; Volkova, Wilson, and Yarowsky, 2013) and 68k first names from the Social Security Administration list of birth names.[1] We train our race and ethnicity prediction models using the Twitter data released by Culotta, Kumar, and Cutler (2015) and Volkova and Bachrach (2015) and supplement these limited resources with race- and ethnicity-labeled name data from the North Carolina Board of Elections,[2] which contains millions of names and corresponding labels. We consider training on both Twitter and auxiliary data, but always evaluate using just Twitter data for the validation and test sets.

We compare our methods against an SVM baseline representative of prior work and a method that makes predictions from the content of 200 tweets per user.

[1] https://www.ssa.gov/OACT/babynames/names.zip

[2] http://dl.ncsbe.gov/index.html?prefix=data/

Table 4-I shows the results of our evaluation on a held-out test set of Twitter users. Compared to the SVM model, our model produces more accurate demographic predictions. The Content baseline outperforms our single-tweet methods but requires 200 tweets per user, making it infeasible in many of the practical scenarios we are interested in. The poor results for predicting race and ethnicity showed the limitations of the available data for this task. This limitation inspiried our later work on data collection and validation for race and ethnicity (see § 4.4.3). We use the best gender classifier from this work in our analysis in Chapter 9, allowing us to conduct our analysis with 133k Twitter API calls instead of 13.3M.

4.4.2 Classifying Twitter Users as Individuals or Organizations

For many applications of social media analyes, we are in particular interested in understanding the opnions or behaviors of a representative sample of individuals, in the same vein as we would want to conduct a sample of the population (Smith, Mazzuchi, and Broniatowski, 2020). In such an analysis, it becomes important to differentiate between Twitter accounts that represent an individual's perspective and those that are being used by an organization or are entirely automated. Demographic classifiers assume that Twitter accounts are linked to a single individual, which is clearly false given the existence of bots and brand marketing.

In Wood-Doughty, Mahajan, and Dredze (2018), we consider the task of predicting whether a Twitter user represents an individual person or is an organizational account. We begin with novel data collection strategies. We examine all tweets collected from Twitter's 1% feed in 2017, about 3 billion tweets, and find all users who include in their profile a URL from `linkedin.com` or `lnkd.in`. We extracted the set of unique authors of these tweets, yielding a corpus of 161k users we believe to be individuals. After finishing data collection, we randomly sampled 100 of these accounts and found that all were correctly labeled. We also use the

Model	Balanced	Full
Majority	50.0	89.5
Humanizr	**89.6**	**94.8**
N-gram	85.2	93.8
CNN	84.5	93.4

(a) Results from training on the data from McCorriston, Jurgens, and Ruths (2015). Humanizr uses 200 tweets per user.

Model	Balanced	Full
Majority	50.0	89.5
Humanizr	-	-
N-gram	84.0	94.1
CNN	**85.8**	**94.6**

(b) Results from training on our collected data. Humanizr was not evaluated due to data constraints.

Figure 4-1. Experimental results. In both experiments, the test sets are 20% of the data released by McCorriston, Jurgens, and Ruths (2015).

Twitter Lists functionality to lists labeled with key words relating to individuals or organizations, finding organizations accounts that appear in lists with key terms such as “businesses” or “companies.” Using this approach to gather about 250 lists, we collected 19k accounts labeled as individuals and 28k accounts labeled as organizations.

A drawback of this training data is that is likely unrepresentative of the Twitter user population, raising concerns of selection bias. Accounts which are added into other users’ lists are likely more popular than a randomly-selected account, and individuals who link their Twitter account to a LinkedIn page likely present a more professional appearance in their profile or tweets. This may bias our classifier to misjudge less popular organizational users or the accounts of individuals who do not use Twitter professionally.

In addition to the collected data, we propose two models for classifying users as either individuals or organizations based on a single tweet. We use a SVM following Knowles, Carroll, and Dredze (2016) and a CNN similar to our work in Wood-Doughty et al. (2018) to learn a low-dimensional representation of the user′s name. Our SVM and CNN models also incorporate features extracted from the user fields contained in the metadata of a single tweet object. Some of these features – the ratio of followers to friends, verification status, and the number of tweets – were used in previous work McCorriston, Jurgens, and Ruths (2015). We also introduce new features, such as the presence of personal pronouns (e.g. "my" vs. "our") and the use of repetitive punctuation (e.g. "!!") in users' descriptions.

To evaluate the quality of our data and methods, we consider two questions: First, how well do our proposed models perform on this task, when using only a single tweet per user, compared to the Humanizr method? Second, how useful is the dataset we created for training models to discriminate between organizations and individuals? Table 4-1 (a) shows the results for our experiment answering the first question, by evaluating our methods and prior work on the crowdworker-annotated data from McCorriston, Jurgens, and Ruths (2015). While the Humanizr method slightly outperforms both our n-gram and CNN models, it requires significantly more data per user. Table 4-1 (b) shows the result for our second experiment, evaluating our models trained on our collected dataset. The CNN improves considerably, almost matching the performance of Humanizr. In fact, in the full setting, the difference between the two is not statistically significant[3]. This provides strong evidence that our dataset, while cheaply collected with noisy labels, is valuable for classifying organizations and individuals on a random sample of Twitter.

Subsequent work has compared against our work when proposing new methods (Wang et al., 2019) or has used our classifier to study either individuals and organizations on Twitter (Stewart, Yang, and Eisenstein, 2020; Chandrasekaran et al., 2020). We also use it in our analysis in Chapter 9.

Citation	% Miss	# Users	% W	% B	% H/L	% A
Preoţiuc-Pietro and Ungar (2018)	4.7	3572	80.8	9.5	6.1	3.6
Culotta, Kumar, and Cutler (2015)	60.0	308	50.0	19.5	30.5	0
Volkova and Bachrach (2015)	36.5	3174	48.0	35.8	8.9	3.0
Total Matching Users	-	2.50M	26.8	53.8	11.3	8.1
Query-Bigram	8.1	112k	51.2	40.8	1.4	6.6
Heuristic-Filter	40.6	135k	42.2	45.9	5.6	6.4
Class-Balanced	0.0	31k	25.0	25.0	25.0	25.0

Table 4-II. Previously-published Twitter datasets annotated for race/ethnicity and datasets collected in Wood-Doughty et al. (2021). "% Miss" shows the percent of users that could not be scraped in 2019. "# Users" shows the number users that are currently available. The abbreviations W, B, H/L, and A corresponds to White, Black, Hispanic/Latinx, Asian respectively. Per-group percentages are from non-missing data.

4.4.3 Using Noisy Self-Reports to Predict Race and Ethnicity

Our analyses in Wood-Doughty et al. (2021), shown in §4.4.1, demonstrated the need for better data and methods for predicting race and ethnicity on Twitter. Even the best methods from past work do not even extend to the four most common demographic groups in the United States. The top three rows of Table 4-II show the relatively small amount of data from past work, suggesting the need for a new approach for collecting a large-scale dataset of users labeled for race and ethnicity. Ideally, such a dataset should cover the categories of standard demographic panels, should be sufficiently large to train accurate systems, and be created with a methodology that is easily reproducible to provide for more up-to-date datasets in the future as tweets are deleted over time.

We introduce a novel method for collecting data that addresses these needs by scraping Twitter users who self-report their racial identity in their Twitter profile description. We begin with simple keyword matching which matches 2.5M users but contains many false positives. We then use a collection of filters to improve the precision of our users labeled by our method. This produces the datasets in the

[3] p=0.36 when using a two-proportion t-test. For the balanced setting, Humanizr's 89.6% is significantly better than the best CNN's 85.8%, with p=0.014 using the same test.

	Imbalanced prediction						Balanced prediction					
	Names		Unigrams		BERT		Names		Unigrams		BERT	
Dataset/Baseline	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc
Random	.250	25.0	.250	25.0	.250	25.0	.250	25.0	.250	25.0	.250	25.0
Majority	.224	**80.8**	.224	80.8	.224	**80.8**	.100	25.0	.100	25.0	.100	25.0
Crowd	.268	74.9	.432	**83.2**	**.402**	74.8	.213	.322	.343	40.9	.402	43.7
QB	**.335**	71.7	.394	71.4	.371	61.0	.316	.377	.406	46.5	.461	48.3
Crowd+QB	.331	74.3	**.460**	78.4	.383	62.4	.276	.344	.453	47.6	.484	50.1
HF	.324	64.4	.401	72.4	.346	62.3	.308	.377	.418	47.3	.408	44.1
Crowd+HF	.198	54.0	.449	76.9	.360	62.1	.149	.233	**.466**	**50.9**	.441	47.4
CB	.299	49.4	.300	43.3	.285	39.0	.379	.381	.463	48.9	.474	49.0
Crowd+CB	.249	35.9	.449	74.6	.349	52.0	**.386**	**.390**	.465	48.9	**.514**	**52.6**

Table 4-III. Experimental results for baseline methods, models trained on the crowdsourced datasets, and models trained on our self-report datasets. The best result in each column is in bold.

bottom three rows of Table 4-II.

To validate the quality of our datasets, we train demographic classifiers on each of our datasets and evaluate them on a held-out test set of Twitter survey responses in which users explicitly self-report their identity. By comparing methods trained on our collected data against models trained on the existing crowdworker-labeled data, we can understand whether our data provides a useful resource for demographic classifiers.

Table 4-III shows the comparison of three different methods trained on either our datasets, previously-collected crowdsourced datasets, or a combination of the two. Because the test set was extremely imbalanced (81% of respondents were white), we consider both an evaluation on the whole test set and an evaluation on a balanced, subsampled test set. We see that in the imbalanced evaluation, no method is able to dramatically improve above the baseline method of simply the majority class. However, in the balanced case in which a method must perform well on all demographic subgroups, models trained on our collected data demonstrate large improvement over those trained on only the previously-published datasets.

We have used the classifiers developed in this work to study the Me Too social movement (Mueller et al., 2021).

4.5 Twitter User Behavior and Demographics

The work highlighted in this chapter has enabled better analyses of opinions and behaviors of Twitter users. However, even with accurate demographic inference tools, there may be other confounding factors that make it difficult to contextualize studies across demographic groups. Since social media analysis relies on *how* people use platforms, variations in usage behaviors by different demographic groups could introduce biases in analyses and alter conclusions. For example, if one group tends to use Twitter nicknames more frequently, a name-based demographic classifier may make more errors on members of that group. Alternatively, if we use profile pictures to infer demographics and users of one demographic are less likely to share pictures of themselves, our results may under-represent that group. Pavalanathan and Eisenstein (2015) studied these issues for geolocation algorithms, finding that classifiers which infer users' locations identify a target population that differs from the general population of Twitter. A Pew Report survey indicated that social media users' privacy settings do vary across demographics, but did not look at specific behaviors (Madden, 2012).

In Wood-Doughty et al. (2017) we use a suite of four classifiers for Twitter user demographics to label a million Twitter users randomly sampled from the Twitter 1% stream and analyze the co-occurrence of those labels with one another and with many simple Twitter behavior metrics such as follower-count or tweets-per-month.

		Gender				Ethnicity				Account Age	
Behavior/Data	**All**	**F 2**	**F 3**	**M 2**	**M 3**	**W**	**B**	**HL**	**A**	**O**	**N**
% of dataset	100	27.0	6.8	31.8	7.9	43.4	28.9	15.3	12.3	50.0	50.0
% with tweets from 2017	82.0	81.9	81.8	81.9	82.0	82.0	82.0	81.9	82.0	81.9	82.0
% with profile image	95.4	96.3	96.6	95.2	97.8	93.9	95.4	97.9	98.0	97.3	93.5
% with profile URL	20.8	21.3	26.5	23.7	29.3	16.8	20.3	26.1	30.0	25.1	16.6
% with description	78.0	76.1	81.0	77.0	80.7	74.1	79.1	80.7	85.3	81.0	75.0
% with profile location	53.6	54.9	62.6	57.3	66.1	48.0	53.5	61.7	63.3	58.6	48.7
% with geotagging enabled	33.1	39.1	47.5	36.0	45.2	28.2	31.0	45.4	40.0	47.2	39.1
% with 1+ geotagged tweet	7.9	10.8	15.5	10.0	14.9	6.1	6.8	13.0	10.8	11.4	4.5
% with Carmen country	17.2	23.8	32.2	22.5	32.8	15.1	15.8	24.7	18.8	21.0	13.5
% with Carmen city	8.6	11.7	16.2	11.9	18.4	7.6	8.2	11.5	9.6	11.1	6.2
(m) % from Android sources	30.5	32.0	30.0	30.3	27.8	28.8	28.7	36.6	30.9	27.2	34.6
(m) % from iPhone sources	36.9	37.9	40.7	33.5	34.0	39.5	39.7	31.2	32.1	37.7	36.0
(m) % from desktop web	9.0	9.4	10.4	11.5	15.5	7.4	7.5	12.4	12.2	9.7	8.2

Table 4-IV. Behavior across groups. For gender groups, 'M' stands for Male, 'F' for Female. '2' indicates that at least three gender classifiers agreed on the label; '3' indicates that all four did. For ethnicity groups, 'W' stands for White/Caucasian, 'B' for Black/African-American, 'HL' for Hispanic/Latino, and 'A' for Asian. For age (of account) groups, 'O' stands for old (user joined before Oct. 2014), 'N' for new. (m) indicates that a percent or average was computed via micro-averaging across users' tweets; all others are macro-averaged across users. Entries that require multiple tweets per user or timezone data are computed by ignoring the users for which that data is unavailable, which may introduce bias.

Table 4-IV shows a selection of these results. We find differences in these indicators across demographic groups, suggesting that there may be underlying differences in how different demographic groups use Twitter.

A striking result is that we find significant behavior differences between groups with differing levels of classifier consensus. All four classifiers we use predict user gender; for users on which two classifiers agree on a gender label, certain behaviors are far less likely than for users on which at least three classifiers agree on a gender label. For example, users labeled as women by at least three classifiers enabled geotagging 47.5% of the time, whereas users labeled as women by only two classifiers did so only 39.1% of the time. Such results strongly suggest that the features relied upon by Twitter demographic classifiers are confounded with Twitter behaviors. Such confounding again highlights the need for causal reasoning. If we want to draw conclusions about a hypothetical probability distribution without missing data, measurement error, or selection bias, we need to clearly state our assumptions and develop methods that correct for these biases and provide robust estimates. These concerns motivate the work we introduce in Chapters 6 and 9.

Chapter 5

Challenges of Using Text Classifiers for Causal Inference

5.1 Introduction

We can now return to our motivating examples from §1.1 and introduce a framework for combining ML and NLP methods into causal analyses. Recall from Chapter 2 that causal analyses typically use low-dimensional structured variables, such as clinical markers and binary health outcomes. Such analyses require assumptions about the data-generating process (DGP), which are often simpler with low-dimensional data. While language is inherently high-dimensional, NLP systems like those discussed in Chapter 3 can be used to process raw text to produce structured variables. This Chapter is adapted from Wood-Doughty, Shpitser, and Dredze (2018), published at EMNLP 2018.

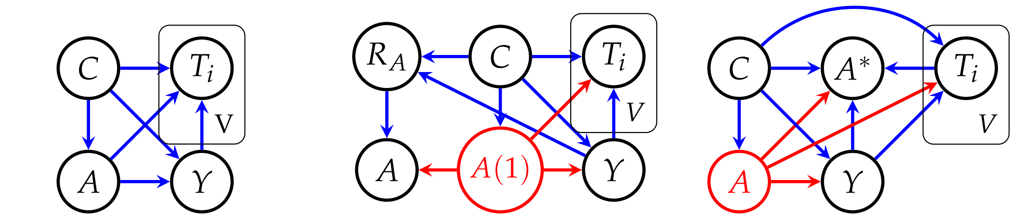
For example, work on identifying undiagnosed side effects from electronic health records (EHR) has used text classifiers to produce clinical variables from the raw text (Hazlehurst, Naleway, and Mullooly, 2009).

NLP provides a natural way to incorporate text data into causal inference models. We can produce low-dimensional variables using, for example, text classifiers, and then run our causal analysis. However, this straightforward integration belies several potential issues. Text classification is not perfect, and errors in a NLP algorithm may bias subsequent analyses. Causal inference requires understanding how variables influence one another and how correlations are confounded by common causes. Classic methods such as stratification provide a means for handling confounding of categorical or continuous variables, but it is not immediately obvious how such work can be extended to high-dimensional data.

Recent work has approached causal inference in high-dimensional domains by using random forests (Wager and Athey, 2017) and other methods borrowed from machine learning (Chernozhukov et al., 2016). But even compared to an analysis that requires hundreds of confounders (Belloni, Chernozhukov, and Hansen, 2014), NLP models with millions of variables are very high-dimensional. While physiological symptoms reflect complex biological realities, many symptoms such as blood pressure are one-dimensional variables. While doctors can easily quantify the effect of high blood pressure on some outcome, can we use the positive sentiment of a restaurant review to estimate a causal effect? More broadly, is it possible to employ text classification methods in a causal analysis?

We explore methods for the integration of text classifiers into causal inference analyses that consider confounds introduced by imperfect NLP. We draw on the assumptions discussed in Chapters 2 and 3 and discuss when those assumptions may or may not be reasonable. We draw on the causal inference literature to consider two modeling aspects: missing data and measurement error. In the missing data formulation, a variable of interest is sometimes unobserved, and text data gives us a means to model the missingness process. In the measurement error formulation, we use a text classifier to generate a noisy proxy of the underlying variable.

We highlight practical considerations of a causal analysis with text data by conducting analyses with simulated and Yelp data. We examine the results of both formulations and show how a causal analysis which properly accounts for possible sources of bias produces better estimates than naïve methods which make



(a) Simple Confounding with Text

(b) Missing Data with Text

(c) Measurement Error with Text

Figure 5-1. DAGs for causal inference, updated from Figure 2-2 to include text as a variable. In the Yelp experiments we discuss in §5.3.2, T_i influences Y and not the other way around.

unjustified assumptions. We conclude by examining how our approach may enable new research avenues for inferring causality with text data. The methods we develop in this chapter are explored further in Chapters 6, 7, and 9.

5.2 Causal Models for Text Data

Recall from Chapter 2 that conceptualizations of missing data and measurement error offer a means to identify causal effects when certain variables are sometimes or always unobserved. However, either approach depends on the data we have available and the assumptions we are willing to make about the data-generation process.

Building on the DAGs we introduced in Figure 2-1, we add new variables to represent text, producing the models in Figure 5-1a. The arrows from the structured variables (A, C, Y) to the text represent the assumption that the text is *caused* by these variables. This could make sense in our motivating examples if the content of a clinical note reflects prior information about a patient or if a social media user's demographics influence their posting behavior. We will use the text to recover the causal relationship between A and Y.

We represent text as an arbitrary set of V variables, which are independent of one another given the non-text variables. In our implemented analyses we will represent text as a bag-of-words, wherein each T_i is simply the binary indicator of the presence of the i-th word in our vocabulary of V words, and $\mathbf{T} = \cup_i T_i$. The restriction to simple text models allows us to explore connections to causal inference applications, though future work could relax assumptions of the text models to be inclusive of more sophisticated text models (e.g. neural sequence models (Lai et al., 2015; Zhang, Zhao, and LeCun, 2015)), or consider causal relationships between two text variables.

In both the missing data and measurement error frameworks below, we will assume that our treatment A is sometimes or always unobserved, and we are using a text classifier to predict A from the information contained in the text.

5.2.1 Missing Data

To show how we might use text data to recover from missing data, we introduce missingness for A from Figure 5-1a to get the model in Figure 5-1b. In the setting of our clinical notes motivating examples, the DAG in Figure 5-1b indicates that our treatment (vitamin D deficiency) is sometimes missing, and its probability of missing (only) depends on C and Y. The missing arrow from $A(1)$ to R_A encodes the Missing-at-Random assumption, which is sufficient to make it possible to identify the full data distribution from the observed data.

The causal effect of A on Y in Figure 5-1b is identified as τ_{MD}, given in Eq. 2.16 in Figure 2-4. The derivation is identical to that in §2.2.1, except we add the text $\mathbf{T}$ into the requisite distributions.

5.2.2 Measurement Error

We can model text data with measurement error by introducing a proxy A^* to the model in Figure 5-1c. We assume that the proxied value of A^* can depend upon all other variables, and that we will be able to estimate $p(A^*, A)$ given an external dataset, e.g. text classifier accuracy on held-out data.

Suppose we again want to estimate the causal effect of A on Y, but A never appears in our dataset. This might be impossible unless we have some way to infer A from the clinical text available. If we have training dataset of clinical notes annotated with patient's vitamin D levels and the text of the notes from their primary care physician. If the physician's notes contain information about whether they take a multivitamin or questions about their diet, we may be able to train a classifier that could predict, with some accuracy, whether they are at risk of a vitamin D deficiency.

Working from the derivation for matrix adjustment in binary models given by Pearl (2010), we identify the causal effect of A on Y (Figure 5-1c) as τ_{ME} (Eq 2.17 in Figure 2-4.) The derivation is identical to that in §2.2.2, except we add the text $\mathbf{T}$ into the requisite distributions.

As discussed in §2.4, our conceptualization of measurement error can be tied to questions of internal validity and construct validity from Cook, Campbell, and Shadish (2002). The assumption that we can measure $p\left(A^*|A\right)$ can satisfy questions of internal validity, because we do not have to assume that A and A^* are equivalent. However, the assumptions we have introduced do not tell us how the unobserved A we are "recovering" from the noisy A^* is connected to any underlying higher-order concept. In our example, it may be that vitamin D deficiency is itself a proxy used by clinicians to assess a patient's general level of nutrition (Schöttker et al., 2014), or that a patient's own knowledge of their vitamin D levels (which might be reflected in clinical notes) is unreliable (Amiri et al., 2017). Understanding whether a study of clinical notes can provide valid causal inferences requires both provable internal validity under a set of assumptions, as well as a careful discussion with domain experts of whether these assumptions are plausible.

Missing Data

$$\begin{aligned} C &\sim \text{Ber}(0.4) \\ A(1) &\sim \text{Ber}(-0.3C + 0.4) \\ Y &\sim \text{Ber}(0.2C + 0.1A + 0.5) \\ T_i &\sim \text{Ber}(0.5 + u_i A + v_i C) \\ R_A &\sim \text{Ber}(0.7 + 0.2C - 0.4Y + \sum_i w_i T_i) \end{aligned}$$

Measurement Error

$$\begin{aligned} C &\sim \text{Ber}(0.4) \\ A &\sim \text{Ber}(-0.3C + 0.4) \\ Y &\sim \text{Ber}(0.2C + 0.1A + 0.5) \\ T_i &\sim \text{Ber}(0.5 + s_i C + u_i A + v_i Y) \end{aligned}$$

Figure 5-2. The synthetic data generating-processes. Ber(p) is a Bernoulli distribution with probability p.

5.3 Experiments

We now empirically evaluate the effectiveness of our two conceptualizations (missing data and measurement error) for including text data in causal analyses. We induce missingness or mismeasurement of the treatment variable and use text data to recover the true causal relationship of that treatment on the outcome. We begin with a simulation study with synthetic text data, and then conduct an analysis using reviews from `yelp.com`.

5.3.1 Synthetic Data

We select synthetic data so that we can control the entire data-generation process. For each data row, we first sample data on three binary variables (A, C, Y) and then sample V different binary variables T_i representing a V-vocabulary bag-of-words. A graphical model for this distribution appears in Figure 5-1a. We augment this distribution to introduce either missing data (Figure 5-1b) or measurement error (Figure 5-1c.) For measurement error, we sample two datasets. A small training set which gives data on $p(A, C, Y, \mathbf{T})$ and a large test set which gives data on $p(C, Y, \mathbf{T})$.

In Figure 5-2, s_i, u_i and v_i are the effect of C, A, and Y on the probability of word

T_i; each is drawn from $\mathcal{N}(0, \zeta)$, a parameter which controls how correlated words are with the underlying variables. When ζ is close to 0, the words are essentially random. When ζ is large, the words are essentially deterministic functions of the underlying variables. Similarly w_i is the effect of word T_i on R_A, and is drawn from $\mathcal{N}(0, \eta)$.

For both settings, we set vocabulary size to 4,334 and $\zeta = 0.5$. For the missing data setting, we set $\eta = 0.1$. We picked these constants by empirically finding a reasonable middle ground between the text data providing only noise and being a deterministic function of their parents. We picked all other constants such that the naïve correlation $p(Y \mid A)$ was a poor estimate of the counterfactual $p(Y(a))$ in the full-data setting.

5.3.2 Yelp Data

We utilize the 2015 Yelp Dataset Challenge[1] which provides 4.7M reviews of local businesses. Each review contains a one- to five-star rating and up to 5,000 characters of text. Yelp users can flag reviews as "Useful" as a mark of quality. Using this dataset – and under the dubious assumption that these are the only variables that matter – we will try to estimate the causal effect of writing a positive (versus negative) review on receiving a Useful flag.

We extract treatment, outcome, and confounder variables from the structured data. The treatment is a binarized user rating that takes value 1 if the review has four or five stars and value 0 if the review has one or two stars. Three-star reviews are discarded from our analysis. The outcome is whether the review received at least one Useful flag. The confounder is whether the review's author has received at least two Useful flags across all reviews, according to their user object. In our data, 74.2% of reviews were positive, 42.6% of reviews were flagged as Useful, and

[1] yelp.com/dataset/challenge

56.7% users had received at least two such flags. We preprocess the text of each review by lowercasing, stemming, and removing stopwords, before converting to a bag-of-words representation with the 4,334 word vocabulary of all words which appeared at least 1000 times in a sample of 1M reviews.

Based on this $p(A, C, Y, \mathbf{T})$ distribution, we assume the DGP that matches Figure 5-1a and introduce missingness and mismeasurement as before, giving us DGPs matching Figures 5-1b and 5-1c.

Our intention is not to argue about a true real-world causal effect of Yelp reviews on peer behavior: we do not believe that our confounder is the only common cause of the author's rating and the platform's response. We leave for future work a case study that jointly addresses questions of identifiability and estimation of a real-world causal effect. In this work, our experiments focus on a simpler task: can a correctly-specified model that uses text data effectively estimate a causal effect in the presence of missing data or measurement error.

5.3.3 Models

We now introduce several baseline methods which, unlike our correctly specified models τ_{MD} and τ_{ME}, are not consistent estimators of our desired causal effect. We would expect that the theoretical bias in these estimators would result in poor performance in our experiments.

5.3.3.1 Baseline: Naïve Model

In both the missing data and measurement error settings, our models use some rows that are full observed. In missing data, these are rows where $R_A = 1$; in measurement error, the training set is sampled from the true distribution. The simplest approach to handling imperfect data is to throw away all rows without full data, and calculate Eq 2.15 from that data. In Figure 5-3, these are labeled as

`*.naive`.

5.3.3.2 Baseline: Textless Model

In Figure 5-1b, if we do not condition on T_i to d-separate $A(1)$ from its missingness indicator, that influence may bias our estimate. While we know that ignoring text may introduce asymptotic bias into our estimates of the causal effect, we empirically evaluate how much bias is produced by this "Textless" model compared to a correct model. This is labeled as `*.no_text` in Figure 5-3 (a).

In principle, we could conduct a measurement error analysis using a model that does not include text. In practice, we found we could not impute A^* from C and Y alone. The non-textual classifier had such high error that the adjustment matrix was singular and we could not compute the effect. Thus, we have no such baseline in our measurement error results.

5.3.3.3 Baseline: `no_y` and `unadjusted` Models

In Figure 5-1b, we must also condition on C and Y to d-separate $A(1)$ from its missingness indicator. In our misspecified model for missing data, we do not condition on Y, leaving open a path for $A(1)$ to influence its missingness. In Figure 5-3 (a), this model is labeled as `*.no_y`.

When correcting for measurement error, a crucial piece of the estimation is the matrix adjustment using the known error between the proxy and the truth. A straightforward misspecified model for measurement error is to impute a proxy for each row in our dataset and then calculate the causal effect assuming no error between the proxy and truth. This approach, while simplistic, can be thought of as using a text classifier as a proxy without regard for the text classifier's biases. In Figure 5-3 (b), this approach is labeled as `*.unadjusted`.

5.3.3.4 Correct Models

Finally, we consider the estimation approaches presented in §5.2.1 and §5.2.2. For the missing data causal effect (τ_{MD} from Eq 2.16) we use a multiple imputation estimator which calculates the average effect across 20 samples from $p(A|\mathbf{T}, C, Y)$ for each row where $R_A = 0$. For the measurement error causal effect (τ_{ME} from Eq 2.17), we use the training set of $p(A, C, Y, \mathbf{T})$ data to estimate $\epsilon_{c,y}$ and $\delta c, y$ and the larger set of $p(C, Y, \mathbf{T})$ data to estimate $q_{c,y}$ and $p(C)$.

These models are displayed in Figure 5-3 (a) as `*.full` and in Figure 5-3 (b) `*.adjusted`.

5.3.4 Evaluation

Each model takes in a data sample with missingness or mismeasurement, and outputs an estimate of the causal effect of A on Y in the underlying data. Rather than comparing models' estimates against a population-level estimate, we compare against an estimate of the effect computed on the same data sample, but without any missing data or measurement error. This 'perfect data estimator' may still make errors given the finite data sample. We compare against this estimator to avoid a small-sample case where an estimator gets lucky. In Figure 5-3, we plot data sample size against the squared distance of each model's estimate from a perfect data estimator's estimate, averaged over ten runs. Figure 5-4 shows a second set of experiments using a larger vocabulary.

Figure 5-4 shows the results of a second set of experiments, which are identical to those described in §5.3 except the vocabulary size is now 53,197 instead of 4,334. For the Yelp data, the larger vocabulary consists of all words which appear at least ten times in a sample of 1M reviews. As the larger vocabulary introduced greater memory requirements, we did not run these experiments with as large of datasets.

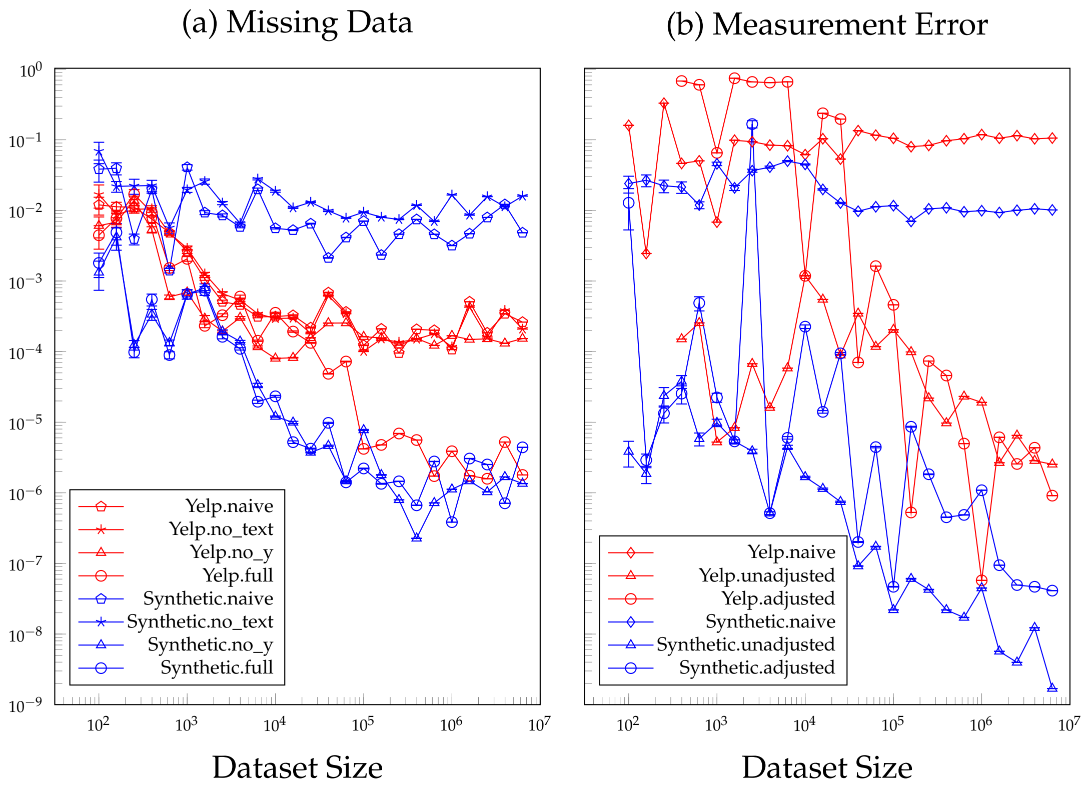


Figure 5-3. Experimental results. Squared distance (y-axis, lower is better) of the estimated causal effect from τ_{SC} calculated from the full data with no missing data or measurement error. Error bars (negligible for larger datasets) are 1.96 times standard error across 10 experiments.

5.4 Results

Figures 5-3 and 5-4 show our experimental results. Both plots use a logarithmic scale for both axes to visualize large-scale trends, showing sample size and causal estimation error. For unbiased methods, we would expect squared error to decrease as sample size increases. For the most naive methods, we see as expected that error does not substantially decrease as sample size increases.

Given that our correctly-specified models are provably unbiased, we would expect them to outperform misspecified models. However, for any given dataset, asymptotic results provide no finite-sample guarantees.

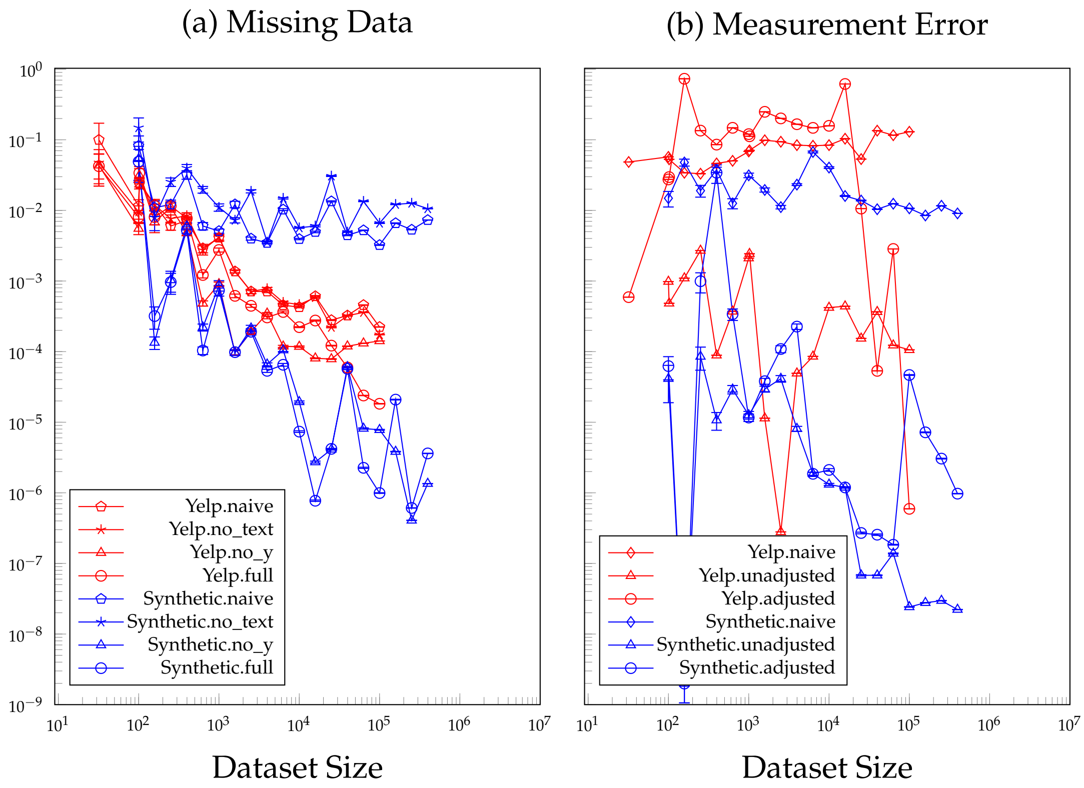


Figure 5-4. Experimental results with a vocabulary of size 53,197. Squared distance (y-axis, lower is better) of the estimated causal effect from τ_{SC} calculated from the full data with no missing data or measurement error. Error bars (negligible for larger datasets) are 1.96 times standard error across 10 experiments.

5.4.1 Missing Data

The missing data (MD) experiments suggest that the correct `full` model does perform best. The `no_y` model performs approximately as well as the correct model on the synthetic data, but not on the Yelp data. The difference between the `no_y` and `full` missing data models is simply a function of the effect of Y on R_A. We could tweak our synthetic data distribution to increase the influence of Y to make the `no_y` model perform worse.

When we initially considered other DGPs for missing data, we found that when we reduced the influence of the text variables on R_A, the `no_text` and `naive` models approached the performance of the correctly-specified model. While intuitive, this reinforces that the underlying distribution matters a great deal in how modeling choices may introduce biases if incorrectly specified.

5.4.2 Measurement error

The measurement error results tell a more interesting story. We see enormous fluctuations of the `adjusted` model, and in the synthetic data, the `unadjusted` model appears to be quite superior.

In the synthetic dataset, this is likely because our text classifier had near-perfect accuracy, and so simple approach of assuming its predictions were ground-truth introduced less bias. A broader issue with the `adjusted` model is that the matrix adjustment approach requires dividing by (potentially very small) probabilities, this sometimes resulted in huge over-corrections. In addition, since those probabilities are estimated from a relatively small training dataset, small changes to the error-estimate can propagate to huge changes in the final casual estimate.

This instability of the matrix adjustment approach may be a bigger problem for text and other high-dimensional data: unlike in our §2.2.2 example of BMI and obesity, there are likely no simple relationships between text and clinical variables. However, instead of using matrix adjustment as a way to recover the true effect, we may instead use it to bound the error our proxy may introduce. As mentioned by Pearl (2010), when $p(A \mid A^*)$ is not known exactly, we can use a Bayesian analysis to bound estimates of a causal effect. In a downstream task, this would let us explore the stability of our `adjusted` results.

5.5 Related Work

There is a conceptually related line of work in the NLP community on inferring causal relationships expressed in text (Girju, 2003; Kaplan and Berry-Rogghe, 1991).

However, this work is fundamentally different. Rather than identify casual relations expressed via language, we are using text data in a causal model to identify the strength of an underlying causal effect.

At the time this work was initially published, only a few papers had considered the possibilities for combining text data with approaches from the causal inference literature. Landeiro and Culotta (2016) and Landeiro and Culotta (2017) explored text classification when the relationship between text data and class labels are confounded. Other work has used propensity scores as a way to extract features from text data (Paul, 2017) or to match social media users based on what words they write (De Choudhury et al., 2016). The only work we know of which seeks to estimate causal effects using text data focuses on effects *of* text or effects *on* text (Egami et al., 2018; Roberts, Stewart, and Nielsen, 2018). In our work, our causal effects do not include text variables: we use text variables to recover an underlying distribution and then estimate a causal effect within that distribution.

Since this work was published, there has been a great deal of interest at the intersection of these methods. Keith, Jensen, and O'Connor (2020) provides a comprehensive overview of methods that use text as a means to correct for unobserved confounding, including the use of contextualized embeddings to model text (Veitch, Sridhar, and Blei, 2020) or an adversarial matching approach (Yao et al., 2019). In Chapter 7, we compare our measurement error formulation with alternative methods in an evaluation using better synthetic data.

5.6 Future Directions

While this chapter addresses some initial issues arising from using text classifiers in causal analyses, many challenges remain. We highlight some of these issues as directions for future research.

One challenge identified by this work is that the provably-unbiased estimator in the measurement error setting is outperformed by the `unadjusted` estimator. Chapter 6 provides a closer look at the reason behind this behavior, and proposes sensitivity analyses that help mitigate this issue for downstream applications.

We provided several proof-of-concept models for estimating effects, but our approach is flexible to more sophisticated models. For example, a semi-parametric estimator would make no assumptions about the text data distribution by wrapping the text classifier into an infinite-dimensional nuisance model (Tsiatis, 2007). This would enable estimators robust to partial model misspecification (Bang and Robins, 2005).

Choices in the design of statistical models of text consider issues like accuracy and tractability. Yet if these models are to be used in a causal framework, we need to understand how modeling assumptions introduce biases and other issues that can interfere with a downstream causal analysis. To take an example from the medical domain, we know that doctors write clinical notes throughout the healthcare process, but it is not obvious how to model this DGP. We could assume that the doctor's notes passively record a patient's progression, but in reality it may be that the content of the notes themselves actively change the patient's care; causality could work in either direction. In Chapter 7 we introduce work that explores how different assumptions about the relationship between text and structured variables in a DGP affect causal analyses.

New lines of work in causality may be especially helpful for NLP. In this work, we used simple logistic regression on a bag-of-words representation of text; using state-of-the-art text models will likely require more causal assumptions. Nabi and Shpitser (2017) develops causality-preserving dimensionality reduction, which could help develop text representations that preserve causality.

Finally, we are interested in case studies on incorporating text classifiers into real-world causal analyses. Many health studies have used text classifiers to extract clinical variables from EHR data (Meystre et al., 2008). These works could be extended to study causal effects involving those extracted variables, but such extensions would require an understanding of the underlying assumptions. In any given study, the necessity and appropriateness of assumptions will hinge on domain expertise. The conceptualizations outlined in this chapter, while far from solving all issues of causality and text, will help those using text classifiers to more easily consider research questions of cause and effect.

Chapter 6

Sensitivity Analyses for Incorporating Machine Learning Predictions into Causal Estimates

6.1 Introduction

This chapter immediately builds off the measurement error framework introduced in §5.2.2. Our experiments in Chapter 5 showed that a theoretically-justified causal estimator does not always outperform a more naive approach. Therefore, we seek to understand the relationship between the predictive accuracy of a machine learning (ML) classifier and the empirical behavior of a causal estimator that relies on that classifier. This chapter was adapted from Wood-Doughty, Shpitser, and Dredze (2020), published at the CDML workshop at NeurIPS 2020.

As we have discussed in Chapter 3, ML methods are widely studied, and often demonstrate exceptional predictive accuracy, but such performance does not provide guarantees on the consistency of downstream analyses (Obermeyer and Emanuel, 2016). For a causal analysis that we hope can inform clinical decision-making, how accurate does the ML classifier need to be to produce a result we can trust? (Jiang et al., 2018; Chen and Asch, 2017). We cannot expect a simple answer such as, "doctors should only trust analyses that use a classifier with greater than 95%

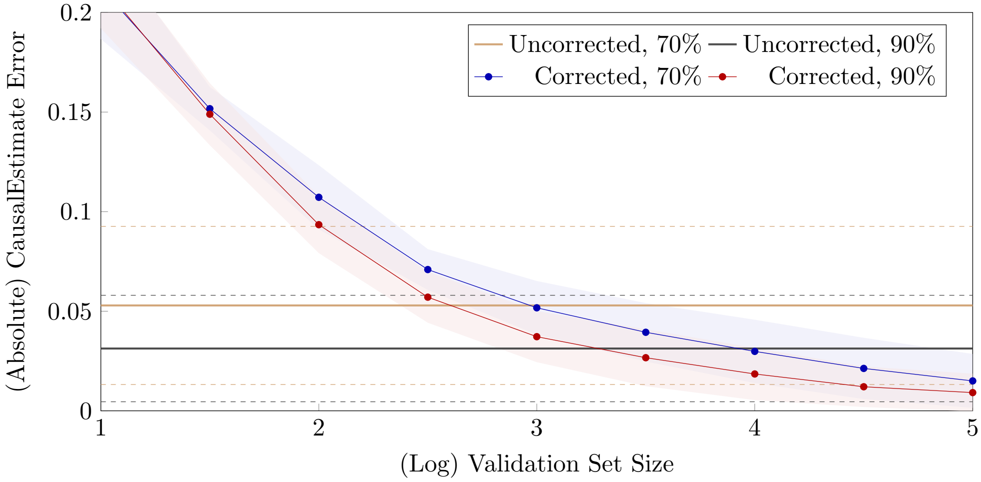


Figure 6-1. The empirical motivation for this work. Ground truth causal effect is at $y = 0$. X-axis shows log size of the validation set used to estimate $p(U^* \mid U)$, the error rate of the classifier. Each line shows the causal error for an estimator that relies upon a classifier with fixed accuracy (either 70% or 90% accurate). For the corrected estimators which rely upon $p(U^* \mid U)$, causal error decreases as the validation size increases. When the validation set is too small, a naive uncorrected estimator outperforms a theoretically-sound corrected estimator. Experimental details are in §6.4.

accuracy.” Instead, the error rate dictates what analyses are possible. We seek to connect classifier error to uncertainty in downstream causal analyses.

As we have shown in Chapter 5, our measurement error framework produces an estimator that is a function of the ML classifier’s predictions and an estimate of its error rates, drawing on measurement error literature (Pearl, 2010; Miao, Geng, and Tchetgen, 2018). We adopt this framework but use simulation studies to show this estimator performs poorly in many finite-sample cases. We introduce three sensitivity analyses to quantify the uncertainty of this estimator. We evaluate the coverage properties of these methods on our synthetic datasets and show they enable more robust analyses. We demonstrate our methods and discuss how their assumptions map onto a specific Twitter dataset and demographic classifier.

6.2 Background and motivation

We return again to the motivating example of a retrospective analysis of electronic health records (EHR). If a variable such as socioeconomic status (SES) is a common cause of our exposure (vitamin D deficiency) and our outcome (preeclampsia), we need to account for SES in our analysis. However, in many EHR datasets SES may not be explicitly recorded, but rather only indirectly noted in free-text clinical notes (McVeigh et al., 2016; Wu et al., 2013). To avoid the cost and privacy concerns of human annotators reading clinical notes and inferring structured variables, we need methods that harmonize with a valid causal analysis.

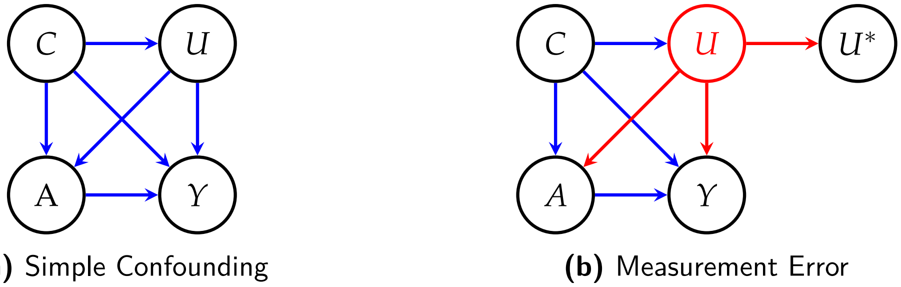
Causal inference uses observational data to reason about hypothetical interventions; in our example, "does vitamin D deficiency increase the risk of preeclampsia?" Causal models and their requisite assumptions are often represented in directed acyclic graphs (DAGs) (Pearl, 2009). In such a model, we can connect counterfactual random variables to observed random variables with assumptions. Assuming all relevant variables are observed, we can use the *g-formula* to write a causal effect as a function of observed data (Robins, 1986). In our example, suppose the only common causes of vitamin D deficiency (A) and preeclampsia (Y) are age (C) and SES (U). Then Figure 6-2a represents a causal DAG, and the counterfactual $Y(a)$, represents "onset of preeclampsia if vitamin D deficiency, possibly contrary to fact, had been set to a." Its expectation is identified via the g-formula as $E[Y(a)] = \sum_{c,u} E[Y \mid A = a, c, u] p(c, u)$. If U is never observed and we cannot infer it, then $E[Y(a)]$ is *not identified* and we cannot proceed with a causal analysis. We are interested in cases in which we do not observe U but an ML classifier can produce a noisy proxy variable, U^*, for U. Figure 6-2b shows a causal DAG in which the U is unobserved but we have access to a proxy U^*. While the g-formula cannot be used in this model, we can identify the counterfactual $Y(a)$ from Figure 6-3 (Pearl, 2010).

Questions of mismeasured data and measurement error have been widely studied in diverse fields. Measurement error has been a central concern in epidemiology research for decades (Fleiss and Shrout, 1977; Willett, 1989; Carroll et al., 2006; Cessie et al., 2012). The statistics literature has also considered questions of measurement error, from parametric models (Stefanski and Carroll, 1985) to semiparametric models (Sinha and Ma, 2014; Wang and Wang, 2015). Most relevant to our work is (Yang and Ding, 2019), which considers causal inference under unobserved confounding when that confounder is observed in a validation set[1]. Measurement error concerns arise in the use of structural equation models across disciplines (Grewal, Cote, and Baumgartner, 2004; Bollen, Gates, and Fisher, 2018). Information systems (IS) and management science has modeling noisy measurements, e.g. in studying consumer preferences (Eliashberg and Hauser, 1985). Recent IS work has also considered a setting similar to ours, where measurement error occurs due to machine learning model predictions (Yang et al., 2018; Yang et al., 2019). In this work, we draw from recent work on non-parametric identification of causal effects under measurement error (Pearl, 2010; Kuroki and Pearl, 2014; Oktay, Atrey, and Jensen, 2019).

This work closely follows Chapter 5 in considering measurement error as an approach to account for errors produced by natural language processing (NLP) classifiers. Throughout, we will assume that we have access to a classifier f that produces U^* with some (unknown) error distribution $p(U^* \mid U)$. We will also assume we have a small validation dataset with full data on $p(U)$ which we can use to estimate $p(U^* \mid U)$. Then, we have a large dataset without U on which we can apply our classifier f to produce a dataset $p(C, A, Y, U^*)$. Using the 'effect

[1] A primary difference in our methods is that our approach is applicable (under a non-differential error assumption) when our validation set only contains U and U^*.

restoration' approach proposed by (Pearl, 2010), we can then estimate our desired causal effects.



(a) Simple Confounding

(b) Measurement Error

Figure 6-2. Causal DAGs. In (b), missing arrows to U^* from C, A, Y represents a non-differential error assumption (see § 6.6.1). Throughout this chapter, we assume for simplicity of presentation that all variables are binary, though the measurement error correction only relies on U being discrete. Recent work has explored recovering from measurement error more generally (Miao, Geng, and Tchetgen, 2018).

The theoretically-sound estimator we introduced in Chapter 5, however, has a counter-intuitive empirical trend. It can be outperformed in practice, even at large finite samples, by a naive estimator that assumes $U^* = U$ (Oktay, Atrey, and Jensen, 2019). To understand this limitation, we conceptualize their method as a two step approach: the classifier which produces a $p(U^*, C, A, Y)$ distribution and a 'corrector' which estimates $p(U^* \mid U)$ and adjusts the causal estimate. Each of these steps is imperfect, e.g. due to finite sample variability. The classifier error is how often the predicted U^* differs from the true U, and depends on the size of the training data. The corrector error is the difference between the true error rate $p(U^* \mid U)$ and our estimate of that error rate from the examples in our validation dataset.

Using simulation studies that we will discuss in §6.4, we show in Figure 6-1 that the method fails if and only if the corrector step fails. If we have low corrector error, our causal estimate will be accurate; and regardless of classifier accuracy, high corrector error will bias our estimates. Thus, we should seek to quantify the uncertainty of the corrector step. Rather than trust an uncertain point estimate, we

$$
\begin{aligned}
\tau_{U^*} = \sum_{c,u} \Bigg[& \frac{p(c,u^*) - \epsilon_u P(c)}{1 - \epsilon_u - \delta_u} \\
& \cdot \left(\frac{p(Y=1,u^* \mid A=1,c) - \epsilon_u p(Y=1 \mid A=1,c)}{p(u^* \mid A=1,c) - \epsilon_u} - \frac{p(Y=1,u^* \mid A=0,c) - \epsilon_u p(Y=1 \mid A=0,c)}{p(u^* \mid A=0,c) - \epsilon_u} \right) \Bigg]
\end{aligned}
$$

Figure 6-3. Estimand (τ_{U^*}) for the causal effect of A on Y in the DAG given in Figure 6-2b. Define $\epsilon_u = p(U^* = u \mid U \neq u)$ and $\delta_u = p(U^* \neq u \mid U = u)$. All variables are assumed binary for simplicity of presentation. The derivation is a slight modification of the measurement error derivation shown in § 2.2.2 and discussed in Chapter 2.

want reliable bounds on the causal effect. We introduce three sensitivity analyses that propagate uncertainty from the corrector step to our final causal estimate.

6.3 Sensitivity analyses for the error estimate

A causal analysis typically outputs a parameter estimate that reflects some real-world phenomenon which may be impossible to further validate. If we know our estimator can be unreliable under certain conditions, how do we know when to trust that a causal estimate is accurate?

Consider a plug-in estimator for τ_{U^*} identified via the functional in Figure 6-3 (with the derivation following (Pearl, 2010) as in § 2.2.2). The 'corrector' step of this estimator involves dividing by an estimate of the error term; small changes to that estimate may result in large changes to the overall causal estimate. A sensitivity analysis for the estimate of the error rate allows us to explore how the final estimate would change as the error rate estimate changes. Rather than accepting a point estimate as our 'best guess,' our uncertainty in $p(U^* \mid U)$ should be represented in an interpretable manner. To explore ways to make these methods more robust and interpretable, we introduce three sensitivity analyses that can capture the uncertainty in the $p(U^* \mid U)$ estimate.

Each of our sensitivity analyses will introduce a sensitivity parameter, γ, which controls the trade-off between the width and coverage of our intervals[2]. Our methods are designed such that $\gamma = 0$ returns a point estimate (no interval) and as we increase γ, the interval widens. As our outcome Y is binary, a maximally-wide interval for τ_{U^*} spans from -1 to 1.

An alternative to the sensitivity analysis approach taken here is to obtain confidence intervals for τ_{U^*} using ideas from the post-selection inference literature (Berk et al., 2013; Reid, Taylor, and Tibshirani, 2017; Lee et al., 2016). Popular existing methods of this type have often been applied in parametric settings, and do not translate in a straightforward way to the setting we consider here, where estimation of functionals corresponding to causal effects does not employ parametric nuisance models.

While our sensitivity analyses provide a means of bounding our causal estimand by encoding the uncertainty of the classifier's error rate, another approach to bounding the causal effect follows from the long history of work on partial identification of causal effects in the presence of measurement error (Manski, 1990; Balke and Pearl, 1997; Drton, Sturmfels, and Sullivant, 2009; Evans, 2016). Whereas our estimand is *in theory* point-identified and our approaches help quantify the *finite-sample* uncertainty, methods on partial identification provide theoretical bounds on our causal parameter. While in general point identification is preferrable, partial identification may be a (or the only) possible approach under weaker assumptions. Depending on the domain knowledge or whether we have access to the $p\left(U^*|U\right)$ error rate, the assumptions we make in this chapter may not be plausible.

In particular, we highlight Finkelstein et al. (2020) as a potential direction for future work that could combine our sensitivity analyses with the bounds introduced by partial identification. In many clinical settings, direct knowledge of the $p\left(U^*|U\right)$ error rate may be impossible. While work on proximal causal inference can recover

[2] Each method also uses, but is empirically robust to changes in, a hyperparameter k.

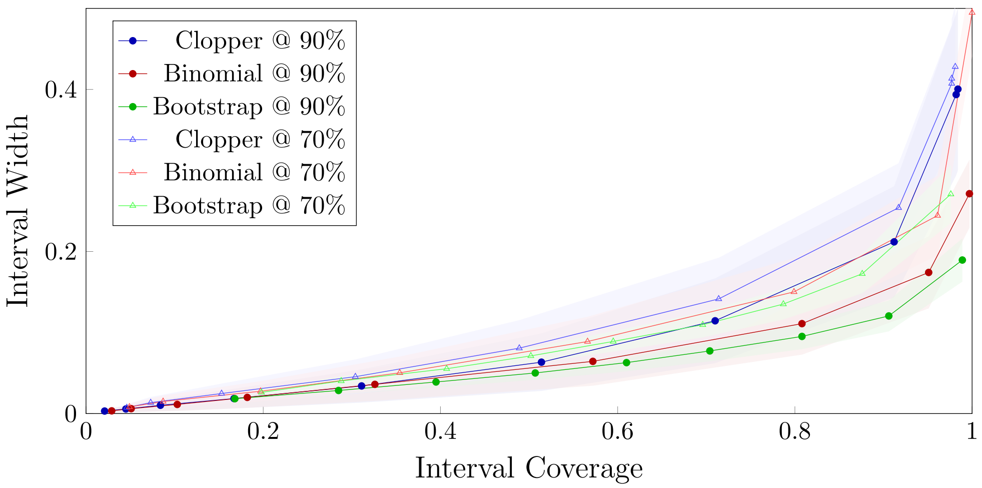


Figure 6-4. Comparison of proposed sensitivity analyses for the corrector step. Clopper refers to the Clopper-Pearson confidence interval, Binomial to the binomial sampling of error rates, and Bootstrap to a non-parametric bootstrap resampling. Noisy-oracle classifier has either 70% or 90% accuracy. As the sensitivity analyses' hyperparameter increases, both width and coverage increase. Bootstrap provides the best trade-off between coverage and width.

from measurement error or unmeasured confounding when multiple proxies are available, such methods require additional assumptions that may not always be plausible (Tchetgen et al., 2020). Finkelstein et al. (2020) allows a measurement error analysis to first collect a list of plausible assuptions based on domain expertise, and then to explore options for bounding the causal effect under (only) those assumptions if point identification is not possible. Such an approach would in general be orthogonal to the sensitivity analyses we introduce in this chapter, as they provide an estimate of finite-sample uncertainty, rather than placing bounds on the causal parameter directly.

The next three subsections discuss theoretical and empirical trade-offs of each of the three sensitivity analyses we introduce; we evaluate each on synthetic data in § 6.4.

6.3.1 Bootstrap resampling

Our first approach to a sensitivity analysis draws from non-parametric bootstrap (Hall, 1988). A classical statistical approach would bootstrap the entire estimator, retraining ML methods many times on many resampled datasets for the estimate of the causal estimand. Given that many ML models can take from hours to months to train, bootstrapping our entire analysis from training data to causal estimate is unrealistic (Krizhevsky, Sutskever, and Hinton, 2012; Devlin et al., 2018). But it is computationally easy to calculate our estimate of the error rate $p(U^* \mid U)$ and use it to compute our estimand.

Our bootstrap sensitivity analysis involves resampling k validation datasets of the same size as the original validation dataset. On each bootstrapped validation set, we calculate our error distribution $p(U^* \mid U)$ and use it to compute τ_{U^*}. This method gives us k τ_{U^*} estimates. To build the curve in Figure 6-4, we plot the intervals given by the middle $\gamma_{\text{Bootstrap}}$% of these k estimates; when $\gamma_{\text{Bootstrap}}$ = 0, our interval has width 0, and when $\gamma_{\text{Bootstrap}} = 100$ our interval endpoints are the min and max of these k estimates. For our experiments, we also set $k = 100$ and find that this allows for good coverage of our causal effect in simulated studies. Depending on the practical setting, good coverage may be possible with a lower k or may require a larger k. A disadvantage of this method is that it requires full access to the validation set. If we are using a classifier validated on data that is proprietary, private or otherwise inaccessible, we would need to collect a new validation set.

6.3.2 Binomial sampling of error rates

Our second sensitivity analysis again relies on sampling, but instead samples synthetic error rates from a binomial distribution. Our point estimate of the error rate $p(U^* = u' \mid U = u)$, assuming U is binary, consists of counting which validation set examples our classifier got right. We can model this as a binomial distribution $B(n, p)$, where 'success probability' p is our point estimate of $p(U^* = c' \mid U = c)$ and the 'number of trials' n is the number of validation set examples where $p(U = c)$. Our sensitivity analysis samples $n\tilde{p} \sim B(n, p)$ and constructs a new error rate $p(U^* \mid U) = \tilde{p}$, which we use to calculate our estimand τ_{U^*}.

To construct an interval from this approach, we sample k such error rates[3] and use the 2.5 and 97.5 percentiles as the bounds of our interval. To trade off between width and coverage in Figure 6-4, we redo this method several times, replacing the true validation dataset size n with a smaller, "synthetic sample size," n'. As this n' decreases, the variability in the sampled error rates increases, which widens the resulting interval. We let $n' = n^{1/(1+\gamma_{\text{Binomial}})}$; as γ_{Binomial} increases from zero, the "synthetic sample size" decreases from n. This approach has the advantage of not requiring access to the full validation set; in addition to the $p(U^* = u' \mid U = u)$ point estimate, we only need to know n, the number of validation set examples that estimate was computed on. However, this method makes the assumption that each error rate is binomially-distributed around its mean.

6.3.3 Clopper-pearson confidence interval

Our third approach also makes a binomial assumption about the error rates. We can replace each $p(U^* \mid U)$ estimate with an interval, and then propagate the uncertainty of each interval into the final causal estimate. Because our error rates are proportions, one reasonable choice of interval is the Clopper-Pearson method (Clopper and Pearson, 1934). If we have computed a point estimate of $p(U^* = c' \mid U = c) = p$ using n examples in our validation set, we compute a 95% Clopper-Pearson interval as the 2.5 and 97.5 percentiles of a Beta distribution with parameters $(np, n - np + 1)$ and $(np + 1, n - np)$. To propagate this interval

[3] We find that $k = 100$ allows for intervals with good coverage in our experiments.

for our error rate to an interval for our final causal estimand, we chosen $k = 20$ evenly-spaced values along the interval for each error rate, and then use all combinations of those error-rate values to estimate τ_{U^*}. Finally, we use the min and max of these resulting estimates as the bounds of our interval. To trade off between width and coverage in Figure 6-4, we take the same approach as for the binomial sampling method, replacing the true n with a smaller $n' = n^{1 \big/ (1+\gamma_{\text{Clopper}})}$. As γ_{Clopper} increases from zero, the resulting interval widens. The method has the same advantage as the binomial approach in that it only requires knowing the size of the validation set.

Note that for all three analyses we have thus-far assumed $p(U^* \mid U, C, A, Y) = p(U^* \mid U)$; this is known as a 'non-differential error' assumption in the measurement error literature (Carroll et al., 2006). If this is not true, we have differential error: rather than estimating two error rates $p(U^* = 0 \mid U = 1)$ and $p(U^* = 1 \mid U = 0)$, a k-variable DAG requires estimating 2^k error rates or positing a model for the error rate. We evaluate our methods in differential error settings in § 6.6.1.

6.4 Synthetic experiments

We conduct several simulation studies on synthetic datasets to explore the behavior of the measurement error estimator with and without our sensitivity analyses. The goal of these experiments is to understand where past work fails, and how our proposed analyses demonstrate improvements in the robustness and interpretability of the causal estimates. We first need to parameterize the distributions from which our synthetic datasets are drawn. We build on our work from Chapter 5, but allow for much broader evaluations. Rather than limiting ourselves to a single data-generating distribution, we can sample arbitrarily-many $p(C, A, Y, U)$ distributions and then sample data from each. In our experiments, we evaluate each method on ten different distributions. To make comparisons more consistent, we restrict $p(C, A, Y, U)$ such that the true causal effect of A on Y is equal to 0.1.

6.4.1 Synthetic evaluation of τ_{U^*}

We now return to the estimator proposed in past work, and re-examine Figure 6-1 to show how it fails in certain settings. We split the estimator into two steps, the classifier which produces $p(U^*, C, A, Y)$ and the corrector which estimates $p(U^* \mid U)$ and adjusts the causal estimate. To highlight the sensitivity of the corrector, we replace the classifier with a 'noisy oracle' with a fixed classification accuracy. This means that in $p(U^*, C, A, Y)$, U^* takes the same value as U with a fixed probability, regardless of all other variables. We then sample a validation dataset from $p(U, C, A, Y)$ and use it to estimate the error rate $p(U^* \mid U)$. As our validation dataset grows, we should expect our error rate estimate to converge to the true accuracy.

Figure 6-1 shows overall causal estimate error as we increase the size of the validation set. We compare the τ_{U^*} (Corrected) estimator against a naive (Uncorrected) estimator that assumes $U = U^*$. We plot both such estimators for a noisy oracle classifier with two accuracies: 70% and 90%. The Uncorrected estimator ignores the validation set entirely, so has constant error. The Corrected estimator improves with additional validation data, but can perform worse than the Uncorrected estimator.

This motivates our need for a sensitivity analysis. If we use such an estimator and wish to draw real-world conclusions from a causal estimate, we need the ability to trust the robustness and reliability of that estimate. If the theoretically-consistent Corrected estimator and the naive Uncorrected estimator disagree, what can we do? We need our sensitivity analyses to inform our degree of uncertainty in the final causal effect.

We now provide a synthetic evaluation of our three proposed sensitivity analyses.

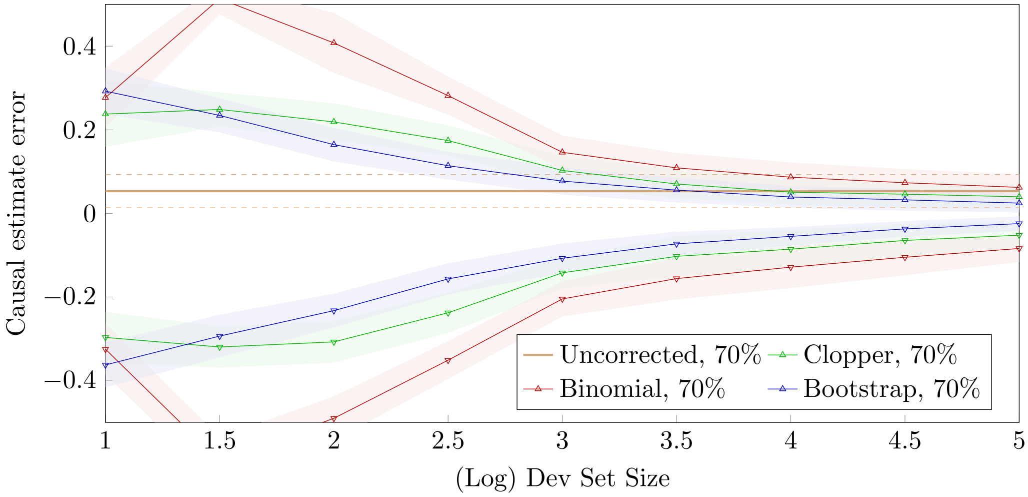


Figure 6-5. Our sensitivity analyses provide upper/lower bounds that capture the correction step's uncertainty. The true causal effect is at $y = 0$. All experiments consider data with a known classifier accuracy (70%); the validation set is only used to estimate classifier error rates. Each line and shaded bounds represent the mean/stdev calculated from 100 simulations on 10 distributions. Uncorrected estimator bounds are represented by the dashed line. In the limit our bounds converge to the truth.

Each of the proposed methods have advantages and disadvantages, which may change how they perform. Each method produces bounds on the final causal effect; we want to understand the width of those bounds and whether those bounds contain the true effect. Figure 6-4 investigates the trade-off between interval width and coverage of the true causal effect. Each analysis is run 100 times on each of ten different distributions, and we calculate the mean and standard deviation across those ten distributions.

For the bootstrap method, we take $k = 100$ bootstrap resamples and sweep over $\gamma_{\text{Bootstrap}}$ to the intervals given by truncating the empirical bootstrap distribution at different percentiles. Each percentile produces an upper and lower bound, which corresponds to a single dot in Figure 6-4, highlighting the width and coverage of those bounds. For our sampling experiments, we sweep over γ_{Binomial}, sampling $k = 100$ error rates to produce an interval for each effective validation size n'. For our interval method, we sweep over γ_{Clopper} and calculate $k^2 = 400$ causal effects for each n' to produce our interval.

It is trivial to get 100% coverage with an interval that covers all possible causal effects. It is also trivial to have a perfectly narrow interval, but with 0% coverage. We see two expected trends: as we increase the oracle classifier accuracy from 70% to 90%, all methods improved dramatically; as we increase the hyperparameter for each method, we see coverage increase but interval width also increases. With a 90% accurate classifier, we see relatively small differences between the different sensitivity analyses; at 70% accuracy, Bootstrap demonstrates the best performance. For real without known causal effects, we cannot know how interval coverage varies with its width. A robust analysis should use domain knowledge to inform such sensitivity analyses; see §6.6 for more discussion.

We now combine our proposed sensitivity analyses into the overall estimator to show how our methods can improve downstream analyses. Figure 6-5 combines our proposed methods into the analyses previously shown in Figure 6-1. Using a noisy oracle estimator with 70% accuracy we estimate the error rate on a validation set. We use our analyses to produce bounds for our causal estimate. For all validation set sizes, our upper and lower bounds contain the truth, and the interval width provides an interpretable quantification of our uncertainty. As the validation set size increases, our bounds tighten around the truth.

When conducting a downstream analysis, our hyperparameters let us control the trade-off between width and coverage. If we examine a simulation study that attempts to mirror a real-world dataset, a sensitivity analysis on the synthetic data can inform our interpretations with real data. In Chapter 7, we evaluate our measurement error methods on a wider array of synthetic data.

6.5 Causal effect modification of gender in Twitter data

We now use a real-data analysis to demonstrate how our methods can inform the reliability of a causal effect analysis in a complex domain. Social media, like EHR notes, is noisy but a valuable source of data. We focus in particular on public perception of vaccines, as such opinions are critically important to public health and have been influentially studied in social media (Dredze et al., 2016). Looking at users who tweet about vaccines, we ask: Does the perceived gender of a tweet's author affect the relationship between popularity and engagement?

From a Twitter stream spanning 2014 to 2019, we collect 21k tweets using the pro-vaccine #VaccinesWork hashtag, and 1.2k tweets with the anti-vaccine #CDCWhistleBlower. Our outcome Y is the whether a tweet receives at least one like, our treatment A is if the author has 300+ followers. The author's verification status is an always-observed confounder, C. Our U^* is the binary gender prediction of a demographics classifier (Knowles, Carroll, and Dredze, 2016), which is a noisy proxy for the perceived gender of the author U. If the effect of A on Y varies with the value of U, we say U 'modifies the effect' of A on Y (Knol and VanderWeele, 2012).

We next describe the data collection and preprocessing steps taken to set up our analysis, and include an examination of whether our choice of thresholds for making continuous variables binary affect our analyses.

6.5.1 Twitter Data Collection and Preprocessing

We collect tweets collected from the Twitter streaming API that mention vaccine-relevant keywords (e.g. "vaccine," "immunization") from November 2014 to April 2019. We select tweets containing two hashtags strongly associated with pro-vaccine (#VaccinesWork) and anti-vaccine (#CDCWhistleBlower) tweets. For the 1.8M

	Women	Men	Total
Unpopular	56.75	41.99	50.84
Popular	53.77	54.31	54.02
Total	55.47	48.05	52.29

(a) $E[Y = 1 \mid A, U^*]$; probability that #CDCWhistleBlower users receive likes.

	Women	Men	Total
Unpopular	41.96	35.87	39.17
Popular	53.79	49.09	51.49
Total	47.70	42.72	45.34

(b) $E[Y = 1 \mid A, U^*]$; probability that #VaccinesWork users receive likes.

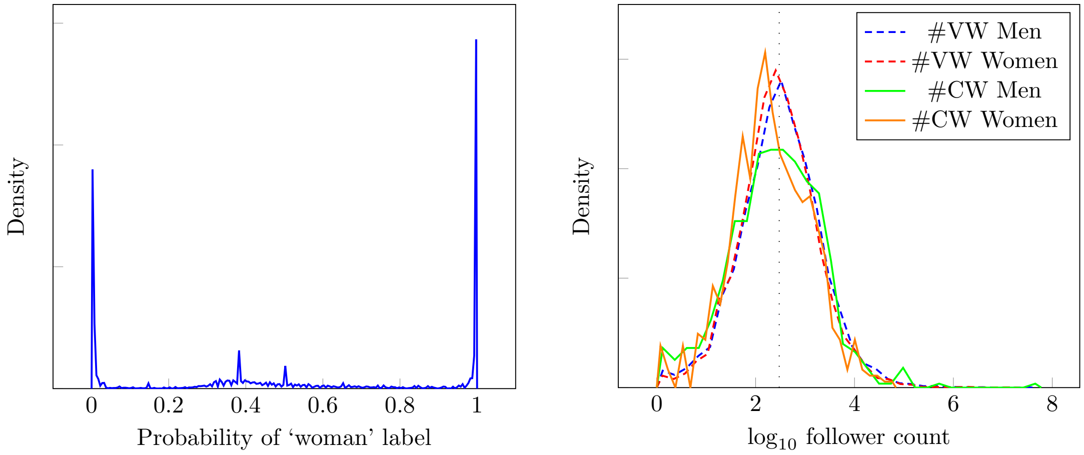
Table 6-I. The percent of tweets that receive at least one like, stratified by classified gender and popularity. Table 6-II shows the joint distribution of gender and popularity (cutoff set at 300 followers.) Each user is assigned the maximum likelihood gender label, and probabilities shown are marginalized over verification status. Datasets are created after all preprocessing steps listed in § 6.5.

tweets and retweets containing these hashtags, we re-download them using the Twitter API and remove tweets that have been deleted from the platform.[4] We recursively extract tweets from the `retweeted_status` and `quoted_status` fields and keep all unique tweets which contained one of the two hashtags. This produced 404k unique tweets for #VaccinesWork and 236k for #CDCWhistleBlower. To study general individuals on Twitter sharing vaccine information, we further filter the dataset given several criteria. First, we only consider accounts representing individuals (not organizations). We use an individual versus organization classifier (Wood-Doughty, Mahajan, and Dredze, 2018) to remove tweets that are not from individuals. [5] This removes 94k tweets, 20% of the #VaccinesWork and 13% of the #CDCWhistleBlower. Next, we remove accounts dedicated solely to the promotion of vaccination information since we are interested in general users, not vaccine-specific accounts. We remove users who posted more than ten such tweets. This yielded tweets from 21k #VaccinesWork and 1.2k #CDCWhistleBlower *users*.

Finally, we obtain perceived author gender using a gender classifier (Knowles, Carroll, and Dredze, 2016), which infers a probability that a user is a 'man' or 'woman.' We use the probability that that user is labeled as a woman by the classifier,

[4] Tweets and accounts can be deleted; this is most common with spam or bot removal.

[5] Future work could also model the measurement error in this step of our analysis.



(a) Distribution of continuous gender label.

(b) Distribution of follower-count by gender and hashtag. The vertical dotted line is at 300 followers.

Figure 6-6. Distribution of the predicted gender label and conditional distribution of follower-count given gender.

which gives us a gender label between 0 and 1 to use in our analysis. Figure 6-6a shows the distribution over the gender label probability for all users. The raw data for our treatment (follower-count) and outcome variables (likes received) are both discrete (integers) variables. Since our analysis framework assumes fitting the joint density of binary variables, we convert follower-count (A) and likes received (Y) into binary variables by binning. We remove users with fewer than 10 or more than 10k followers, and split the remainder at a cutoff of 300 followers; $A = 1$ if a tweet's author has more than 300 followers, and $A = 0$ otherwise. As the majority of tweets in our dataset receive no likes, we define $Y = 1$ if a tweet receives at least one like, and $Y = 0$ otherwise. A possible concern with any binning process is that it assumes homogeneity within each group. For example, if it were the case that all men with more than 300 followers actually had 3,000 followers but all women with more than 300 followers only had 400 followers, then any differences we attributed to gender might actually be attributable to the heterogeneity within our 'popular' treatment category. However, Figure 6-6b shows the distribution of follower-count and likes-received is fairly similar for men and women.

	Women	Men	Total
Unpopular	32.63	21.77	54.40
Popular	24.53	21.07	45.60
Total	57.16	42.84	100.00

(a) $p(A, U^*)$ distribution of 1,092 #CDCWhistleBlower users.

	Women	Men	Total
Unpopular	27.07	22.85	49.91
Popular	25.53	24.56	50.09
Total	52.60	47.40	100.00

(b) $p(A, U^*)$ distribution of 19,890 #VaccinesWork users.

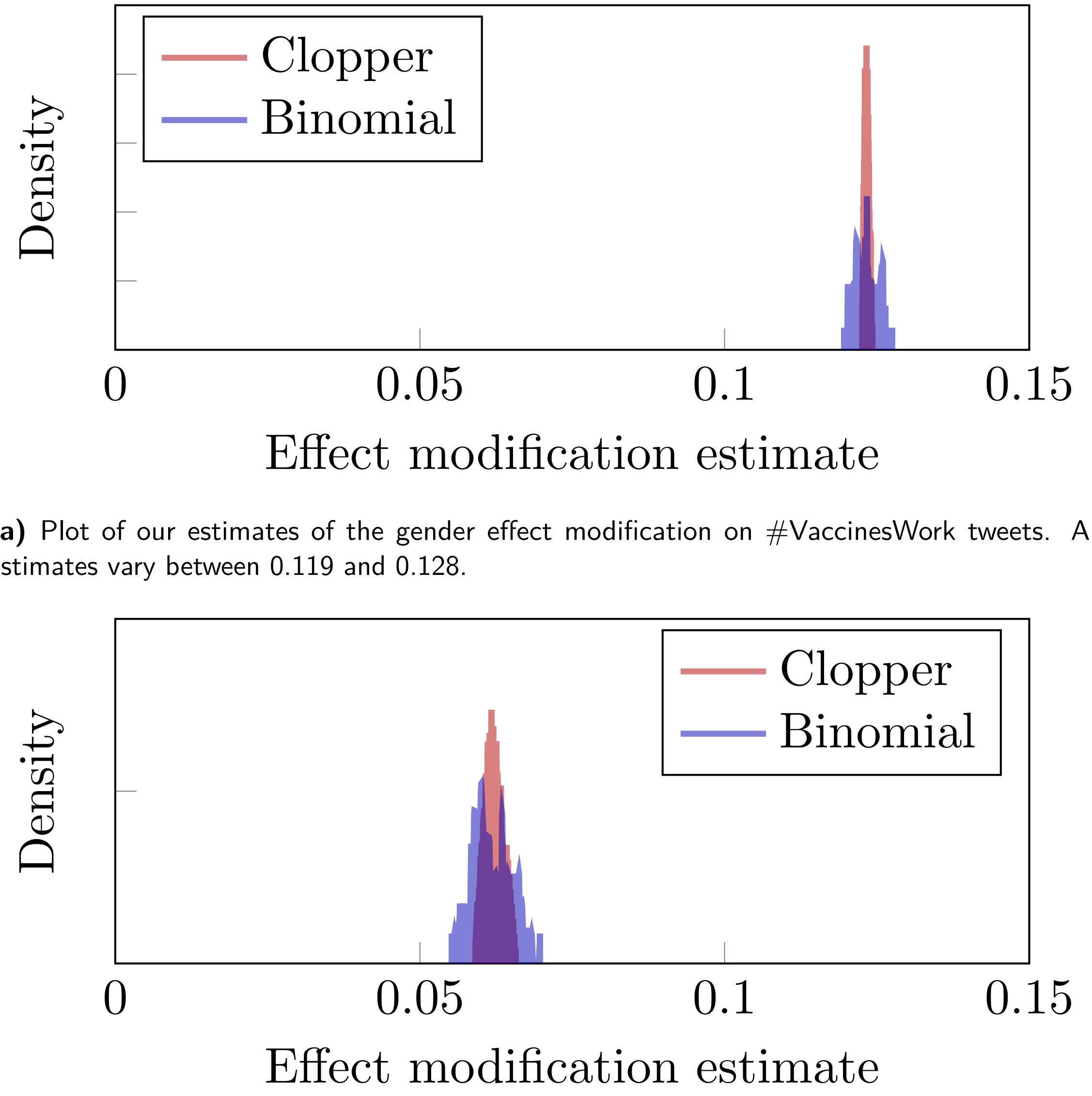
Table 6-II. Complement to Table 6-I. Tables (a) and (b) show the joint distribution of popularity and inferred gender.

Choices such as the cutoff between popular and unpopular users may unduly influence the results of our analysis. To consider the influence such choices may have, we provide additional details of our data. Figure 6-6a demonstrates that treating the gender classifier output as a binary label for the purposes of Figures 6-7a and 6-7b does not throw away too much information. Figure 6-6b demonstrates that the follower distributions conditional on (binarized) gender and hashtags are quite similar, and we should not expect any particular cutoff to conflate popularity with gender.

Tables 6-I(a-b) summarizes the data. For #VaccinesWork tweets, popular users clearly receive more engagement than unpopular users, and women receive more engagement than men. For #CDCWhistleBlower tweets, both such statements are true marginally, but tweets by unpopular women have 3% *more* engagement than those by popular women. However, these tables do not factor in the uncertainty of our gender classifier, which we now consider with our sensitivity analysis.

6.6 Sensitivity analysis of gender effect modification

We conduct a sensitivity analysis of the Twitter data that accounts for the error rate of our gender classifier to estimate the effect modification of gender. The estimand for effect modification is simply a rearrangement of terms of τ_{U*}, giving the difference between two conditional causal effects: does high-popularity increase



(a) Plot of our estimates of the gender effect modification on #VaccinesWork tweets. All estimates vary between 0.119 and 0.128.

(b) Plot of our estimates of the gender effect modification on #CDCWhistleBlower tweets. All estimates vary between 0.055 and 0.070.

Figure 6-7. Estimating τ_{U^*} using our approach on our Twitter datasets.

engagement more for men than it does for women?

The perceived gender classifier released by Knowles, Carroll, and Dredze (2016) was estimated to have error rates of $p(U^*\!=\!1|U\!=\!0)\!=\!17.0\%$ and $p(U^*\!=\!0|U\!=\!1)\!=$ 19.3% on a validation set of 52k users. As discussed in §6.3, our sensitivity analysis requires a hyperparameter choice that encodes some prior uncertainty. We choose

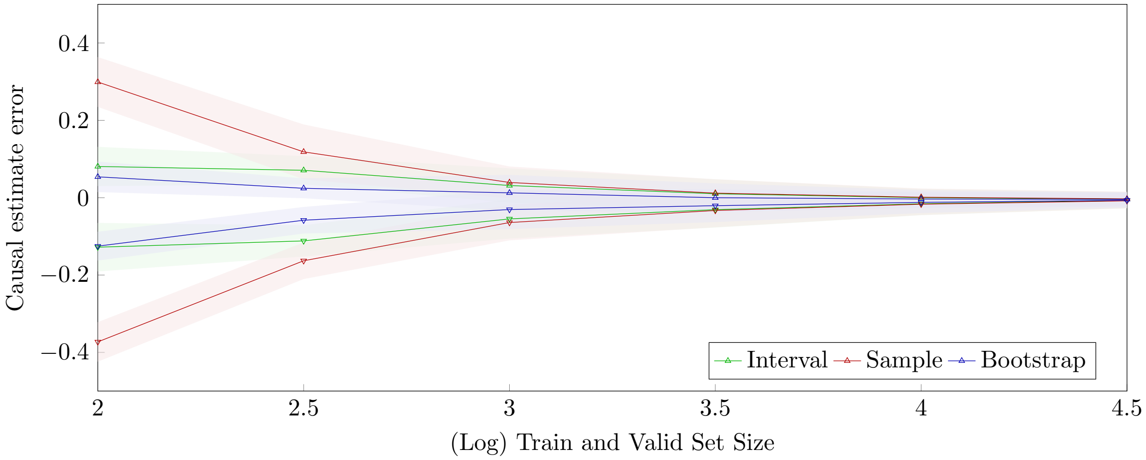


Figure 6-8. Our sensitivity analyses with a classifier trained on synthetic data, where our methods assume non-differential error. Our methods are over-confident and converge so as to not contain the true causal effect. Each line and its bounds represent the mean and standard deviation calculated from 100 simulations on ten different distributions.

$\gamma_{\text{Bootstrap}}$ and γ_{Clopper} following the synthetic results in § 6.4, which equates to an effective validation size of 900 users. While the public nature of Twitter data means we could in fact re-collect this dataset and use our bootstrap method, we do not for the purposes of this analysis. In many real-world cases, we could not have full access to the validation set.

A histogram of the outputs of both methods are shown in Figures 6-7a and 6-7b. Interpreting these plots relies upon connecting our domain knowledge to our methodology. How should an analyst choose γ for these sensitivity analyses? If our binomial sampling method produces an outlier estimate of -0.5, can we safely disregard it? Answering such questions in any applied analysis relies on domain knowledge. Suppose previous work suggests that a particular classifier's error rate varies greatly by domains, then we may want to choose a conservatively large hyperparameters for our analyses. In contrast, if we are confident that past work has established the true classifier accuracy, we may be willing to trust tighter bounds. Our Twitter data and sensitivity analyses give very tight bounds, but we could further validate our approach by collecting another validation set on our

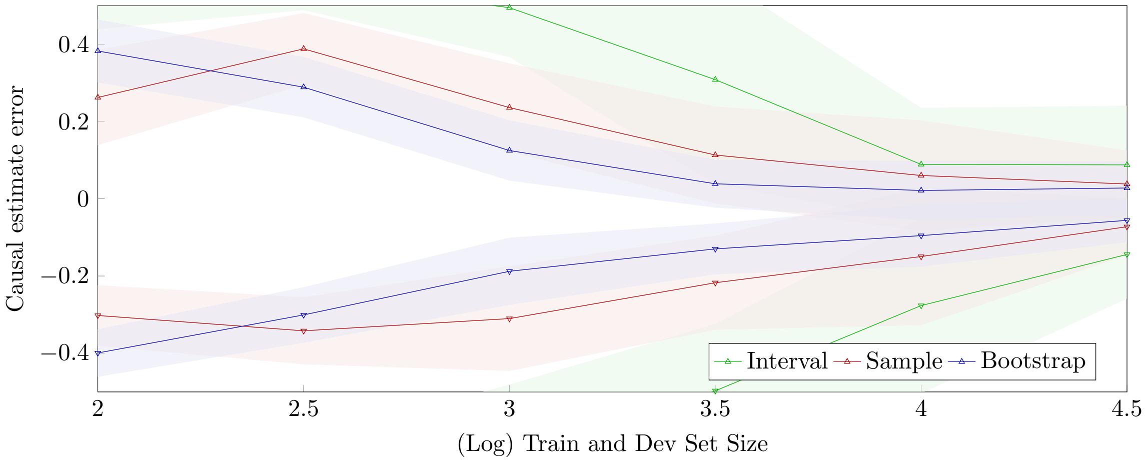


Figure 6-9. Our sensitivity analyses with a classifier trained on synthetic data, where our methods assume differential error. Our methods tend to be under-confident, with the interval method providing uninformative bounds for many validation set sizes. Each line and its bounds represent the mean and standard deviation calculated from 100 simulations on ten different distributions.

vaccine-specific data.

While parametrizing uncertainty in our analyses is certainly helpful, it does not obviate the need to draw on domain expertise for interpretability. Past public health research has found conflicting results on whether gender plays a significant role in vaccine skepticism or decision-making (Nagata et al., 2013). However, the existence of vaccination gender disparities (Abat and Raoult, 2018) and the need to effectively communicate to diverse audiences (Chen and Dredze, 2018; Nan, 2012) necessitates further study of gender differences in vaccination trends.

6.6.1 Differential measurement error

All of our analyses as presented rely upon an assumption that the measurement error is *non-differential*, meaning that the error rate is independent of A, C, Y. If this is not the case, our measurement error correction becomes more difficult; we must model the error rate as it depends on those variables. Much applied work on measurement error assumes non-differential error, as causal effects can be unidentified without such an assumption (Butler et al., 1987; Carroll et al., 2006).

Each of our sensitivity analyses need to adapt to differential measurement error in different ways. Full differential error means that for a C confounder of dimension k, we need to estimate 2^{k+2} error rates even in the fully-binary case. The Bootstrap analysis is identical, but may require many more samples to cover the variability of differential error. The Binomial sampling approach can simply sample these many error rates, but again it may take many samples to cover the space of possible causal effects. Our Clopper-Pearson approach becomes quickly intractable to compute as the dimension of the causal DAG increases. If we want to consider all combinations of interval endpoints for a DAG with k variables, we must calculate 2^{2^k} endpoint combinations. This is 65k calculations for 4 variables, and many billions for 5 variables. Future work could explore better ways to balance coverage, interval width, and computational tractability in the differential error setting.

Comparing Figures 6-8 and 6-9 shows the how our estimates change when we assume or do not assume differential error for a trained classifier. When our methods try to account for the need to estimate additional error rates, they converge more slowly, with the Clopper-Pearson interval approach providing uninformative bounds when the validation set is small. In Chapter 9, we further explore how the assumption of non-differential error affects an application of our measurement error method to a different Twitter dataset.

6.6.2 Analysis limitations

Our Twitter case study demonstrated the efficacy of our sensitivity analysis, but we caution against drawing conclusions about users' behavior. We estimate that the effect modification is robust to bias from the gender classifier, but not other assumptions in our analysis: gender is not binary and we do not differentiate between perceived and self-identified gender (Hamidi, Scheuerman, and Branham, 2018; Frohard-Dourlent et al., 2017; Butler, 1988). While conceptualizing the inherent uncertainty of gender prediction in a measurement error framework is better than taking its predictions as truth, but could still cause harm if used to misgender individual users (Keyes, 2018). Second, data processing details can change the outcome, e.g. removing retweets and prolific users or how we model follower-counts and like-responses. For example, follower-count and like-response are only noisy proxies of an underlying concept of user and tweet status. We may have added new bias by artificially binning these two variables into binary categories. While we could bin these into a larger number of discrete variables, our matrix adjustment approach needs additional assumptions for real-valued distributions (Miao, Geng, and Tchetgen, 2018). Third, we restrict our analysis to a three-variable causal model, an over-simplification of social media behaviors. There are likely other unobserved confounders, such as malicious foreign actors (Broniatowski et al., 2018). Additionally, if the classifier error rate is correlated with user popularity, our correction step may introduce new biases. Finally, we treat tweets as independent samples, ignoring network effects of the platform which may correlate user behaviors.

Many of these implicit assumptions are unrealistic and likely introduce some bias into our conclusions, though such assumptions are ubiquitous in social media analyses. We accept those assumptions to focus on explicitly addressing measurement error induced by an imperfect classifier.

6.7 Conclusions

We have presented a new measurement error formulation that provides a means for incorporating estimated errors of ML classifiers into a causal analysis framework. Our formulation provides a more robust framework for reaching causal conclusions using classifier predictions. This work creates new opportunities for the analysis of causal factors in a variety of domains, including in computational social sciences that rely on analyses of high-dimensional data such as text. We highlight our methods on synthetic and real that highlights the interpretability provided by our sensitivity analyses.

There are several directions along which further work can extend our framework. The sensitivity analysis could be extended to work with non-binary classifications, high-dimensional C or U vectors, differential measurement error, or interference. Each of these research directions would expand the real-world applications of our methods, and would benefit from our contributions.

6.8 Broader impact

We have introduced three sensitivity analyses for understanding the uncertainty of methods that rely on machine learning model predictions to estimate causal effects.

Many existing ML applications need causal reasoning to inform possible interventions (Stern and Price, 2019). There are many examples of predictive models performing well in an experimental setting but poorly in real-world settings (Subbaswamy, Schulam, and Saria, 2019). Our methods provide a generally-applicable method for quantifying the uncertainty that is introduced into causal analyses that depend upon a trained machine learning model.

A primary implication of our methods is that they enable causal analyses that rely on trained machine learning models. ML methods trained on high-dimensional features in large datasets can extract information at a scale that cannot be matched by human experts. If such models, e.g. in medical image analysis (Shen, Wu, and Suk, 2017) or in processing EHR notes (Rajkomar et al., 2018), can be incorporated into a principled causal framework, many new analyses will be possible.

Despite this promise, we must be wary of overconfidence in ML methods that perform well in experiments but poorly in practice. Our methods make assumptions about the relationship between a machine learning classifier's models and the variables relevant to the ultimate causal analysis. If a causal analysis informs healthcare policy, it must address broader concerns of fairness and transparency (Nabi, Malinsky, and Shpitser, 2019). If a classifier's errors disproportionately affect demographic groups, the resulting biases may go unnoticed unless an effort is made to look for them.

There are several initiatives that could guide this line of work towards better societal impacts. From a technical side, we could benefit from new connections between theoretical convergence rates of estimators and the empirical bounds we focus on in this work. In practical settings, we need ways to calibrate methods to optimize for societal goals. If a deployed model informs medical interventions, we need clearly-established guidelines for evaluating the costs of over- or under-confidence of bounds on a causal estimate. Such initiatives require the collaboration of ML researchers and the domain experts who may hope to apply their models.

Chapter 7

Generating Synthetic Text Data to Evaluate Causal Inference Methods

7.1 Introduction

We have thus far explored approaches for combining ML and NLP methods into causal analyses. While we applied our methods to a real-world dataset of Twitter posts in §6.5 and 6.6, we have mostly demonstrated efficacy of our methods using synthetic data. In §2.3, we discussed the role of synthetic data for evaluation in the field of causal inference. Because causal methods rely on untestable assumptions about the data-generating process (DGP) that produced the data, evaluating the empirical behavior of a method requires complete knowledge of the DGP. Such a setting is only possible with synthetic data, because of the inherent noise and uncertainty of any real-world dataset. This chapter introduces more complex synthetic text DGPs and uses them to evaluate causal methods. The work in this chapter is under review, and is available as a preprint at Wood-Doughty, Shpitser, and Dredze (2021).

Text data provides a particularly difficult domain for evaluating causal methods because it requires modeling causal relationships between structured variables and text: "what caused the author to write the text this way?" While there is plentiful text data for training predictive models, we cannot directly measure the underlying processes that humans use to produce or adapt their language in complex domains. Synthetic DGPs need to balance 'realism and control' (Wendling et al., 2018): the goal of producing realistic text data against the competing goal of completely specifying the causal effects that produce the text. Past methods evaluated on synthetic data have only satisfied one such goal, either by producing particularly unrealistic text with known effects (Yao et al., 2019; Wood-Doughty, Shpitser, and Dredze, 2018; Johansson, Shalit, and Sontag, 2016) or using real-world text without a fully-specified DGP (Veitch, Sridhar, and Blei, 2020; Mozer et al., 2018; Weld et al., 2020).

In this chapter, we introduce a synthetic framework for evaluating causal methods that incorporate text data, exploring desiderata of synthetic text DGPs and tradeoffs between competing goals. We introduce two nontrivial synthetic DGPs, one which samples a bag-of-words from an Latent Dirichlet Allocation (LDA) topic model, and another which samples full sentences from GPT-2 (Blei, Ng, and Jordan, 2003; Radford et al., 2019). These two underlying generative models allow us to test how causal methods perform when their assumptions are violated (e.g. whether word order matters) (Wallach, 2006). We use our framework to compare four causal methods that rely on text, addressing a known gap in empirical evaluation of such methods (Keith, Jensen, and O'Connor, 2020). We explore how existing methods' empirical performance depends on their assumptions and show that when the causal estimator depends on a text classifier model, better classification accuracy of that classifier does not necessarily imply better causal estimates. We release our code and synthetic datasets to facilitate further development and evaluation of causal methods for language data.

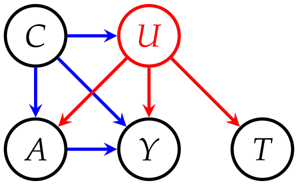


Figure 7-1. The causal DAG we consider. A is our treatment, Y is our outcome, C and U are confounders, and T is the raw text which is influenced by U. The counterfactual $p(Y(a))$ cannot be non-parametrically identified from $p(C, A, Y)$ alone due to unobserved confounding from U. Methods may make parametric assumptions on the relationship between T and U in order to estimate the causal effect, or assume knowledge of $p(U|T)$. We parameterize $p(T|U)$ with text generation models in § 7.2. We discuss the limitations of this DAG model and extensions to other models § 7.7.1.

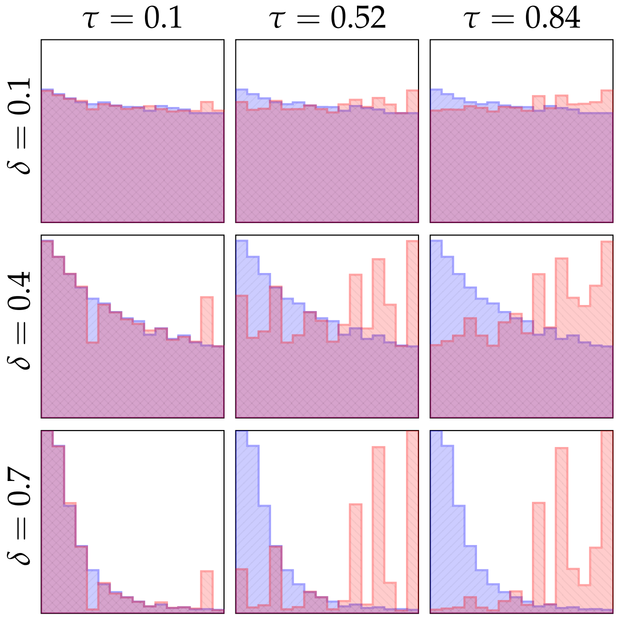


Figure 7-2. Causal effect strengths and Trivial text generation. Blue and red bars correspond to $U = 0$ and $U = 1$ respectively. As τ increases, the ranked preferences between $U = 0$ and $U = 1$ diverge. As δ increases, the distribution is puts more weight on the ranked preferences. The x-axis indexes the 16 words in the vocabulary, with each bar indicating the probability that a word shows up at least once in a 16 word sequence. When $\tau = 0.1$ and $\delta = 0.1$, the distributions are close to uniform and almost entirely overlap. In all plots the $\tilde{V}_0$ order matches the x-axis order. As τ increases, the $\tilde{V}_1$ order diverges. As δ increases, both distributions become more concentrated on higher-ranked words.

7.2 Causal Effects in Text Generation

We return again to our motivating clinical example. In Figure 7-1, the treatment A is a binary measure of vitamin D deficiency and the outcome Y is the onset of preeclampsia. C and U, age above 35 years and socioeconomic status (SES), are confounders that influence both A and Y. We define $Y(1)$ as a counterfactual random variable representing "preeclampsia status if a patient, possibly contrary to fact, had a vitamin D deficiency." This counterfactual variable's distribution can be identified as:

$$p(Y(a)) = \sum_{C,U} p\left(Y|A=a,C,U\right) p\left(C,U\right) \tag{7.1}$$

We'll again suppose SES is not directly recorded in structured (i.e. tabular) records, but can be inferred from physician's text notes about the patient. While for simplicity we will assume A, C, U, and Y are binary variables, we let T denote the raw text of the clinical notes. The edge from U to T assumes that the clinician's note-taking is influenced by the underlying U value; the lack of edges between $\{A, C, Y\}$ and T reflects a simplifying assumption. The relationship between U and T is complex and essential to the methods we will consider.

Recent work in natural language generation has introduced language models with enormous empirical gains in perplexity and according to human judgments (Radford et al., 2019; Hashimoto, Zhang, and Liang, 2019; Brown et al., 2020). Language models generate text sequences token by token, where token i is sampled conditional on the previous $i-1$ tokens and the first token is often sampled conditional on some initial context. These existing methods, however, do not produce datasets with known causal effects on text itself; we must first produce a formal definition for the causal effect of a structured variable on the text generation process. In our clinical example, such an effect represents how a doctor's notes would have changed had a patient, *counterfactually*, been of high SES. By controlling the effect of U on T in

τ_{word}	0.1	0.52	0.84
δ_{word}			
0.1	0.54	0.60	0.62
0.4	0.75	0.97	0.98
0.7	0.78	1.00	1.00

.

Table 7-I. $p\left(U|T\right)$ classifier accuracy for the nine examples of our trivial DGP. When both τ and δ are small, accuracy is near random chance. If one of the two parameters is large, accuracy improves; if both are large, accuracy nears 100%

Figure 7-1, we can evaluate how causal methods perform when their assumptions are met or violated.

We want our marginal $p(T)$ to conform to a language model that generates text according to a learned distribution, but want to parameterize $p\left(T|U\right)$ such that we can force the generation to smoothly diverge from its learned distribution to depend on U. We want a causal effect of U on T to make some words or topics more likely and others less so. That is, texts generated when $U = 1$ should be quantitatively and qualitatively different from texts when $U = 0$. We will introduce τ and δ as hyperparameters that control our causal effects. Intuitively, τ controls rankings over the vocabulary; the larger τ is, the more the ranked preference for $U = 0$ differs from that of $U = 1$. We can conceptualize δ as controlling how much the model indulges its preference; the larger δ is, the more likely $p\left(T|U = u\right)$ samples according to these ranked preferences rather than from the pre-trained language model distribution.

To formalize this, let $V = \{x_1, \ldots, x_N\}$ be a vocabulary of N words. The learned language model provides an initial distribution $p(V)$ and uses it to generate the sequences that comprise $p(T)$. Let $\tilde{V}$ be an ordering over V For a binary U, we choose two orderings, $\tilde{V}_{u=1}$ and $\tilde{V}_{u=0}$. Our τ parameter controls the correlation between those two orderings. When $\tau = 0$, the orderings are the same; when $\tau = 1$, they are exact reversals of each other. For a given τ, we sample these orderings such that their Kendall Tau correlation is approximately $1 - 2\tau$.

For a given $\tilde{V}$ and our choice of δ, we will construct a new distribution over the vocabulary. Define $f_{\tilde{V}}(x_i)$ as a mapping from a vocabulary item x_i to the position of that item in the ordering. If x_{42} is the first item in the $\tilde{V}$ ordering, then $f_{\tilde{V}}(x_{42}) = 1$. Now define a 'modified Zipfian distribution' as $p(x_i) \propto f_{\tilde{V}}(x_i)^{-\delta/(1-\delta)}$, When $\delta = 0$, this is simply a uniform distribution over the vocabulary; when $\delta = 1$, it is a point mass on the first item in its preference.[1]

Now, given our language model's learned $p(V)$, we construct a new distribution:

$$p'(x_i; \tilde{V}, \delta) \propto p(x_i) \otimes f_{\tilde{V}}(x_i)^{-(\delta/1+\delta)} \tag{7.2}$$

where $\otimes$ indicates element-wise multiplication. The distribution p' represents an average between the initial $p(V)$ and the modified Zipfian defined by $\tilde{V}$ and δ. We define a function h which takes in an initial text generation distribution $p(V)$, and values for τ and δ and returns new distributions p'_u following Eq. (7.2). We write this as:

$$h : (p(V), \tau, \delta) \rightarrow \{p'_0(V; \tilde{V}_0, \delta), p'_1(V; \tilde{V}_1, \delta)\} \tag{7.3}$$

Both τ and δ live in the $[0, 1]$ domain. We can conceptualize τ as controlling the 'preference' over words in the vocabulary and δ as controlling the 'strength' of that preference. If either hyperparameter is 0, the structured variable U has no effect on the text generation. If $\tau = 0$ then $\tilde{V}_1 = \tilde{V}_0$; while δ will change the word probabilities, it will change them equally for either value of U. Similarly, if δ is 0, then no matter how different $\tilde{V}_1$ is from $\tilde{V}_0$, $h(p(V), \tau, 0)$ ignores those preferences and returns the language model's learned $p(x_i)$.

Figure 7-2 shows how δ and τ control a trivial text generation model. We sample nine datasets of 10k sequences of 16 tokens. Our initial $p(V)$ distribution is simply

[1] For any value of δ, we will normalize p to be a distribution with probabilities in $[1e^{-10}, 1 - 1e^{-10}]$.

	δ_{word}	0.1	0.2	0.2	0.2	0.5	0.5
	τ_{word}	0.1	.005	0.05	0.1	.005	0.05
δ_{topic}	τ_{topic}						
0.1	0.1	0.56	0.55	0.58	0.64	0.63	0.94
0.2	0.005	0.56	0.54	0.59	0.64	0.63	0.94
0.2	0.05	0.63	0.62	0.65	0.70	0.69	0.95
0.2	0.1	0.65	0.65	0.67	0.72	0.73	0.95
0.5	0.005	0.65	0.64	0.68	0.73	0.71	0.96
0.5	0.05	0.87	0.86	0.88	0.89	0.91	0.99
0.5	0.1	0.97	0.97	0.97	0.97	0.98	0.99

Table 7-II. $p\,(U|T)$ classification accuracy for LDA text. Increasing τ and δ values lead to increased classification accuracy, with exceptions when δ increases but τ decreases. If either the topic or word effects are particularly large, classification accuracy exceeds 90%; when both are large, it quickly approaches 100%.

uniform over the vocabulary of 16 tokens. Each cell in the figure shows how the Trivial $p\,(T|U)$ distributions change as we vary δ and τ. When δ is large but τ is small, some words are much more likely than others, but the two distributions only differ on a single word. When τ is large but δ is small, the distributions differ by a small amount on many words.

If we want to explore how causal methods perform in Figure 7-1, we can control the $p\,(T|U)$ distribution with δ and τ. As we turn to more complicated $p(V)$ distributions, we want a better way to interpret the text generated with a given choice of these hyperparameters.

Our approach differs from past (semi-)synthetic text datasets for causal evaluation. In Wood-Doughty, Shpitser, and Dredze (2018) (§5.3), we sampled synthetic 'texts' in a bag-of-words manner similar to our Trivial distribution above, except without the ability to control the strength of the $p\,(T|U)$ relationship. Veitch, Sridhar, and Blei (2020) used real text from Reddit or academic papers and sampled synthetic outcomes conditional on metadata related to each text, but without the ability to measure or specify the causal relationship between the text and its metadata. Weld et al. (2020) generate semi-synthetic data by inserting template-based posts into the actual post history of a social media user. These synthetic interventions are discrete, however; there is no way to specify a real-valued causal effect and manipulate it arbitrarily. The flexibility of our approach allows us to explore how methods perform as we vary the causal effect on the text.

7.2.1 Classification Accuracy and δ, τ

The δ and τ hyperparameters completely control the effect of the structured variables on the text, but are not particularly interpretable. How do we know if particular δ or τ values are realistic? What values best mimic a real clinical notes DGP?

Rather than adapt our hyperparameters to a specific natural language domain, we will use text classification accuracy as a lens that can be equally applied to both synthetic and real-world text. Given a synthetic dataset, we will train a classifier with T as the features and U as the labels. Considering the accuracy of such a classifier will let us compare a synthetic dataset to a real dataset; past work has extensively considered the task of classifying clinical concepts from unstructured text (Liu, Zhang, and Razavian, 2018; Meystre et al., 2008; Afzal et al., 2018; Savova et al., 2010). A synthetic dataset in which a text classifier achieves 99% accuracy is unrealistic, implying δ and τ are too large. Similarly, if δ and τ are too small, a $p\left(U|T\right)$ classifier will be no better than chance.

Table 7-I shows binary classification accuracy of a simple bag-of-words model trained on the datasets from Figure 7-2. We use a train/dev/test split of 8k/1k/1k sequences for this and all subsequent text classification experiments. Accuracy improves above random chance as either δ or τ increase, and quickly maxes out when both are large. Classification accuracy on this task provides a useful way to abstract away the underlying DGP as we introduce more complicated synthetic datasets.

δ_{word}	0.0	0.2	0.2	0.2	0.5	0.5	0.5	0.5	0.5	0.7	0.7
τ_{word}	0.00	.025	.025	0.15	0.05	0.05	0.05	0.15	0.15	0.15	0.15
δ_{topic}	0.7	0.7	0.7	0.7	0.5	0.7	0.7	0.7	0.9	0.7	0.9
τ_{topic}	0.45	0.15	0.45	0.15	0.05	0.05	0.45	0.05	0.15	0.15	0.15
Accuracy	0.62	0.54	0.64	0.62	0.65	0.64	0.70	0.85	0.85	0.78	0.79

Table 7-III. $p\,(U|T)$ classification accuracy for GPT-2 text. Accuracy is much lower than on trivial or LDA data. Increasing τ and δ values generally leads to increased classification accuracy, but this is not monotonic. When increasing $(\delta_{\text{word}}, \tau_{\text{word}})$ from (0.5, 0.15) to (0.7, 0.15) and reducing $(\delta_{\text{template}}, \tau_{\text{template}})$ from (0.9, 0.15) to (0.7, 0.15) we see a notable decrease in accuracy even when $(\delta_{\text{template}}, \tau_{\text{template}})$ returns to (0.9, 0.15). This is because GPT-2 word and template effects can conflict; because the language model tries to maintain grammatical structure, certain templates make it unlikely to sample certain words.

7.2.2 LDA with Causal Effects

For a slightly more complicated synthetic DGP, we consider Latent Dirichlet Analysis (LDA), one of the most widely-used models of text (Blei, Ng, and Jordan, 2003). It provides a generative model of text that clusters the distribution over the vocabulary into a distribution over topics. While the LDA model ignores word order, so each sampled word drawn from the trained model is independent. This results in generated texts that have no grammatical structure. We train an LDA model on a set of 250,000 documents which was released as part of the training data for GPT-2 (Radford et al., 2019).

To define a $p\,(T|U)$ distribution that uses LDA, we will define causal effects for both the words and the topics. Let V_{word} be the word vocabulary and V_{topic} be the set of learned topics. Then $p_{\text{LDA}}(V_{\text{topic}})$ is LDA's learned baseline distribution over the topics, and $p_{\text{LDA}}(V_{\text{word}} \mid t \in V_{\text{topic}})$ is the learned distribution over words for topic t.

We introduce causal effects with h from (7.3). To sample a word from our modified LDA model when $U = u$, we first sample a topic t from $h(p_{\text{LDA}}(V_{\text{topic}}), \tau_{\text{topic}}, \delta_{\text{topic}})$. Then, instead of sampling from the original LDA

δ_{w}	The child was known for ...
0.0	his role in the very real Peter Pan film that skyrocketed
0.1	his role in the flamboyant sleuth Jackie Turner's hit
0.15	his German business, and her books were sold in Bavaria
0.25	her ability to play, run and shoot gags involving giant
0.4	her ability to see. She began training one more spring
0.45	her ability to see in one eye; her ability conquer magic
0.5	her ability to disown her magic ability and her identify
0.6	her ability one ability her magic ability her magic

Figure 7-3. DistilGPT-2 generation when we fix the random seed, template, and $\tilde{V}_{\text{word}}$ but vary δ_{word}. We construct $\tilde{V}_{\text{word}}$ so the most-preferred words are *her*, *magic*, and *ability*. The model switches from *his* to *her* pronouns as δ increases. As δ further increases, sentence fluency decreases.

distribution, $p_{\text{LDA}}(V_{\text{word}} \mid t)$, we sample from $h(p_{\text{LDA}}(V_{\text{word}} \mid t), \tau_{\text{word}}, \delta_{\text{word}})$.

How do these τ and δ hyperparameters control the generated text? Table 7-II shows text classification results. We see that in general, larger τ and δ lead to higher accuracy, yet there are exceptions. Within a given row or column, when δ increases but τ decreases, we see a brief drop in accuracy. We can conceptualize this with the plots in Figure 7-2; as δ increases the effect of U on T grows and the word distribution changes from its learned distribution, but as τ decreases it decreases the difference between the $U = 0$ and $U = 1$ 'preference' distributions. If we plot τ_{word} against δ_{word} and hold topic effects constant, we would see that accuracy monotonically increases as either word effect hyperparameter increases.

7.2.3 GPT-2 with Causal Effects

One of the primary drawbacks of LDA is that it only models topic, and has no sense of word order or syntax. Therefore, we consider a more complex DGP by extending our synthetic data framework to more complicated neural models that are widely used for text generation.

GPT-2 is a large neural language model that has improved the state-of-the-art on several benchmark evaluations (Radford et al., 2019). It uses 1.5-billion parameters to encode a context sentence into an internal representation and then uses that representation to predict a distribution over the next word in the sentence. Once a word has been sampled from that distribution, it is fed back into the model as additional context, and the sampling process continues. Word-order is thus intrinsic to the sentences generated by GPT-2. To save computation time, we use a smaller 82M parameter DistilGPT-2 model (Sanh et al., 2019). We discuss extensions to more recent neural language models in § 7.7.2.

While the model can take as input an arbitrary context sentence or phrase, we follow Sheng et al. (2019) and use a set of simple templates to seed the generation of the GPT-2 model. The templates are a combination of a subject (e.g. 'the person') and the beginning of a verb phrase (e.g. 'was known for'). Our V_{template} has 60 templates. We treat GPT-2 as a black-box which inputs a distribution over these 60 templates and outputs a distribution over the words in the vocabulary. As with our LDA model, we will introduce causal effects which influence these inputs and outputs, but otherwise leave the model untouched.

We start with an initial uniform distribution over the 60 templates. From an initially uniform $p_{\text{GPT-2}}(V_{\text{template}})$, we sample a template t from $h(p_{\text{GPT-2}}(V_{\text{template}}), \tau_{\text{template}}, \delta_{\text{template}})$. Then, we feed that template into the GPT-2 model as context, and it produces a distribution over words: $p_{\text{GPT-2}}(V_{\text{word}} \mid t)$. We then sample the first word from $h(p_{\text{GPT-2}}(V_{\text{word}} \mid t), \tau_{\text{word}}, \delta_{\text{word}})$. We then feed that sampled word, w_1, back into the GPT-2 model and sample the next word, conditioning on both the template and the first sampled word, from $h(p_{\text{GPT-2}}(V_{\text{word}} \mid w_1, t), \tau_{\text{word}}, \delta_{\text{word}})$.

Table 7-III shows how text classification accuracy changes as our τ and δ parameters change. As in Table 7-II, larger τ and δ values lead to better classification accuracy, with some exceptions. Every $p\,(U|T)$ accuracy drop on LDA data in Table 7-II co-occurred with a drop in a τ or δ effect. With GPT-2, we see one case where causal effects strictly increase but text classification accuracy decreases. When $\tau_{\text{word}} = 0.15$ and $\delta_{\text{template}} = 0.7$ and δ_{word} increases from 0.5 to 0.7 and τ_{template} increases from 0.05 to 0.15, text classification accuracy drops from 85% to 78%. A likely explanation for this is that the GPT-2 templates do not affect individual word probabilities, but provide context that affects the entire sequence. The template fragment 'worked as a' likely increases occupation-related words, where the fragment 'was known for' may not. These non-monotonic effects may complicate the ability of our simple bag-of-words model to differentiate the two distributions.

We also see that while the formal definitions of τ and δ are the same between LDA and GPT-2, the values must be much larger for the classifier to reach 90% test set accuracy. This reflects the mismatch between the bag-of-words assumption of our text classifier and the more complex text sequences of GPT-2.

As GPT-2 produces more fluent text than LDA, we can also visualize the effect of δ_{word} by slightly varying its value while repeatedly sampling from the model. Figure 7-3 shows how the generation changes when we fix the template and GPT-2's random seeds, and increase δ_{word} for a given $\tilde{V}_{\text{word}}$ preference.

7.3 Causal Methods with Text

We have introduced a framework for producing datasets where we can provide fine-grained control over how structured variables influence the text. We can use this framework to evaluate existing methods for estimating causal effects with text data. We will first provide an overview of four such approaches, and then use our framework to conduct a range of simulation studies that explore how well these methods perform as we vary the $p(T|U)$ relationship.

Each method relies on sample-splitting for robust inference (Chernozhukov et al., 2016; Anderson and Magruder, 2017). In particular, we will split dataset in half, use one split to train and validate a simple bag-of-words logistic regression model, and then use the other split to estimate our causal effect. Then we will flip the splits to get a second effect estimate on the first split, and then report the average of the two. As we only use simple models for these evaluations, we leave full implementation and training details to our released code.

7.3.1 Matching with Text

Matching is a popular causal method (Stuart, 2010), which has been recently applied to text datasets (Roberts, Stewart, and Nielsen, 2018; Mozer et al., 2018; Yao et al., 2019; Wang and Culotta, 2019). Matching adjusts for confounding by estimating the causal effect among patients who are similar, where similarity can be defined by confounders or by their propensity to have received the treatment. We consider two types of text matching: propensity score matching and representation matching.

If U were observed in Figure 7-1, valid propensity score matching would proceed by learning a model for $p\left(A|C,U\right)$ and matching patients based on the estimated propensity. With U unobserved, we will instead match on a propensity score modeled as $p\left(A|C,T\right)$. This method will be biased in general because matching requires the true propensity score. However, if there exists a function that maps our estimated $p\left(A|C,T\right)$ to the true propensity $p\left(A|C,U\right)$, this approach can be unbiased. To implement this method, we model the propensity $p\left(A|C,T\right)$ with a bag-of-words classifier. We then match on the estimated propensity using full matching as implemented in the `R` package `optmatch`, following Mozer et al. (2018).

Representation matching attempts to adjust for confounding by matching patients on their covariates (C, U) and then taking $p\left(Y|A\right)$ within each matched group as an unbiased estimate of $p\left(Y(a)\right)$. As U is unobserved, we can instead match on both C and a learned representation of T. The intuition is that if two patients have similar T

representations, they are likely to have the same value of U. However, this method will be biased in general if two values U can produce the same T representation. For our experiments, we use an LDA topic model representation of T and perform full matching using cosine similarity, following (Mozer et al., 2018).

7.3.2 Conditioning on Text

Rather than matching on the propensity score, we can directly use it in an inverse propensity weighting (IPW) model (Rosenbaum and Rubin, 1983). This approach reweighs the observed data by the inverse of the true propensity model; if the true propensity $p(A|C,U)$ is used, this is a consistent estimator for Eq. (7.1). When we replace $p(A|C,U)$ with $p(A|C,T)$, our estimates are no longer guaranteed to converge to the ground truth. Instead, we must assume that if the effect of U on T is strong, then the learned propensity score will suffice to reweigh the examples. This approach is similar to the bag-of-words method used by Veitch, Sridhar, and Blei (2020). Initial experiments, we found that more powerful neural models performed poorly on our datasets of only 10k examples. This method follows other work in controlling for high-dimensional confounders (Hill, Weiss, and Zhai, 2011; McCaffrey, Ridgeway, and Morral, 2004; Low, Gallego, and Shah, 2016).

Our implementation again models $p(A|C,T)$ as a bag-of-words classifier. We truncate propensity weights and report the mean of 100 bootstrap estimates (Lee, Lessler, and Stuart, 2011).

7.3.3 Imputing with Text

Our fourth causal method is the measurement error approach developed throughout this thesis in Chapters 2, 5, and 6. We assume access to a text classifier model $p(U|T)$ that can impute U^*, a noisy proxy for the true U. The method uses the classifier

	Representation			Propensity			IPW			Measurement		
τ_{word}	0.1	0.52	0.84	0.1	0.52	0.84	0.1	0.52	0.84	0.1	0.52	0.84
δ_{word}												
0.1	0.19	0.19	0.19	0.17	0.16	0.16	0.19	0.18	0.18	0.11	0.03	0.03
0.4	0.18	0.03	0.02	0.14	0.05	0.05	0.17	0.06	0.04	0.03	0.01	0.00
0.7	0.16	0.01	0.01	0.12	0.05	0.05	0.14	0.01	0.01	0.01	0.00	0.00

Table 7-IV. Causal estimation error for the four estimation methods on our trivial DGP. All methods approach zero error as δ and τ values increase.

and an estimate of the error rate of the classifier to correct for the bias induced by the imperfect classifications (Pearl, 2010). Importantly, this approach requires more information than text matching or IPW, as we must have access to either a pre-trained classifier with known error rate or enough labeled data $p(U, T)$ to train a classifier. In many cases, such labeled data may be difficult or impossible to collect. We train a logistic regression classifier for $p\left(U|T\right)$, using half the training split to train the classifier, and the other half to estimate the classifier's error rates.

7.4 Evaluating Causal Methods with Text

We have thus far introduced a framework for producing synthetic text datasets and discussed four past methods that have been proposed for estimating causal effects from text datasets. We will now apply each of these four methods – text propensity score matching (Prop), text representation matching (Rep.), IPW, and measurement error (ME) – to the synthetic datasets we have introduced. Our released code reproduces these experiments.

7.4.1 Structured Variable Distribution

In §7.2, we introduced hyperparameters that control the causal effect of a structured variable on a text generation model. To build our datasets, we first define $p(Y, A, C, U)$ and then define the text distribution $p\left(T|U\right)$. We limit ourselves to the DAG in

Figure 7-1 and only consider binary structured variables.

We choose the parameters of $p(Y, A, C, U)$ randomly, subject to three constraints. First, we ensure that the true distribution-level causal effect (7.1) is equal to 0.1; given C and U, the treatment increases the likelihood of the outcome by 0.1. Second, we ensure that our dataset exhibits Simpson's paradox: if we estimate (7.1) **without conditioning** on U, the causal effect should appear to be -0.1. This setup ensures that methods that completely ignore U and T will fail to estimate the causal effect. Finally, we ensure that $p(U = 1) = 0.5$, which makes a majority-guess strategy for inferring U maximally uninformative. These constraints allow for consistency across experimental evaluations; each structured distribution should be comparable.

7.4.2 Reproducibility of Experiments

Because we have a complex method for producing our text distribution $p\left(T|U\right)$ and we enforce non-trivial constraints on $p(Y, A, C, U)$, we carefully seed the random number generation required to produce these synthetic distributions. In particular, our sampling of text distributions and structured distributions are orthogonal. We consider four separate structured distributions that meet our above constraints, which we reuse in our evaluations across all three text distribution settings: the trivial 16-word vocabulary, the LDA model, and the GPT-2 model.

All results in Tables 7-IV, 7-V, and 7-VI show the absolute-value divergence of the methods' estimates from an oracle with access to the full structured distribution $p(Y, A, C, U)$. The causal estimate errors for a given (τ, δ) pair are averaged over the 16 synthetic distributions that combine our four structured distributions and four text distributions.

Representation

	δ_{word}	0.1	0.2	0.2	0.5
	τ_{word}	0.1	0.05	0.1	0.05
δ_{topic}	τ_{topic}				
0.1	0.1	0.19	0.19	0.20	0.16
0.2	0.05	0.19	0.20	0.19	0.16
0.2	0.1	0.19	0.19	0.20	0.16
0.5	0.05	0.15	0.14	0.14	0.12

Propensity

	δ_{word}	0.1	0.2	0.2	0.5
	τ_{word}	0.1	0.05	0.1	0.05
δ_{topic}	τ_{topic}				
0.1	0.1	0.17	0.17	0.17	0.10
0.2	0.05	0.16	0.16	0.16	0.10
0.2	0.1	0.16	0.16	0.16	0.10
0.5	0.05	0.11	0.11	0.11	0.07

IPW

	δ_{word}	0.1	0.2	0.2	0.5
	τ_{word}	0.1	0.05	0.1	0.05
δ_{topic}	τ_{topic}				
0.1	0.1	0.45	0.38	0.40	0.22
0.2	0.05	0.34	0.39	0.44	0.20
0.2	0.1	0.36	0.42	0.34	0.22
0.5	0.05	0.27	0.24	0.29	0.16

Measurement

	δ_{word}	0.1	0.2	0.2	0.5
	τ_{word}	0.1	0.05	0.1	0.05
δ_{topic}	τ_{topic}				
0.1	0.1	0.07	0.03	0.02	0.01
0.2	0.05	0.04	0.04	0.02	0.00
0.2	0.1	0.03	0.03	0.03	0.01
0.5	0.05	0.01	0.01	0.01	0.00

Table 7-V. Estimation error for each causal method on LDA synthetic data, averaged over the combination of four structured distributions and four text distributions for each cell. All methods reduce estimation error as the δ and τ effects increase in strength, but only measurement error achieves near-zero error for any effect strength.

7.4.3 Evaluation with Trivial Text

Table 7-IV shows how the four causal methods perform on the trivial 16-word vocabulary dataset we introduced in § 7.2. We see that when the $p(T|U)$ relationship is very weak ($\delta_{\text{w}} = 0.1, \tau_{\text{w}} = 0.1$), all four methods perform about as poorly as they would if they had ignored the text entirely. As the $p(T|U)$ relationship becomes stronger, all four methods improve. The text matching and measurement error methods are able to perfectly estimate the true causal effect when the effect of U on T becomes overwhelmingly strong. The IPW method does worse, but does correct for the U confounding as the $p(T|U)$ relationship strengthens. It is not surprising that the measurement error approach works here, as Table 7-I and Figure 7-2 showed us that $p(U|T)$ classification can achieve perfect accuracy on this trivial dataset. The success of the text matching approach highlights that even though $p(A|C,T)$ is not the true propensity score, the relationship between U and T is strong enough to allow for the method to correct for the confounding.

7.4.4 Evaluation with LDA Text

Table 7-V shows how the four causal methods perform on synthetic datasets using the LDA text generation we introduce in § 7.2.2. These results are less encouraging. Our text generated from LDA is word-order independent, so simple bag-of-words models $p(A|C,T)$ should be powerful enough to capture the text's complexity. Even so, the matching methods struggle to correct for U's confounding, though they slightly improve as τ and δ increase. Compared to the trivial setting, in LDA there is a less direct relationship between U and the sampled text. Thus Representation matching is more likely to match two texts with different U values, and in Propensity the estimated $p(A|C,T)$ diverges from the true propensity. That Propensity outperforms Representation when it did not for Trivial text suggests that the propensity matching may be more effective given its low dimensionality (Roberts, Stewart, and Nielsen, 2018). The IPW method, on the other hand, does extremely poorly when the effects of U on T are small. Because a naïve estimator that ignores the text can achieve a causal error of 0.20, the IPW estimator actually worsens the confounding bias. The measurement error approach is effective when τ and δ are large enough.

7.4.5 Evaluation with GPT-2 Text

Table 7-V shows how the four causal methods perform on synthetic datasets using the GPT-2 text generation we introduce in § 7.2.3. Here we see that neither the matching nor IPW methods ever noticeably improve. The measurement error method is still effective, but only when the effect of U on T is strongest.

δ_{word}	0.0	0.2	0.2	0.2	0.5	0.5	0.5	0.5	0.5	0.7	0.7
τ_{word}	0.00	.025	.025	0.15	0.05	0.05	0.05	0.15	0.15	0.15	0.15
δ_{template}	0.7	0.7	0.7	0.7	0.5	0.7	0.7	0.7	0.9	0.7	0.9
τ_{template}	0.45	0.15	0.45	0.15	0.05	0.05	0.45	0.05	0.15	0.15	0.15
Representation	0.19	0.20	0.19	0.19	0.19	0.19	0.19	0.18	0.18	0.17	0.16
Propensity	0.16	0.17	0.16	0.17	0.17	0.17	0.16	0.13	0.13	0.10	0.10
IPW	0.18	0.20	0.17	0.19	0.17	0.19	0.16	0.10	0.09	0.12	0.11
Measurement	0.10	0.10	0.04	0.03	0.03	0.04	0.03	0.01	0.02	0.01	0.01

Table 7-VI. Causal estimation error for each method on the GPT-2 synthetic data. The measurement error method estimates approach zero only for the largest values of δ and τ. Neither Propensity nor IPW correct more than half the confounding of a naive estimator, and Representation barely reduces the confounding bias at all.

While GPT-2 clearly does not produce language at the complexity of real-world datasets, we can better understand the assumptions made by these causal models by exploring how they perform as the underlying text generation become more complex. On this data, simple bag-of-words models we consider are not flexible enough to fully capture the complexity of the text. Even though Table 7-III shows us that a bag-of-words classifier can effectively learn this more complicated $p(U|T)$ when the word and template effects are large enough, the $p(A|C,T)$ model learned for the IPW and matching methods does not capture information on the true propensity. The measurement error method and its $p(U|T)$ classifier can provide unbiased estimates, but only when δ and τ effects are strongest.

7.5 Text Classification Accuracy and Estimation Error

Our propensity score matching, IPW, and measurement error methods all rely in part upon a text classifier to estimate the causal effect. However, better performance (as measured by classification accuracy) of this classifier does not necessarily translate into lower causal estimation error. For both propensity score matching and IPW, the text classifier models $p(A|C,T)$. For the measurement error estimator, the

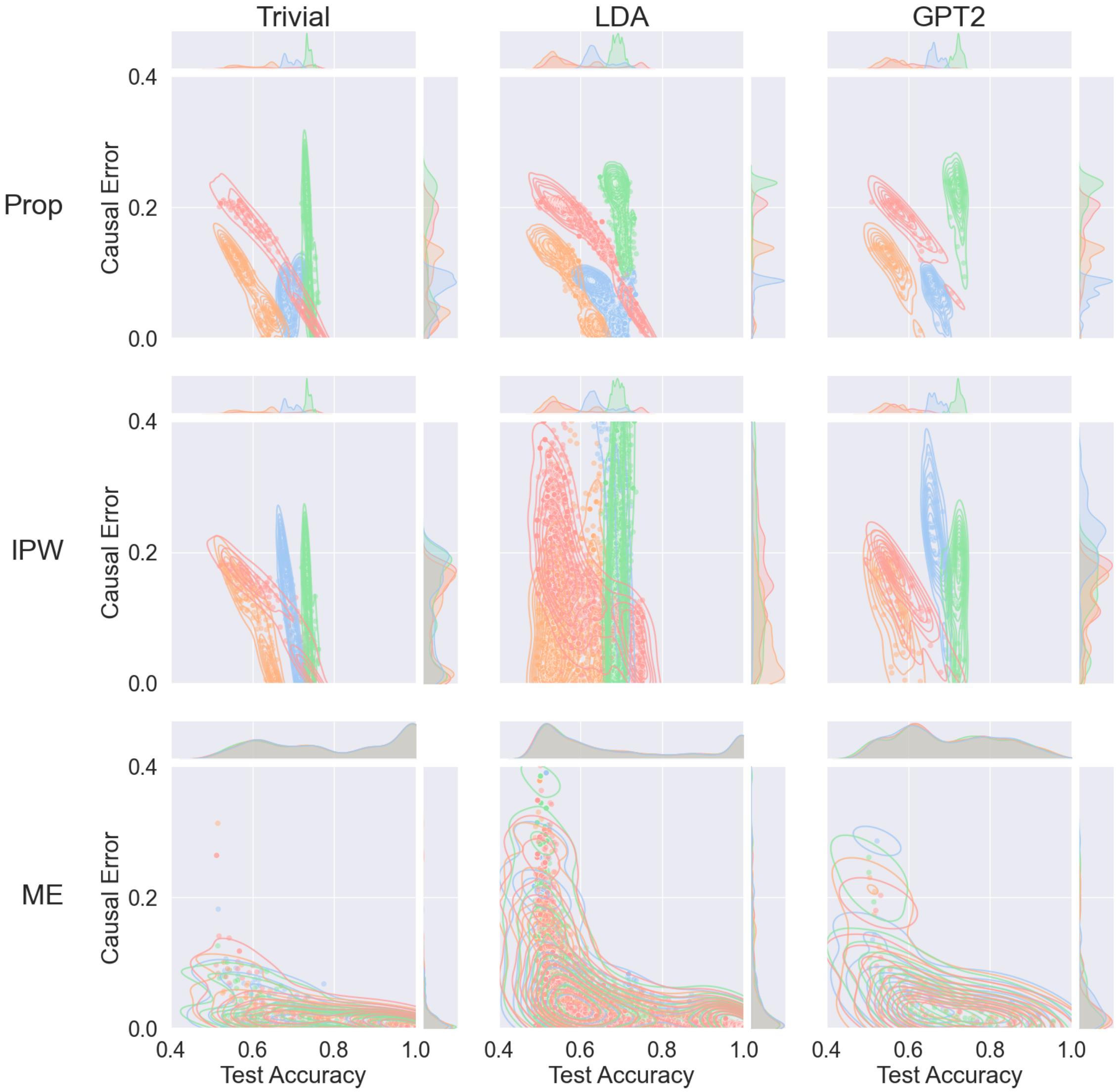


Figure 7-4. Joint and marginal density plots of text classifier accuracy and mean absolute causal estimation error for each DGP and each estimation method that relies on a text classifier. Each dot represents one experiment. Figure 7-5 shows a zoomed-out plot for LDA+IPW; all other plots contain all data. Colors indicate the four structured variable random seeds used to create the true data-generating distributions. For the IPW and Prop methods, the visible clusters show that the relationship between classifier accuracy and causal error is highly dependent on the random seed for structured variables. Thus, for a real-world analysis with an unknown DGP, better classifier accuracy does not imply lower causal error. For the ME method, classifier accuracy and causal error are not clustered by the underlying DGP.

text classifier models $p(U|T)$. For the binary A and U we consider, we can easily characterize these models in terms of their classification accuracy. The density plots

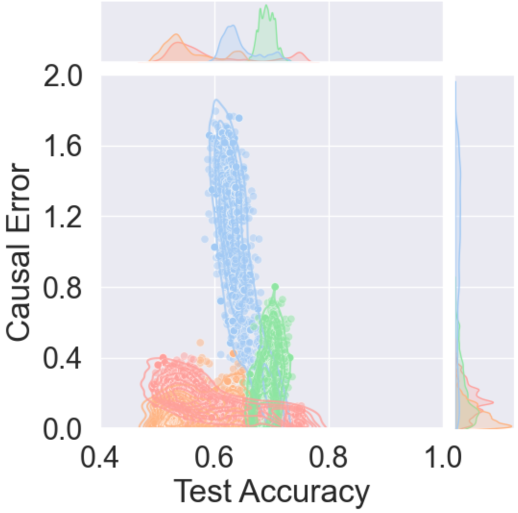


Figure 7-5. Zoomed-out version of Figure 7-4 for for IPW estimator on LDA data. For one random seed for structured variables (the blue cluster), causal error is quite large.

DGP	Prop	IPW	M.E.
Trivial	-0.23	-0.43	-0.57
LDA	-0.14	0.06	-0.58
GPT-2	0.11	0.01	-0.57

Table 7-VII. Pearson correlation between absolute causal estimation error and the test accuracy of the text classifier that the estimation method relies on. On the Trivial text data, all methods have a negative correlation: increased test accuracy implies lower estimation error. As the text DGP increases in complexity to LDA and GPT-2, this correlation dwindles and then reverses for the Propensity and IPW methods, but remains stable for the measurement method.

in Figure 7-4 shows the relationship between text classifier accuracy and the causal estimation error.

Across all three DGPs, we see that when the $p(U|T)$ classifier has accuracy greater than 80%, our estimate of the causal effect is within 0.05 of the truth. If we could achieve 100% classifier accuracy for the measurement method, it would imply that we had access to the true $p(A, Y, C, U)$, and can trivially estimate the causal effect.

However, for propensity and IPW methods, better classification accuracy does

				Labeled $p(U,T)$ examples										
δ_w	τ_w	δ_t	τ_t	50	100	200	300	400	500	1000	1500	2000	2500	5000
0.0	0.00	0.7	0.45	0.19	0.16	0.14	0.11	0.11	0.10	0.08	0.08	0.07	0.11	0.10
0.2	0.03	0.7	0.15	0.19	0.18	0.16	0.17	0.16	0.16	0.16	0.14	0.10	0.10	0.10
0.2	0.03	0.7	0.45	0.18	0.17	0.13	0.09	0.11	0.10	0.10	0.08	0.07	0.06	0.04
0.2	0.15	0.7	0.15	0.19	0.17	0.13	0.15	0.13	0.09	0.06	0.04	0.03	0.03	0.03
0.5	0.05	0.5	0.05	0.17	0.16	0.13	0.11	0.12	0.09	0.04	0.03	0.04	0.03	0.03
0.5	0.05	0.7	0.05	0.18	0.16	0.14	0.13	0.13	0.12	0.06	0.03	0.04	0.04	0.04
0.5	0.05	0.7	0.45	0.18	0.14	0.12	0.10	0.09	0.10	0.06	0.03	0.03	0.03	0.03
0.5	0.15	0.7	0.05	0.10	0.05	0.06	0.04	0.04	0.02	0.02	0.01	0.01	0.01	0.01
0.5	0.15	0.9	0.15	0.11	0.07	0.06	0.05	0.04	0.03	0.03	0.02	0.02	0.02	0.02
0.7	0.15	0.7	0.15	0.10	0.08	0.05	0.04	0.04	0.03	0.02	0.02	0.02	0.02	0.01
0.7	0.15	0.9	0.15	0.10	0.07	0.04	0.04	0.04	0.03	0.02	0.02	0.02	0.02	0.01
$p(U,C,A,Y)$ Baseline				0.28	0.14	0.05	0.03	0.04	0.03	0.03	0.02	0.01	0.01	0.01

Table 7-VIII. Measurement error method's mean absolute estimation error on GPT-2 data as we vary the amount of labeled data used. Train and validation data is split evenly; we train the $p(U|T)$ classifier with half and estimate its error rate on the other half. The last column is equivalent to the last row of Table 7-VI. The $p(U,C,A,Y)$ baseline ignores the text and simply computes the causal effect using Equation 7.1.

not imply lower estimation error. In fact, better classification accuracy of $p(A|C,T)$ is orthogonal to our goals of low causal estimation error. Instead, we need $p(A|C,T)$ to converge to the true $p(A|C,U)$, which is untestable without observing U.

Table 7-VII shows that as we increase the complexity of our DGP from the Trivial text to LDA and then to GPT-2, we can also empirically see that the correlation between classifier accuracy and estimation error degrades for the Propensity and IPW methods. For the Propensity and IPW methods on GPT-2 data, classifier accuracy is positively correlated with estimation error, suggesting that the $p(A|C,T)$ classifier has overfit and diverged from the true $p(A|C,U)$ propensity.

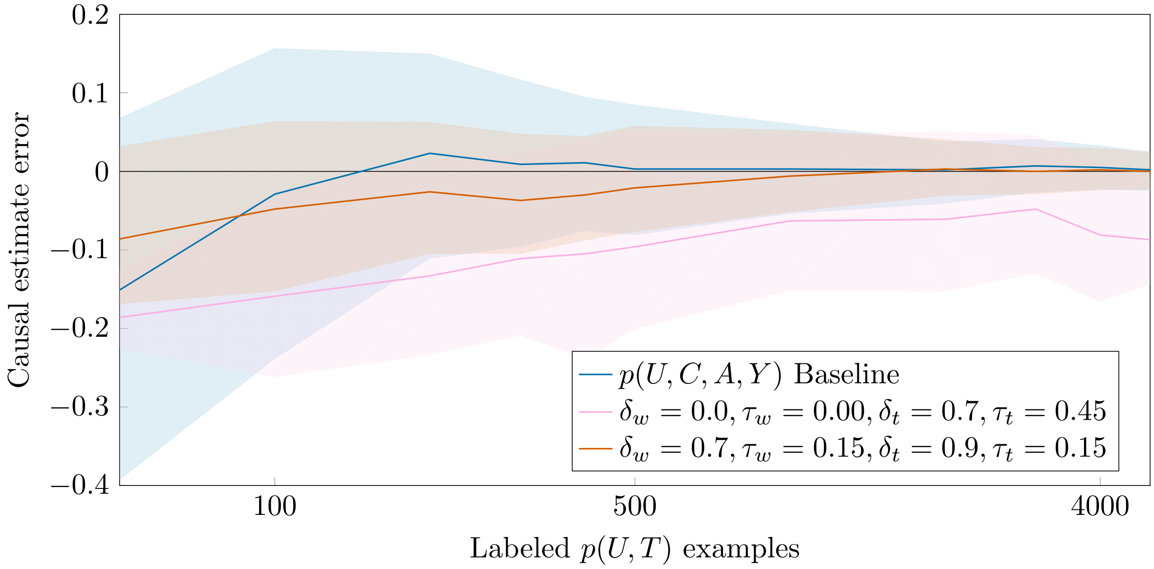


Figure 7-6. A closer look at three rows from Table 7-VIII. Solid line plots mean (not mean absolute) causal error; shaded regions show 95% confidence interval from 100 bootstrap samples. Measurement error results are averaged over four structured variables distributions and four text distributions. The baseline ignores the text and is averaged over four structured distributions. Even for text data with the strongest causal effects we consider, the measurement error approach is not noticeably better than the $p(U, C, A, Y)$ baseline once we have at least 200 labeled examples.

7.6 Availability and Use of Labeled U Data

Our empirical results have demonstrated that the measurement error estimator performs the best on our synthetic datasets. However, this method relies upon access to labeled $p(U|T)$ data. This finding raises two questions: how much labeled data does the measurement error method require, and could other methods perform as well or better if given access to such labeled $p(U,T)$ data?

We run additional experiments where we limit the amount of labeled data that our estimator has access to. Of a dataset of 10,000 total examples, we use n of them to train and validate a classifier $p(U|T)$ and use $(10,000 - n)$ to compute our estimate of the causal effect. Our previous experiments have considered $n = 5,000$; in Table 7-VIII we plot estimation error as we vary n from 50 to 5,000. For the DGPs with the strongest causal effects, the mean absolute error remains small even as we substantially reduce the number of examples. Estimation error on DGPs with weaker causal effects are more sensitive to the number of examples.

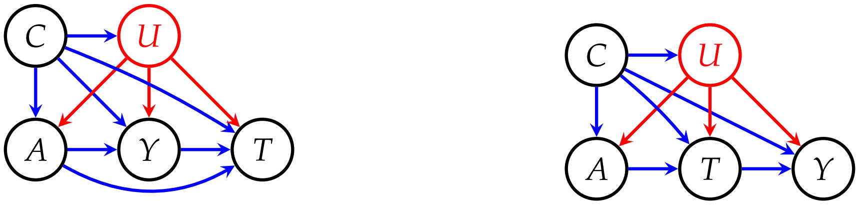
We then compare these evaluations on limited labeled data against a baseline that assumes access to an equal amount of data on the full $p(U, C, A, Y)$ distribution. Suppose we can pay clinicians to annotate n patient records for the unobserved confounder U; should we use those examples to use the measurement error method, or should we just directly compute the causal effect using Equation (7.1), ignoring the text entirely? The $p(U, C, A, Y)$ baseline in Table 7-VIII suggests that as soon as we have at least 200 examples, this baseline is as good on average as the measurement error method, even on DGPs with the strongest $U \rightarrow T$ causal effects. Figure 7-6 shows in more detail this baseline compared against two of the DGPs in Table 7-VIII. In particular, this figure shows the 95% confidence interval for the three methods. For the DGP with large causal effects, the measurement error method is quite comparable to the baseline as $n \geq 500$, but has somewhat smaller confidence intervals at lower-data settings. On the DGP with small $U \rightarrow T$ causal effects, the measurement error method is strictly worse than the baseline.

The measurement error method is the only approach that achieves success on our GPT-2 DGPs, but requires access to $p(U|T)$ labels. If this method can be matched by a baseline that ignores the text entirely, it may seem that incorporating NLP methods into causal inference is not worth the effort. But our results are not entirely pessimistic and the flaws they do reveal point to many opportunities for future work. The $p(U, C, A, Y)$ baseline importantly requires access to the full joint, whereas the measurement error method only requires data on the $p(U|T)$ conditional. This has many practical implications. For example, if researchers at a hospital cannot collect U annotations for their data due to patient privacy restrictions, they still may be able to apply a $p(U|T)$ classifier to that data. Thus if we can leverage existing anonymized clinical datasets as the $p(U, T)$ data, we can produce analysis that would otherwise be impossible.

There are also many opportunities to develop new approaches that outperform the four methods we evaluated. We should expect that some access to labeled data should make it possible to learn a propensity score or text representation that provides for lower estimation error when primarily using data without labeled U. An unsupervised text representation such as LDA could be augmented with labeled $p(U, T)$ so that learned topics are more discriminative of the underlying U (Blei and McAuliffe, 2007). Similarly, if we were given access to some labeled $p(U, T, C)$ data, we could train a propensity score model such that predicted propensities must be roughly equal for examples with the same U. We can also explore approaches that combine these four methods to produce new *multiply-robust* methods. Many causal estimators use multiple models and are provably unbiased if at least one or more of those models are correctly-specified (Bang and Robins, 2005; Vansteelandt et al., 2008). Can we develop a new matching method that are unbiased if either the propensity model *or* representation model are unbiased? Can we effectively combine all four methods we considered into a single multiply-robust estimator?

7.7 Limitations and Extensions

Our evaluation framework and experimental results provide new insights into how existing estimators perform on synthetic text datasets. In generating our synthetic datasets and evaluating these methods, we have made simplifying assumptions. Many of these assumptions may limit the efficacy of our work to certain applications, yet most such assumptions can be relaxed by extending our work.



(a) A DAG in which all structured variables influence text generation.

(b) A DAG in which the text acts as either a treatment or outcome.

Figure 7-7. Causal DAG models to which our evaluation framework could be extended.

7.7.1 Other DAG Models

We only sample datasets from synthetic DGPs corresponding to the DAG model in Figure 7-1. There are of course infinitely many DAG models that could be considered, but we point out a few important generalizations that would complicate our methods for sampling data or evaluating methods.

Figure 7-7a extends Figure 7-1 by adding causal effects from *all* structured variables to the text data. Such a DAG complicates our approach for sampling text from a language model conditional on the structured variables. In § 7.2 we parameterized $p\left(T|U\right)$ with our two types of hyperparameters: δ and τ. Figure 7-7a requires sampling from $p\left(T|U,C,A,Y\right)$, which may require a different hyperparameter formulation. Our implementation assumes U is binary, the immediate extension to a continuous-valued U simply requires replacing the two orderings ($\tilde{V}_{u=1}$ and $\tilde{V}_{u=0}$) with a continuous function of U that outputs an ordering $\tilde{V}_u$. If T is sampled conditional on multiple structured variables, then we need a function that maps from those variables to an ordering over the vocabulary. In such a setting, we need one or more τ hyperparameters that control how sensitive this function is to changes in one or more structured variables.

The DAG in Figure 7-7a also changes the assumptions for the causal methods we consider. The Propensity and Representation methods, like any matching estimator, requires matching only on *pre-treatment* covariates; variables that are non-descendants of the treatment A. Matching on post-treatment variables can introduce significant bias (Rosenbaum, 1984; Stuart, 2010). If the text data is influenced by both U and A, it cannot be easily used for matching. Similarly, for the IPW model (or an outcome model), if the text is a collider (descendant of both A and Y), conditioning on it may introduce bias (Greenland, 2003).

Within the context of the measurement error estimator, Figure 7-7a violates our previous assumption of non-differential measurement error (Carroll et al., 2006; Wood-Doughty, Shpitser, and Dredze, 2018). Thus, rather than estimating two (assuming U is binary) marginal error rates $p\left(U^* = 1|U = 0\right)$ and $p\left(U^* = 0|U = 1\right)$, we must estimate several conditional error rates of the form $p\left(U^* = u'|U = u, A, C, Y\right)$. Estimating such error rates requires data on the full joint $p\left(U, C, A, Y, T\right)$ which, as discussed in § 7.6, reduces the efficacy of these methods compared to simpler approaches that ignore the text data entirely.

In the DAG in Figure 7-7b, the text T can be seen as a treatment or an outcome; $p(T(a))$ is the counterfactual distribution over T if we intervene on A, and $p(Y(t))$ is the counterfactual distribution over Y if we intervene on T. Because our framework currently does not support sampling structured variables conditional on the text, we cannot sample from $p\left(Y|T, U, C\right)$. The causal estimators we consider do not make the necessary assumptions to estimate the high-dimensional effects of A on T or of T on Y (Nabi, McNutt, and Shpitser, 2017; Egami et al., 2018).

7.7.2 Other Language Models

Recent years have seen an explosion in both the frequency and size of neural language models (Bender et al., 2021). While the only such model we have considered is a compressed version of GPT-2 (Sanh et al., 2019; Radford et al., 2019), our framework for adding causal effects can be easily extended to new language models such as GPT-3 or Switch-C (Brown et al., 2020; Fedus, Zoph, and Shazeer, 2021). All our approach assumes is that the model takes as input an initial context and then, for each word, outputs a distribution over the vocabulary. Our causal effects simply adjust the distribution over context inputs and the distribution over the word logits.

Other work on language modeling has focused on *controllable* text generation which can produce sentences that follow a specified style (Xu et al., 2020; Keskar et al., 2019; Kedzie and McKeown, 2020). For example, the approach from Dathathri et al. (2019) specifies topic (e.g. politics) and a sentiment (e.g. negative) which guides the text generation. Such an approach could help generate synthetic datasets which are more domain-specific (see § 7.7.4). In any future work analyzing synthetic text generated from large-scale language models, researchers should be careful to examine how such models learn and reproduce societal biases encoded in the training data (Sheng et al., 2019; Bender et al., 2021).

7.7.3 Better Estimators

We have mentioned in § 7.6 that future work should consider multiply-robust estimators with better asymptotic properties. Our evaluations could also be extended by implementing more flexible (e.g. neural) nuisance models that capture relationship between the structured variables and the text. Veitch, Sridhar, and Blei (2020) proposed causal methods that leverage existing text embeddings which have been widely successful in many predictive tasks. Such neural models may require new assumptions – such as with respect to smoothness (Farrell, Liang, and Misra, 2021) – but have demonstrated empirical performance greatly surpassing that of the bag-of-words logistic regression models we have considered (Rajpurkar et al., 2016). Such neural models often require large datasets for training or pre-training, and in our initial experiments, such models did not outperform logistic regression on our small datasets. Future work could combine pre-training on large datasets (Lee et al., 2020) with fine-tuning on our small datasets (Jin et al., 2019). We could also compare against stronger baselines that ignore the text but leverage all available data, such as the estimator of Yang and Ding (2020) which combines both a small dataset that includes the unobserved confounder and a large dataset that does not. Such an estimator should outperform the $p(U, C, A, Y)$ baseline we considered in § 7.6 by leveraging the additional data that does not contain U.

7.7.4 More realistic DGPs

Our synthetic DGPs enable new evaluations for causal methods for text, but synthetic data in general is not without its inherent limitations. One barrier that prevents generalizability of results on synthetic data to real-world data is that often synthetic DGPs are explicitly designed to demonstrate the utility of a proposed method, and thus other assumptions that could expose the method's flaws may be ignored by the creator (Gentzel, Garant, and Jensen, 2019). While our framework addresses some of these concerns by making it easy to randomize the DGP parameterization and enabling extensions to new language models, there is more that can be done. Gentzel, Garant, and Jensen (2019) suggests semi-synthetic datasets that, for example, use $p(U, C)$ data from a real-world study and then sample $p(A, Y|U, C)$ synthetically so the causal effects are known (Dorie et al., 2019; Shimoni et al., 2018). While our framework could adopt this approach and use empirical $p(U, C)$ data, if we use empirical text data we lose any knowledge of the causal relationships between text and structured variables.

Within the synthetic framework we have proposed, there are many ways to make our synthetic DGPs more realistic for applications to specific domain areas. We have used EHR data and clinical notes as a motivating example throughout, but our DGPs are unrelated to such applications. Suppose we have an EHR dataset with physiological measurements and clinical notes. If we want to conduct a retrospective causal analysis using text, we might first develop a synthetic DGP that tries to approximate the empirical dataset (Neal, Huang, and Raghupathi, 2020). To adapt the synthetic DGPs from this work to this application, we might consider using a language model fine-tuned on clinical notes (Lee et al., 2020) or adapted to the complex vocabulary and style of the domain (Ruch, Baud, and Geissbühler, 2003; Melamud and Shivade, 2019; Boag, Naumann, and Szolovits, 2016; Choi et al., 2017). If our clinical data has a structured variable U that we believe influences the text T, we might incorporate controllable generation techniques to parameterize $p\left(T|U\right)$ more realistically, for example by choosing a vocabulary preference $\tilde{V}_u$ that reflect which words are more commonly used when describing patients with different values of U. Such adaptations could make inferences drawn from synthetic data more robust or make evaluations more interpretable to domain experts.

7.8 Conclusions

Our experiments demonstrate the importance of accurate assumptions in a causal analysis. All four causal methods can control for unobserved confounding in a trivial text generation setting, but as our generative $p\left(T|U\right)$ increases in complexity, the implicit assumptions of the matching and IPW methods render them biased. Although the matching and IPW methods use the same $p\left(A|C,T\right)$ propensity score model, the matching approaches work are superior in the trivial and LDA settings. Even though the trained models are identical, the underlying assumptions are different. Because it requires additional data, the measurement error approach is able to make fewer assumptions, remaining effective as long as its $p\left(U|T\right)$ classifier is accurate. These results do not imply that text matching and IPW methods *cannot* control for unobserved confounding, but rather that we should be cautious and clear about what assumptions we make about our models and the underlying DGP. Evaluating on synthetic data can help clarify these assumptions.

As NLP research furthers the state-of-the-art in predictive modeling, such tools offer the potential to influence human decision-making and guide our understanding of the world. Such models rely on assumptions that may be irrelevant for a supervised learning benchmark and yet essential to any real-world application. Explicitly adopting a causal inference perspective on natural language datasets can help enable inferences that are robust to confounding or other biases. We hope our evaluation framework and released code will support further research in these directions.

Chapter 8

Proxy Models for Explaining Black-Box Models

8.1 Introduction

This chapter switches from a focus on causal applications of predictive models to the study of interpretability of those models themselves. While interpretability is not exclusively a causal question, it has been formalized in ways that draw heavily from causal inference literature (Sani, Malinsky, and Shpitser, 2020; Broniatowski, 2021). The proxy model approach we introduce in this chapter is inspired by the causal interpretation of time-varying confounding between a time series model's predictions, yet the work presented here does not discuss applications to causal inference. The work in this chapter is adapted from two papers which are under review; one is available as a preprint as Wood-Doughty, Cachola, and Dredze (2021). Chapter 9 will build upon this methodology to connect our proxy models back to our motivating examples of combining NLP and causal inference.

As we have discussed throughout this thesis, starting in Chapter 3, machine learning (ML) methods have demonstrated their ability to make accurate predictions across many domains. ML methods can automate decision-making based on consideration of high-dimensional and heterogeneous data that may be overwhelming for human domain experts. In high-stakes domains such as clinical care or political decision-making, high accuracy alone is not necessarily enough to motivate expert adoption of automated systems (Caruana et al., 2015). If a neural network for medical image analysis arrives at a diagnosis and recommended treatment that contradict a physician's judgment, that doctor may not know how to reconcile the disagreement, especially as modern neural networks – with millions or billions of parameters – may be impossible for a non-expert to understand (Feng et al., 2018). Explainable Artificial Intelligence (AI) can enable models to provide the explanations needed for them to be integrated into decision-making processes (Tonekaboni et al., 2019). Trust in ML models is a function of both accuracy and explainability; model predictions need to be accompanied by an explanation that can be interpreted by the people who will decide whether to act on that prediction.

This need to understand model predictions has motivated a large body of ML research into interpretability and explainability. The need for explanations introduces a number of competing goals and corresponding approaches. First, explainable methods can be divided into two types: explainable-by-design models and post hoc explanations for an existing trained model. Explainable-by-design requires training a model from scratch, which is unattractive when an application-specific model has already been developed and validated. However, many post hoc methods only explain one example at a time, potentially making them slow and inconsistent from example to example (Linden, Haned, and Kanoulas, 2019). A second trade-off considers the competing goals of faithfulness (explanations that accurately convey the decision-making process of the model) and plausibility (explanations that make sense to domain experts). Balancing these goals can be challenging; faithful explanations that accurately convey the reasoning of complex AI systems may be implausible to a domain expert, and vice versa. The local versus global explanation trade-off means that more detailed explanations for a single example (local) can provide greater insights but may also make it more difficult to understand overall trends (global) in the model's behavior if each explanation is overly specific to a single example. Finally, models must also balance sophistication against transparency. Sophisticated methods – such as large neural networks fine-tuned to a specific domain – may yield the best performance on a task, but are often the least able to provide plausible explanations.

We propose to disentangle these competing goals by introducing a *proxy model*. The proxy model itself is a new explainable-by-design model, but is trained to closely mirror the predictive behavior of an existing trained model. We train the proxy model on the *predictions* of the trained ML model, so that the behavior of the proxy model mimics the trained model's behavior, rather than independently modeling the target task. We then rely on the interpretable proxy model to create explanations, allowing the trained model to use sophisticated methods to achieve high accuracy. We apply our approach to two challenging task that rely on deep learning models: forecasting global disruptive political events with an LSTM and predicting medical billing codes from clinical text. For each of these settings, we demonstrate the proxy model's faithfulness to the original trained model by showing it makes similar predictions on held-out data.

Whereas many existing methods were primarily developed for computer vision applications and convolutional neural network (CNN) models, When applied to the time series LSTM, our method can be thought of as *controlling for the confounding effect* of past predictions. In the time series application, we compare our method against several methods from popular explainability toolkits and find that our proxy is several orders of magnitude faster than these methods for generating explanations and produces globally-consistent estimates of feature importance across examples, providing a more holistic summary of the trained model. In the clinical text setting, we show that the proxy model is faithful to the original model and produces plausible explanations, as measured on clinician annotations of generated explanations.

8.2 Background on Explainable AI

Recognition of the importance of explanations has driven a wave of research in Explainable Artificial Intelligence (XAI), the broader field under which interpretable ML falls. We present an overview of major themes in the literature, and direct the reader to recent surveys for more details (Doshi-Velez and Kim, 2017; Guidotti et al., 2018; Gilpin et al., 2018; Arrieta et al., 2020).

Past work distinguishes between "transparent" or "inherently interpretable" models that offer their own explanations, and "post hoc" methods that produce explanations for a separately-trained model. Methods such as logistic regression are often considered transparent or inherently interpretable, because their simplicity allows a domain expert to understand how a change in input would produce a different output (Guidotti et al., 2018). However, even simple models can prove difficult to interpret in certain settings, such as when the model's features are complex (Lipton, 2018). LIME is an example of a post hoc method (Ribeiro, Singh, and Guestrin, 2016); given a trained model of arbitrary complexity it produces explanations for individual predictions. The trade-off in the different methods is that inherently-interpretable methods are often limited in model complexity. Deep neural networks, for example, often demonstrate better performance but are not inherently interpretable (Feng et al., 2018), and typically rely upon post hoc methods to derive explanations (Guidotti et al., 2018).

Lipton (2018) critiques the idea of "inherent" interpretability and argues that methods that are intended to be transparently understood should pursue several traits. These include simulatability, or whether a human can reasonably work through each step of the model's calculations to understand how a prediction is made; decomposibility, or whether each parameter of the model can be intuitively understood on its own; and algorithmic transparency, or whether the model belongs to a class with known theoretical behaviors. Lou, Caruana, and Gehrke (2012) highlights linear and additive models as particularly decomposible (or intelligible) classes of models, because "users can understand the contribution of individual features in the model." Our proposed approach will use a linear model trained on simple features representations to provide a simulatible, decomposible, and transparent method.

Interpretability methods are also distinguished by the form and quality of the explanations they produce. Two primary desiderata for explanations of ML systems are "faithfulness" and "plausibility."[1] A faithful method accurately describes the true machinery of the model's prediction, while a plausible model produces explanations that can be interpreted by a human expert (Jacovi and Goldberg, 2020). A method could be faithful but not plausible, if it accurately explains a model's predictions but does so in terms of high-dimensional feature vectors that a human cannot interpret. Similarly, a method could be plausible but not faithful, if it produces concise natural language summaries that are unrelated to the calculations that produce the model's predictions. Methods should attempt to achieve both goals, but there is a trade-off between the two; explanations typically cannot be both concise and perfectly descriptive. Plausibility, unlike faithfulness, necessarily requires an evaluation based on human perception (Herman, 2017). A strength of our proposed method is that it is designed for plausibility and transparency, but optimized for faithfulness.

[1] Faithfulness is also referred to as validity or completeness; plausibility is alternatively referred to as persuasiveness (Herman, 2017) See Jacovi and Goldberg (2020) for a longer discussion of alternate terminology.

8.3 Proxy Model Explanations

Our proposed proxy model approach is post hoc and seeks to balance faithfulness and plausibility while providing global consistency in explanations. We assume that we have a trained model with good predictive performance but low interpretability. Given this trained model and a dataset on which it can be applied, we train a *proxy model* that takes the same input from the dataset, but uses the trained model's predictions as its labels. In other words, given the dataset's input, the proxy model predicts the outputs of the uninterpretable model. We model the output probabilities to allow the proxy to learn from the detailed behavior of the trained model, rather just the binary decisions. The proxy provides explanations based on its coefficients that can be individually understood as measuring the contribution of a feature to the original model's predicted probabilities. Training the proxy model on predictions from the existing model optimizes for faithfulness by design.

We also want the proxy model to produce plausible explanations and fulfill the criteria from Lipton (2018): simulatibility, decomposibility, and algorithmic transparency. To do so, we restrict our proxy model to a class of models that fulfills these desiderata. The fundamental trade-off here is that if we restrict our model class too much, the proxy will be unfaithful and unable to mimic the behavior of the trained model. But if we allow for a proxy model that is too complex, it may not provide plausible or otherwise desirable explanations. The choice of proxy model requires some consideration of the particular domain, as feature preprocessing and similar details may affect its behavior and explanations.

Our approach is similar to LIME (Ribeiro, Singh, and Guestrin, 2016) in that it learns a simple (linear) model to explain a pretrained model. However, whereas LIME learns a linear model to post hoc explain a single prediction, our linear model is trained to predict and explain the entire dataset of predictions. This has several consequences. Unlike LIME, we do not require sampling perturbed inputs that do not exist in the training data, which can produce contrasts which are misleading or unintuitive (Mittelstadt, Russell, and Wachter, 2019). Slack et al. (2020) showed that LIME can be fooled into providing innocuous explanations for models that demonstrate racist or sexist behavior by exploiting its reliance on perturbations. It also means that our proxy model is given a more difficult task than a LIME model – it may be that a given proxy model is insufficiently flexible to model the complexity of the pretrained model, in which case we can measure this failure in terms of our faithfulness evaluation. Because LIME trains a model linear only in the neighborhood of a given instance, its feature importance scores are difficult to aggregate across a dataset, making extrapolation difficult (Linden, Haned, and Kanoulas, 2019). When our proxy model is faithful to the trained model, our approach gives us explanations that we can expect to apply to future predictions. If the proxy model demonstrates sufficient empirical performance, a domain expert may even prefer to use it in place of the original trained model, an option unsupported by LIME models.

In §8.4 and 8.5, we will introduce two tasks: predicting disruptive political events from news and economic data, and predicting medical diagnostic codes from clinical notes. The political event task highlights how our method exploits a time series setting to account for past model predictions, and the medical coding task allows us to compare against domain expert annotations for model plausibility. For each task, we will apply our proxy model approach to a trained model published in previous work, and evaluate the faithfulness, plausibility, consistency, and runtime of our approach. In both settings, we see ways in which our method improves over existing approaches.

8.4 Proxy Model Explanations of a Time Series LSTM

8.4.1 Introduction

For the first application of our method, we focus on a challenging task with a complex model: forecasting global disruptive political events (i.e. irregular government leadership changes) using a recurrent neural network (RNN). This task highlights the ability of our proxy method to explain a complex model with high-dimensional time series data. Whereas much previous work was primarily developed for computer vision applications and convolutional neural network (CNN) models, our method draws inspiration from causal inference research to control for the confounding effect of past predictions for our time series application.

We compare our proxy model approach for generating explanations against several methods from popular explainability toolkits. In addition to LIME (see §8.2), we consider three widely-used salience methods, which use the gradient of a neural network to determine which features are most influential for a given prediction. These methods take in the trained model and its input and compute gradients of the model's output with respect to its feature values. We use implementations for these three methods found in the `iNNvestigate` package (Alber et al., 2019). The first of these is referred to as "Simple Gradient" or "Gradient" and is just the raw gradient of the prediction with respect to the inputs (Simonyan, Vedaldi, and Zisserman, 2014). Input × Gradient (I × G) takes that same gradient and multiplies it by the values of the input features (Shrikumar et al., 2016). Layer-Wise Relevance Propagation computes gradients in a modified backpropagation approach that maintains a local conservation property, which can be viewed as a sequence of Taylor decompositions (Bach et al., 2015; Montavon et al., 2019). `iNNvestigate` provides many additional methods based on other published work, but we found that all other methods either could not run[2] or duplicated another method's explanations. We leave for future work a comparison against additional methods.

We find that existing methods are inconsistent, making aggregated understanding of feature importance values difficult. Our proxy model is several orders of magnitude faster than these methods for generating explanations and produces globally-consistent estimates of feature importance across examples, providing a more holistic summary of the trained model.

8.4.2 Politically Disruptive Event Forecasting

Our data comes from Global Data on Events, Location, and Tone (GDELT), which includes millions of geolocated events (Leetaru and Schrodt, 2013) in the form of monthly aggregated counts of news articles that reference events in 164 countries. These events cover different types of political actions, e.g. "Threaten to halt negotiations." As the amount of news coverage ingested into the GDELT dataset grows year-to-year from 2002 to 2019, each month's event counts are normalized by the total number of events of that type worldwide. To supplement GDELT's events with basic information about each country, our data also includes features drawn from the World Development Indicators (WDI) (*World Development Indicators* 2021).

Our goal is to utilize these features to forecast Irregular Leadership Change (ILC) events drawn from (Raleigh and Dowd, 2021), defined as a political leadership change that does not follow the country's normal laws or conventions (Beger, Dorff, and Ward, 2016). These events are rare, with only 180 occurring in the data between 2002 and 2019. This complicates training and evaluation, as a model can achieve 99.5% accuracy by never predicting an ILC event.

We begin with "Crystal Cube," (Parrish et al., 2018; Buczak et al., 2021) a

[2] Many methods only work for CNN models. `iNNvestigate` includes 17 variations of the LRP approach, but all return the same explanations for our trained model.

Comparison	Crystal Cube				Ground Truth	
Model	ρ	τ	APS	AUC	APS	AUC
Crystal Cube	1	1	1	1	0.037	0.776
Baseline	0.024	0.020	0.033	0.504	0.009	0.506
Proxy	0.907	0.799	0.904	0.995	0.029	0.770
L1 Proxy	0.923	0.816	0.869	0.989	0.031	0.778

Table 8-I. Faithfulness evaluation using Spearman ρ, Kendall τ, Average Precision Score (APS), and ROC Area Under the Curve (AUC).

pre-existing model developed to predict ILC events from the GDELT and WDI data. Crystal Cube uses an LSTM (Hochreiter and Schmidhuber, 1997) with 1454 input features derived from GDELT article counts and WDI statistics, with an instance representing each month for each country. The model maintains a hidden state that is updated each month before making a prediction, which allows the model to capture ongoing changes in a country's political environment that may foreshadow an ILC event.

Crystal Cube achieves a ROC AUC score of 0.776 and an average precision score (APS) of 0.037 for the task of predicting ILC events. The gap between these metrics is due to the imbalanced nature of the labels. Parrish et al. (2018) introduced the model and demonstrated that it outperformed several forecasting baselines. For comparison, our logistic regression baseline in Section 8.4.4.1 (Table 8-I) achieves an ROC AUC of 0.506 and APS of 0.009. Crystal Cube is a complex neural network with 50 million parameters. However, it is also a carefully-trained, domain-specific model that achieves good performance, and thus we seek to create explanations for it, rather than replace it by training an explainable-by-design model from scratch (Rudin, 2019). This makes it a perfect application for our post hoc proxy model method. We refer readers to (Parrish et al., 2018) for full details on the dataset and model training and evaluation.

	Gradient		I × G		LRP	
Feature (week offset)	Rank	Value	Rank	Value	Rank	Value
Accede to demands for change in leadership (wk-1)	20	1.2e-3	4	4.5e-5	22	5.6e-4
Accede to demands for change in leadership (wk-2)	32	1.1e-3	2	4.9e-5	33	5.2e-4
Accede to demands for political reform (wk-1)	31	1.1e-3	5	4.4e-5	31	5.2e-4
Accede to demands for political reform (wk-2)	58	9.8e-4	3	4.7e-5	61	4.7e-4
Accede to demands for political reform (wk-5)	60	9.8e-4	62	1.4e-5	62	4.7e-4
Appeal for change in leadership (wk-1)	30	1.1e-3	71	1.2e-5	30	5.3e-4
Appeal for change in leadership (wk-2)	3	1.4e-3	36	2.0e-5	3	6.5e-4
Appeal for change in leadership (wk-4)	41	1.0e-3	22	2.8e-5	42	4.9e-4
Appeal for change in leadership (wk-5)	26	1.1e-3	52	1.6e-5	28	5.3e-4
Demand release of persons or property (wk-2)	124	7.9e-4	170	6.0e-6	125	3.8e-4
Previous Model Prediction						
Threaten with administrative sanctions (wk-2)	1	1.8e-3	912	8.4e-7	1	8.6e-4
World Development Index Feature 23	1239	1.1e-4	1	5.4e-5	1246	5.4e-5
World Development Index Feature 3	1363	6.9e-5	29	2.6e-5	1369	3.2e-5

	LIME		Proxy		L1 Proxy	
Feature (week offset)	Rank	Value	Rank	Value	Rank	Value
Accede to demands for change in leadership (wk-1)	85	4.1e-3	174	0.112	2	0.178
Accede to demands for change in leadership (wk-2)	28	4.9e-3	592	0.030	5	0.096
Accede to demands for political reform (wk-1)	154	3.3e-3	388	0.048	14	0.029
Accede to demands for political reform (wk-2)	132	3.5e-3	211	0.092	21	0.021
Accede to demands for political reform (wk-5)	301	2.6e-3	120	0.165		
Appeal for change in leadership (wk-1)	59	4.4e-3	509	0.036		
Appeal for change in leadership (wk-2)	7	6.3e-3	163	0.118		
Appeal for change in leadership (wk-4)	13	5.9e-3	310	0.062	9	0.042
Appeal for change in leadership (wk-5)	112	3.7e-3	145	0.131	6	0.056
Demand release of persons or property (wk-2)	1	8.1e-3	794	0.021		
Previous Model Prediction			72	0.814	1	0.862
Threaten with administrative sanctions (wk-2)	1318	3.2e-4	472	0.039		
World Development Index Feature 23	66	4.3e-3	21	44.661		
World Development Index Feature 3	638	1.7e-3	1	5.1e+2		

Table 8-II. Most important features across methods.

8.4.3 Proxy Model Explanations

We utilize a linear regression as a proxy model to explain Crystal Cube's predictions. Our proxy model takes as input the same data as Crystal Cube but is trained to predict its probabilistic outputs, rather than the binary ground-truth labels. Unlike the sequential structure of Crystal Cube, our proxy predicts the probability of a country's ILC using only the features of a single month. To address the time series nature of the application, we also give the proxy model access to the previous month's prediction probability from Crystal Cube. This draws inspiration from causal inference methods, in that it allows our model to *adjust for confounding* of the Crystal Cube's prior state (Pearl, 2009). Because Crystal Cube is a recurrent model, its month-to-month predictions are highly correlated. Our proxy model seeks to *explain* the recurrent model by pointing to which features in a given month caused the model to change its prediction.

While a linear proxy model is easily interpretable, to be effective it must be *faithful* to Crystal Cube (Jacovi and Goldberg, 2020; Du, Liu, and Hu, 2019). We can explicitly measure the faithfulness of our explanations by evaluating how closely our proxy outputs match those of Crystal Cube using a held-out test set of predictions.

Our training set consists of the GDELT and WDI data from 2002 to 2011, and our test set covers 2012 to 2019. Our loss function is the mean squared error (MSE) between proxy outputs and the real-valued probability outputs of Crystal Cube. To evaluate faithfulness, we compute the held-out Spearman ρ correlation and Kendall τ correlation between Crystal Cube's predictions and the proxy model's outputs. In addition to these regression-based metrics, we threshold[3] Crystal Cube's predictions to create binary labels, and compute ROC AUC score and APS

[3] Using a threshold of 0.22, following (Buczak et al., 2021).

	Gradient	I × G	LRP	LIME	Proxy
I × G	0.076				
LRP	1.000	0.076			
LIME	0.058	0.058	0.058		
Proxy	0.000	0.506	0.000	0.024	
L1 Proxy	0.072	0.169	0.072	0.030	0.000

Table 8-III. Jaccard similarity between methods

comparing our real-valued proxy outputs against these binary labels. Finally, we use the true ILC event labels and compute ROC AUC and APS for the true events.

For simplicity and reproducibility, we use models from the `sklearn.linear_model` package. Our first, denoted 'Proxy' in our tables, is a linear regression model without regularization. To further simplify the proxy model, we introduce an L_1 (Lasso) regularization penalty with α parameter [4] of 5e-5, which induces sparsity into models coefficients. We denote this model as 'L1 Proxy.' We compare both proxies against a baseline logistic regression ('Baseline') that is trained to directly predict the ILC event labels, independently of Crystal Cube. We expect the baseline to be somewhat correlated with Crystal Cube simply because both models address the same task, but our proxy models should be much more faithful.

8.4.4 Evaluation

8.4.4.1 Proxy Model Evaluation

Our first evaluation explores whether our proposed proxy model approach is *faithful* to Crystal Cube. Table 8-I shows our evaluation of Crystal Cube, the logistic regression baseline, Proxy, and L1 Proxy models on how well they predict the held-out test set Crystal Cube predictions and the ground truth labels for ILC events, using the faithfulness metrics described in Section 8.4.4.1. All of these metrics have a maximum of 1.0; Crystal Cube is perfectly correlated with and

[4] We try five α values from 1e-3 to 1e-5 and found 5e-5 balanced sparsity and performance.

predictive of itself.

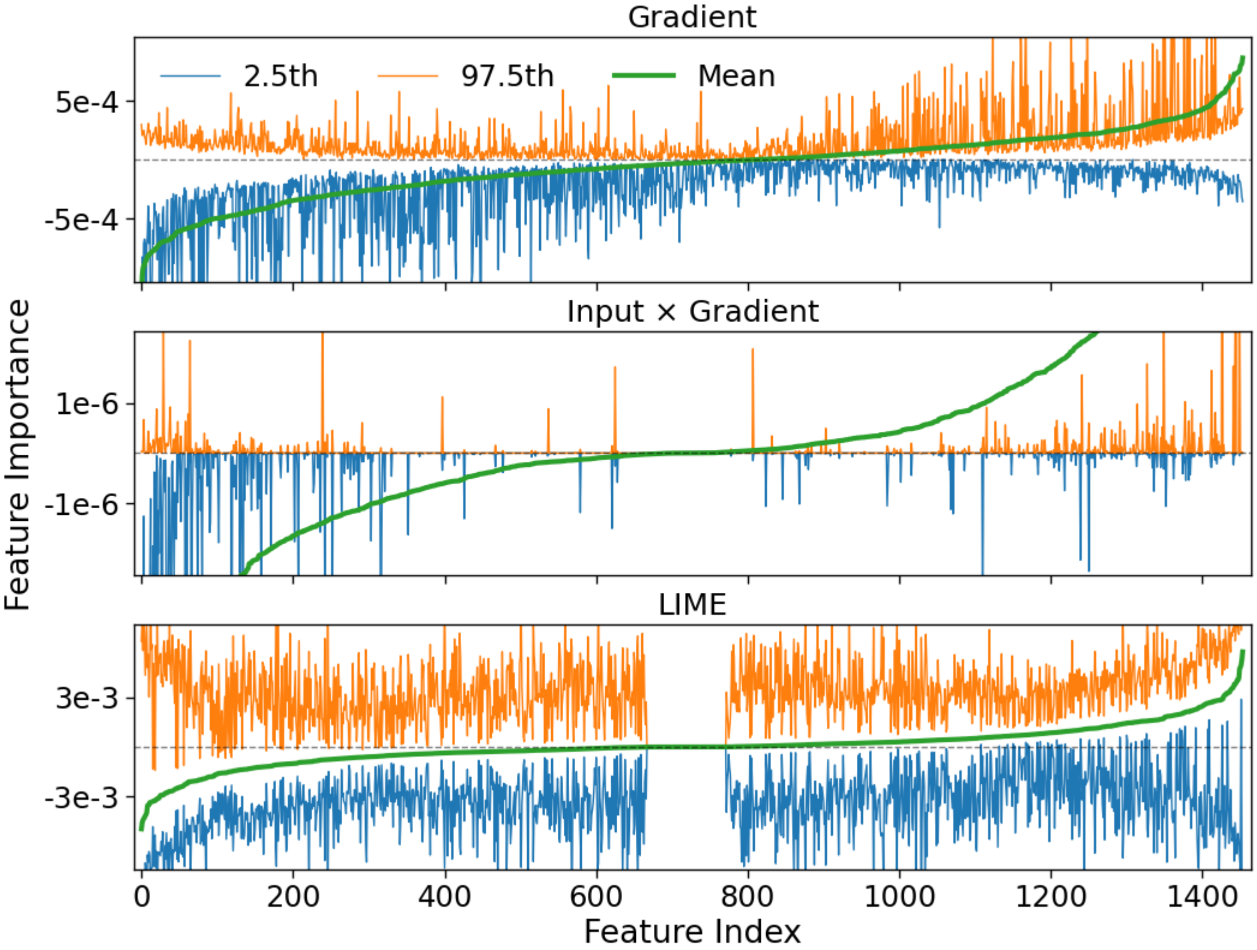


Figure 8-1. Distribution of feature values for local methods. For almost every feature, the empirical distribution of importance scores contains zero between its 2.5th and 97.5th percentiles. Local methods are not consistent in determining whether features make ILC predictions more or less likely. Whitespace appears in I × G and LIME plots because there are 713 and 105 features, respectively, where both the 2.5th and 97.5th percentiles are 0.

The logistic regression Baseline has almost no predictive power for either Crystal Cube's predictions or the true ILC event labels. The unregularized Proxy model scores well on all four faithfulness metrics, while losing 0.008 and 0.006 points on APS and AUC scores respectively for predicting ILC events. Introducing L1 regularization further reduces the APS and AUC scores against Crystal Cube, but increases the correlation scores and the ground truth predictive performance. While these results do not suggest we should simply discard Crystal Cube and use our proxy models to predict events, the proxies perform extremely similar to Crystal Cube across the entire test set. Although L1 Proxy scores slightly lower than Proxy on predictive metrics of faithfulness, with only 40 nonzero features (compared to Proxy's 1455) it provides an even simpler representation of Crystal Cube's behavior.

8.4.4.2 Differences Between Explanations

We now compare our proxy model explanations against four methods for producing post hoc local explanations described in Section 8.2. Each local method uses the trained Crystal Cube model and its input features to individually explain the model's prediction for each country and month. From the `iNNvestigate` library released by Alber et al. (2019), we use the Gradient, I × G , and LRP methods for local explanations. LIME is our fourth local method, using the `LimeTabularExplainer` class in the implementation released[5] by the authors. Due to runtime constraints, we reduced LIME's number of sampled perturbations from the default of 1000 to 500, but otherwise left all defaults unchanged. We run these methods across all countries and months in our train and test sets.

Table 8-II shows the top features across methods when aggregating by taking the mean absolute value of each feature importance across all examples, regardless of the ground-truth ILC event label and whether Crystal Cube correctly predicted it. For each feature, we show its rank and its mean absolute importance according to each method. The table shows the top ten features with the highest average ranking across methods, with the top-ranked feature from each method added if necessary. There are several aspects of these results that stand out:

1. The scale of importance scores is different between methods; I × G has a maximum value of 5.4e-5 compared to Proxy at 5.1e+2.

2. Many of the top features are either "accede to demands" or "appeal for

[5] https://github.com/marcotcr/lime

change" events. However, even if we focus only on the "appeal for change in leadership" events within the same method, we see large variance in the feature importance and ranking between weeks.

3. Gradient and LRP are the most correlated methods, whereas Proxy and L1 Proxy are surprisingly unrelated.

We now more closely examine correlations between methods. Table 8-III shows Jaccard similarities between the top 64 most important features for each method. We rank features by taking the mean of their absolute importance scores across all countries and months to also account for negative values (features that reduce the predicted probability of an ILC event). We compute Jaccard similarities involving the L1 Proxy using only its 40 nonzero features. We see that Gradient and LRP methods are equivalent in terms of Jaccard similarity, even though the scale of their feature importances are quite different in Table 8-II. Except for that pair of methods, no two methods are particularly closely related. I × G most closely overlaps with the unregularized Proxy, followed by the L1 Proxy. Other than these examples, no pair exceeds a modest similarity score of 0.1. Because there is no ground truth for the *correct* way to explain Crystal Cube's predictions, we cannot directly evaluate whether one method provides better explanations than the rest. We can either choose to aggregate across methods (such as in Table 8-II) or use domain knowledge to choose a specific method.

The methods we consider make different assumptions about the nature of how a trained model should be explained, and thus it is not entirely surprising that they produce different explanations. For a domain expert, a greater acceptance of one method's assumptions may give more weight to its explanations. While we do not have expert testimony for this data, we can provide some analyses of the domain-independent pros and cons of the various methods.

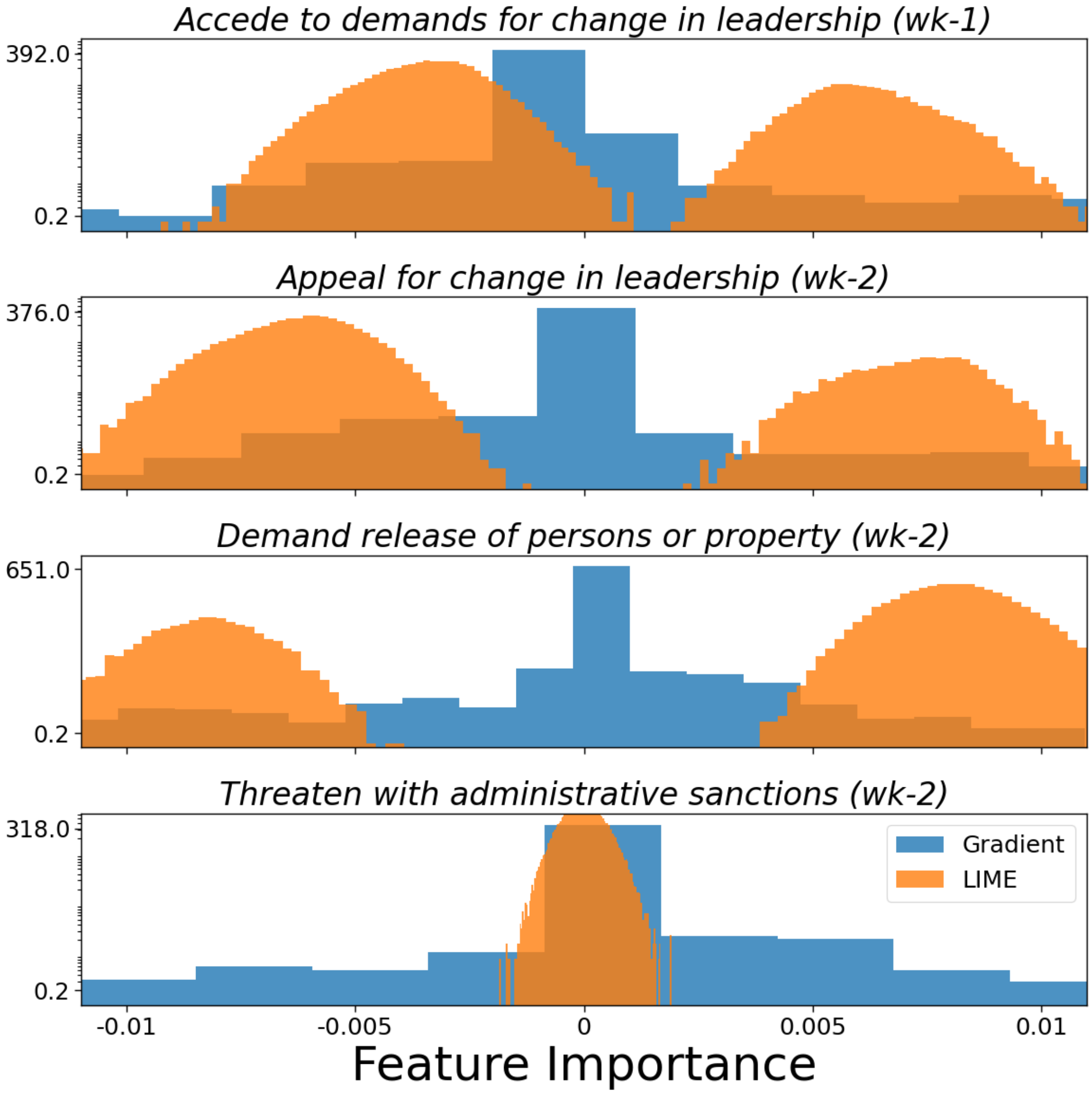


Figure 8-2. Log-scale histograms for the distribution of importance scores for a few top features from Table 8-II. For the Gradient method (and I × G and LRP not shown), almost all importance scores are close to zero, and are otherwise roughly symmetric. For three of the LIME features, the distribution is asymmetric but bimodal.

8.4.4.3 Runtime

A major benefit of our proposed proxy model approach, in addition to its overall simplicity, is its speed. Whereas the other methods we consider make *local* explanations and thus must consider each combination of month and country independently, our proxy model considers all such examples simultaneously. This results in an improvement in runtime of several orders of magnitude. Training a proxy model regression takes only a few minutes on the 20k training examples; once trained, its parameters do not change between explanations. In contrast, the local explanation methods we compare against require at least a few seconds for each prediction they seek to explain. Because our train and test set together contain 164 countries and 215 months, this results in over 40 hours of computation for the Gradient method and over 140 hours for LIME. Although parallelization can decrease the wall-clock time dramatically, these runtime numbers reflect the total time to run on a single CPU. However, the enormous speed advantage of our proxy model approach can enable analyses that might otherwise be too expensive.

8.4.4.4 Consistency and Plausibility

Another differentiating factor between the proxy model and local explanations methods is the variance within the explanations of each method. The proxy model produces only a single importance score for each feature across the entire dataset. Each local method produces an importance score per feature per example, which means we must aggregate those explanations to provide a summary of the overall behavior of the model. In doing so, we must handle variance: what happens if a feature is sometimes highly predictive of an ILC event and sometimes predictive of no event?

Figure 8-1 shows plots of the mean and percentiles for the feature importance values of the Gradient, $\mathrm{I} \times \mathrm{G}$, and LIME methods. In these plots, the x-axis is an ordering over the features based on their mean absolute importance for the given method. For each slice at $X = i$, we see the mean and percentiles for the ith feature. For almost every feature across these methods, its empirical distribution of feature importance contains zero between its 2.5th and 97.5th percentiles. This means that even for features that are *on average* predictive of an ILC event occurring, i.e. the 'useful' features, these features are still often found to be predictive of no event. The way these methods explain these features is inconsistent; it may be locally informative but is not globally reliable.

Figure 8-2 shows a closer look at the distributions of individual features for the Gradient and LIME methods, using a few of the most important features from Table 8-II. The LIME plots show an asymmetric bimodal distribution in three of the four examples, with the feature being labeled as either positive-predictive or negative-predictive, depending on the context. This includes "Demand release of persons or property (wk-2)," the overall most important feature for LIME. The Gradient plots show a nearly point-mass distribution; for "Threaten with administrative sanctions (wk-2)," over 90% of Gradient's feature importance scores are between -1e-5 and 1e-5, and the remaining density is split 56% negative and 44% positive. While we omit LRP and I × G histograms, they are similar[6] to the Gradient plots.

These results indicate that for the local methods, whether and how we aggregate feature importance values can provide vastly different explanations. If a feature's average importance is dominated by a few outliers and is otherwise symmetric and mean-zero, it may be difficult to interpret whether past explanations should influence our understanding of future model predictions. This may jeopardize the ability of these methods to instill trust in domain experts who do not understand

[6] For "Threaten with administrative sanctions (wk-2)," 92% of LRP scores lie in (-1e-5, 1e-5) and the remainder are split 59% to 41%. For "Accede to demands for change in leadership (wk-2)," 96% I × G scores lie in (-1e-7, 1e-7) and the remainder are split 42% to 58%.

the underlying black-box model.

In this context, a benefit of our proxy model approach is its *consistency*. In training our Proxy and L1 Proxy models, we assign a single importance score to each feature. This naturally aggregates the importance of each feature across the entire dataset. The addition of L1 regularization also greatly reduces the number of nonzero features, allowing the L1 proxy to summarize Crystal Cube's global behavior as a collection of 40 features. Looking at features in Table 8-II, we can see that the L1 Proxy accounts for the temporal correlation between Crystal Cube predictions by putting the most weight on the "previous prediction" feature. For a domain expert, this can be interpreted as saying: "Crystal Cube tends to reduce its prediction probability by 13% unless one of these other 39 features is present in this month's data." For the time series application we consider, this may be a helpful insight that might be otherwise unnoticed through the lens of local explanations.

8.4.5 Limitations

The work in this section does have its drawbacks. First, our method does not provide specific details about any individual data instance, which may make it less helpful for understanding outliers or edge cases of the black-box model's behavior. Similarly, we require enough examples on which to train a proxy, whereas local methods can be applied to just a single model prediction. Finally, there is no guarantee that a simple linear model can provide a faithful proxy for a given black-box model. On the other hand, we can empirically evaluate faithfulness on a held-out test set, whereas similar evaluations can be difficult for LIME (Atanasova et al., 2020).

We will now demonstrate the effectiveness of our proxy model approach on a second task, and defer a general discussion of the results of both tasks until §8.6.

8.5 Faithful and Plausible Explanations of Medical Code Predictions

8.5.1 Introduction

For a second application of our proxy model approach, we consider the widely-studied task of medical code prediction (Lima, Laender, and Ribeiro-Neto, 1998; Scheurwegs et al., 2016; Zhang et al., 2017). While ML methods have achieved predictive success on various versions of ICD clinical code assignment, the best-performing methods have been neural networks that are notoriously difficult to interpret. Mullenbach et al. (2018) introduced DR-CAML, a method designed to produce explainable predictions, which outperformed several baselines when evaluated by a clinical expert.

We reproduce this work and compare to our proxy model. We use a linear regression proxy model that learns to mimic the behavior of the trained DR-CAML model. We show that the proxy model is faithful to the original model and produces plausible explanations, as measured on clinician annotations of generated explanations.

8.5.2 Explainable prediction of medical codes

The work in this section closely follows that of Mullenbach et al. (2018). We use the same dataset of clinical texts and associated medical codes (described in § 8.5.4) and compare against their method: Description-Regularized Convolutional Attention for Multi-Label classification (DR-CAML). DR-CAML is a neural model that seeks to produce its own faithful explanations using a per-label attention mechanism that highlights n-grams in the input text that were correlated with the model's predictions. Because DR-CAML has over six million learned parameters, it does not fulfill simulatability or decomposability; a single parameter cannot be understood

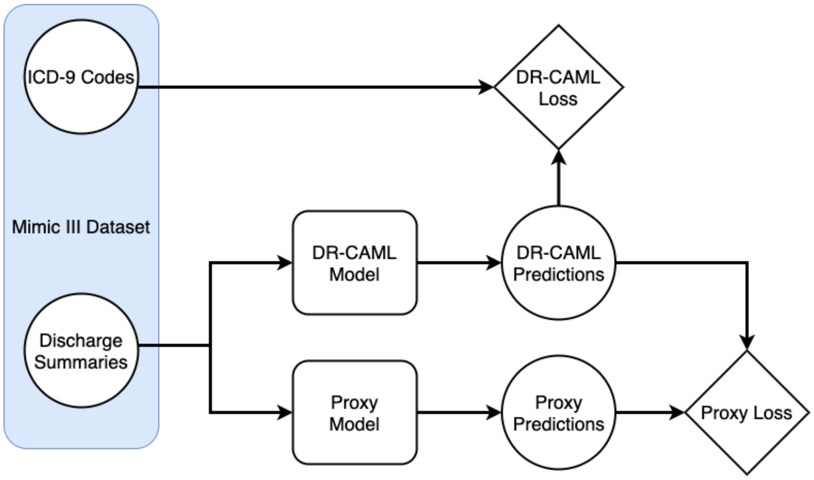


Figure 8-3. Relationship between trained DR-CAML model and proxy model. The proxy model is trained to predict DR-CAML's outputs, rather than the true ICD-9 codes. This optimizes the proxy model for faithfulness.

in any intuitive way. However, the attention mechanism allows for some insight into the model's decision-making, as it indicates which regions of the input text were given more weight in the prediction.

DR-CAML's use of attention to produce explanations has sparked discussion. Jain and Wallace (2019) showed that attention mechanisms can provide misleading explanations that are not faithful to the model's true reasoning. Wiegreffe and Pinter (2019) argued that while the explanations produced by attention may not always be faithful, they are often plausible. This discussion has continued in the interpretable ML literature, with methods demonstrating how attention mechanisms can be useful or deceptive (Zhong, Shao, and McKeown, 2019; Grimsley, Mayfield, and Bursten, 2020; Jain et al., 2020; Pruthi et al., 2020). Creating models that are both faithful and plausible remains a challenge.

8.5.3 Methods

Figure 8-3 gives a visual representation of the proxy model setup. For the medical code classification task, the original model (DR-CAML) is trained on the text of

	AUC		F1		P@n	
	Macro	Micro	Macro	Micro	8	15
Mullenbach et al. (2018)	0.895	0.986	0.088	0.539	0.709	0.561
Wiegreffe et al. (2019)	0.889	0.985	0.080	0.542	0.712	0.562
Ours (using released weights)	0.892	0.978	0.090	0.298	0.636	0.471
Ours (retrained)	0.628	0.884	0.001	0.024	0.042	0.027

Table 8-IV. Published predictive performance of CAML and our replicated results. Our experiments throughout the paper use the model with the released weights, which is closest to the published numbers (despite Micro F1).

discharge summaries and produces a probability for each of the 8,922 possible medical codes. We apply DR-CAML to the texts in MIMIC III (Johnson et al., 2016) and save its continuous-valued probabilities as the labels for our proxy model. We use a linear regression model trained on a bag-of-words representation of the clinical texts to predict these DR-CAML probabilities. We train 8,922 proxy models, one for each medical code in the dataset's labels. We implement our method using the linear `SGDRegressor` model from `sklearn` (Pedregosa et al., 2011), and apply a log transform to the model's probability outputs and train the proxy to minimize squared loss. We release the code for training and evaluating our method.

By applying our proxy model method to the DR-CAML model from Mullenbach et al. (2018), we enable an evaluation of both faithfulness and plausibility. We evaluate whether our model is faithful by seeing how closely its outputs match the predictions of DR-CAML. Because DR-CAML was designed to be interpretable using its attention mechanism, we can compare its explanations against those produced by our proxy. In the next two sections, we introduce our evaluation for the proxy model's faithfulness to the DR-CAML model and the plausibility of its explanations.

	Logistic	Proxy	DR-CAML
Macro AUC	0.596	0.901	0.906
Micro AUC	0.889	0.967	0.972
Macro F1	0.033	0.142	0.224
Micro F1	0.278	0.326	0.536
Prec @ 8	0.547	0.483	0.701
Prec @ 15	0.413	0.407	0.548

Table 8-V. Comparison of the logistic baseline, the proxy model, and DR-CAML to true ICD labels. Although the logistic model was trained for this specific task and the proxy model was not, the proxy model outperforms the baseline in terms of AUC and F1. The proxy model's outputs are unnormalized, which partially explains the gap between its F1 scores, which are computed with a threshold of 0.5, and its AUC scores, which are invariant to normalization. This lack of normalization may also explain the proxy model's low precision scores, as each code is predicted independently of the others.

8.5.4 Faithfulness evaluation

The MIMIC-III dataset contains anonymized English-language ICU patient records, including physiological measurements and clinical notes (Johnson et al., 2016). Following Mullenbach et al. (2018), we focus on discharge summaries which describe a patient's visit and are annotated with ICD-9 codes. There are 8,922 different ICD-9 codes that describe procedures and diagnoses that occurred during a patient's stay. The manual assignment of these codes to patient records are required by most U.S. healthcare payers (Topaz, Shafran-Topaz, and Bowles, 2013). We duplicate the experimental setup of Mullenbach et al. (2018) which uses the text of the discharge summaries as input to the DR-CAML model, which then is trained to predict all ICD-9 codes associated with that document. After applying their pre-processing code to tokenize the text, the dataset contains 47,724 discharge summaries divided into training, validation, and test splits.

Our proxy model is the combination of 8,922 linear regression models trained to predict DR-CAML's log probability for each ICD-9 code. After a brief grid search

	Regression			Classification			
				AUC		F1	
Model	Spearman	Pearson	Kendall	Macro	Micro	Macro	Micro
Logistic	0.036	-0.195	-0.135	0.734	0.936	0.012	0.353
Proxy	0.498	0.794	0.608	0.980	0.995	0.052	0.416

Table 8-VI. Comparison of the logistic baseline and the proxy model to the DR-CAML predictions. For the F1 evaluation, we threshold the unnormalized proxy outputs at 0.5. The logistic model was trained to predict the ICD codes; the proxy model to predict DR-CAML's predictions. As expected, the proxy model dramatically outperforms the logistic baseline in terms of faithfulness to the DR-CAML model.

on the validation set, we chose to apply L1 regularization with $\alpha = 0.0001$ for each regression. To establish that this collection of linear regressions is faithful to the trained DR-CAML model, we want to show that it makes similar predictions across all ICD-9 codes on held-out data. Recall from Figure 8-3 that the proxy is trained not to predict the true ICD-9 codes but to output the same label probabilities as DR-CAML. In fact, the proxy model never sees the true ICD-9 codes. We evaluate faithfulness by comparing the outputs of DR-CAML and the proxy model on the held-out test set. If the two systems produced identical outputs on held-out data, we would say that the proxy was perfectly faithful. We make this comparison in three different ways – first using regression metrics that compare the continuous outputs of the two models, then using classification metrics with binarized DR-CAML predictions, and finally by using the proxy model's outputs as predictions for the true ICD-9 codes. For all these comparisons, we use a logistic regression baseline that is trained to directly predict the ICD-9 codes without knowledge of DR-CAML's predictions. While we would expect the logistic baseline's predictions to be somewhat correlated with those of DR-CAML, we would not expect the baseline to be faithful.

Our first evaluation uses regression metrics that assess the correlation between the proxy's predictions and DR-CAML's predicted probabilities. We use Spearman and Pearson correlation coefficients and the non-parametric Kendall Tau rank correlation. These metrics range from -1 to 1 with 1 indicating perfect faithfulness. Regression results are on the left side of Table 8-VI.

Our second evaluation treats DR-CAML's predictions as binary labels based on whether they exceed the threshold used by Mullenbach et al. (2018) to compute F1 scores. We then evaluate the faithfulness of our proxy model by treating its outputs as unnormalized probabilities and using classification metrics such as F1 score. These metrics range from 0 to 1, where perfectly faithful predictions would have 1.0 AUC and F1 scores. The proxy model is considered faithful if it correctly predicts whether DR-CAML will make a binary prediction. We again use the logistic regression baseline. Classification results are on the right side of Table 8-VI.

Finally, we use the proxy model's predictions to predict the ground-truth ICD code labels and compare its predictive performance against that of DR-CAML in Table 8-V. While the proxy model was not trained using these labels, we can use its predictions as unnormalized probabilities for these codes. By comparing against the logistic regression baseline (a linear model of equal complexity), we can see whether our training setup allows the proxy model to learn a better predictor.

Our results show that the proxy model is quite faithful to the DR-CAML model. Compared to the logistic regression baseline, the proxy model is dramatically better on all metrics in Table 8-VI. Combining the results from Tables 8-VI and 8-V we can see that on AUC metrics, the proxy model is closer to the DR-CAML predictions than DR-CAML is to the ground-truth labels. The proxy model also outperforms the logistic regression baseline in the classification metrics (AUC and F1), indicating that the proxy model is more faithful to the DR-CAML predictions. In Table 8-V, we see a large gap between its performance on the AUC metrics and the F1 and precision metrics. This is likely because the outputs of the proxy model are not normalized to be valid probabilities and AUC is invariant to normalization, unlike

934.1: “Foreign body in main bronchus”		
Mullenbach et al. (2018)		
CAML	(HI)	… *line placed bronchoscopy performed showing* **large mucus plug on** *the left on transfer to* …
Cosine		… *also needed medication to help* **your body maintain your** *blood pressure after receiving iv* …
CNN		… *found to have a large* **lll lingular pneumonia on** *chest x ray he was* …
Logistic		… *impression confluent consolidation involving nearly* **the entire left lung** *with either bronchocentric or vascular* …
Ours		
DR-CAML	0.38	… *line placed bronchoscopy performed showing* **large mucus plug on** *the left on transfer to* …
Logistic	0.28	… *tube down your throat to* **help you breathe you** *also needed medication to help* …
Proxy	0.38	… *a line placed bronchoscopy performed* **showing large mucus plug** *on the left on transfer* …

Table 8-VII. Comparison of the clinical evaluation from Mullenbach et al. (2018) with our plausibility evaluation. The first four systems for each example are directly copied from Table 1 of Mullenbach et al. (2018). The (HI) and (I) labels in the second column indicate whether the clinician labeled those explanations as Highly Informative or Informative. The three systems below the dotted line are from our evaluation, for which the second column indicates the probability output of our plausibility classifier. Table 8-VIII shows two additional comparisons. For these comparisons, the proxy and DR-CAML produce almost identical explanations; additional comparisons between DR-CAML and the proxy are shown in Table 8-IX.

F1 and precision.

Rudin (2019) critiques post hoc methods in general, arguing that “if we cannot know for certain whether our [post hoc] explanation is [faithful], we cannot know whether to trust either the explanation or the original model.” Because no post hoc method can ever be perfectly faithful to an original model, we believe our approach to explicitly measuring faithfulness provides a useful approach for understanding whether the proxy is “faithful enough” for a given application. It also allows for a prediction-specific analysis – if we wish to use the proxy model to explain a high-stakes prediction made by DR-CAML, we can first check to see whether the two models agree upon that specific prediction.

In applications where explainability is essential, our proxy model could be used as a more interpretable replacement for a high-performing black-box model. In such a case, a domain expert might care less about the evaluation of faithfulness in Table 8-VI and more about the ground-truth predictive performance evaluated in Table 8-V. We leave for future work the challenge of whether a proxy model produced by our method could be fine-tuned to improve its performance at predicting ground-truth ICD codes.

8.5.4.1 Reproducing CAML predictive performance

The trained DR-CAML model released by Mullenbach et al. (2018) produced predictions that matched their published F1 and ROC scores. However, we were unable to precisely replicate the outputs of the CAML model. Table 8-IV shows the scores published by Mullenbach et al. (2018) as well as those for a CAML reimplementation done by Wiegreffe et al. (2019). We include the scores we observe using the model weights released on GitHub as well as the scores for a model we retrained from scratch. We use the released model instead of the retrained model as its performance is much closer to published work.

8.5.5 Plausibility Evaluation

Explanations are considered plausible if they can be reasoned about by a person. Thus, evaluating plausibility is typically more difficult than faithfulness, because it requires input from annotators (Herman, 2017). Furthermore, an explanation that is plausible to a domain expert may not be plausible to a layperson. Mullenbach et al. (2018) evaluated the plausibility of CAML's explanations by collecting annotations from a clinician. Wiegreffe and Pinter (2019) argued that the attention mechanism of CAML and DR-CAML generally provide plausible explanations, even if they at times are not faithful to the model's internal decision-making. For each model they considered, they extracted an explanation in the form of a 14-token subsequence

442.84: “Aneurysm of other visceral artery”

Mullenbach et al. (2018)		
CAML	(I)	*… and gelfoam embolization of right* **hepatic artery branch pseudoaneurysm** *coil embolization of the gastroduodenal …*
Cosine		*… coil embolization of the gastroduodenal* **artery history of present** *illness the pt is a …*
CNN		*… foley for hemodynamic monitoring and* **serial hematocrits angio was** *performed and his gda was …*
Logistic	(I)	*… and gelfoam embolization of right* **hepatic artery branch pseudoaneurysm** *coil embolization of the gastroduodenal …*
Ours		
DR-CAML	0.55	*… gelfoam embolization of right hepatic* **artery branch pseudoaneurysm coil** *embolization of the gastroduodenal artery …*
Logistic	0.57	*… biliary stents hx cbd r* **colonic fistula r colectomy** *partial l nephrectomy for renal …*
Proxy	0.55	*… embolization of right hepatic artery* **branch pseudoaneurysm coil embolization** *of the gastroduodenal artery history …*

428.20: “Systolic heart failure, unspecified”

Mullenbach et al. (2018)		
CAML		*… no mitral valve prolapse moderate* **to severe mitral regurgitation** *is seen the tricuspid valve …*
Cosine		*… is seen the estimated pulmonary* **artery systolic pressure is** *normal there is no pericardial …*
CNN		*… and suggested starting hydralazine imdur* **continue aspirin arg admitted** *at baseline cr appears patient …*
Logistic	(HI)	*… anticoagulation monitored on tele pump* **systolic dysfunction with ef** *of seen on recent echo …*
Ours		
DR-CAML	0.38	*… anticoagulation monitored on tele pump* **systolic dysfunction with ef** *of seen on recent echo …*
Logistic	0.36	*… seen the mitral valve leaflets* **are mildly thickened there** *is nomitral valve prolapse …*
Proxy	0.38	*… anticoagulation monitored on tele pump* **systolic dysfunction with ef** *of seen on recent echo …*

Table 8-VIII. Two additional comparisons of the clinical evaluation from Mullenbach et al. (2018) with our plausibility evaluation. See Table 8-VII for details.

296.20: “Major depressive affective disorder, single episode, unspecified”

DR-CAML	0.47	*...diagnosis overdose of medications narcotics* **benzodiazepine suicide attempt chronic** *migraine headaches depression stage iv ...*
Proxy	0.33	*... up from the medications you* **were evaluated by psychiatry** *and will be transferred to ...*

455.0: “Internal hemorrhoids without mention of complication”

DR-CAML	0.38	*... and she then underwent a* **colonoscopy with gi that** *also did not detect evidence ...*
Proxy	0.52	*... past medical history diverticular disease* **diverticulitis sbo anxiety hemorrhoids** *past surgical history s p ...*

244.9: “Unspecified acquired hypothyroidism”

DR-CAML	0.48	*... baseline mental status oriented x1* **hyperlipidemia hypertension hypothyroidism s** *p cataract surgery social history ...*
Proxy	0.29	*... twice daily lasix mg po* **tid levothyroxine mcg daily** *discharge medications none discharge disposition ...*

592.0 : “Calculus of kidney”

DR-CAML	0.30	*... if you develop any of* **these symtpoms please call** *the office or go to ...*
Proxy	0.46	*... the colon gastroesophageal reflux asthma* **irritable bowel syndrome gastroparesis** *osteoporosis anxiety and or depression ...*

Table 8-IX. Differing explanations and classifier scores between DR-CAML and the proxy.

taken from the discharge summary. The clinician read all (anonymized) four explanations and the corresponding ICD code and rated each explanation as either “informative” or not. CAML was rated slightly more informative than logistic regression and CNN baselines. Tables 8-VII and 8-VIII shows explanations produced by Mullenbach et al. (2018)’s methods as well as the ones we consider in this work.

The format of Mullenbach et al. (2018)’s plausibility evaluation does not easily lend itself to replication. While the authors shared their annotations with us, missing metadata prevented a direct reproduction of their analysis. Additionally, since the clinical annotator considered explanations in a comparative setting, we cannot easily add our proxy model as another method using the same annotations. Therefore, we replicate this evaluation by using a classifier to predict synthetic labels as to whether the clinical domain expert *would have* labeled our models' explanations as plausible. Using BioWordVec embeddings released by Zhang et al. (2019), the text of the ICD-9 code description, and the 14-gram explanation produced by each model from Mullenbach et al. (2018), we train a classifier that predicts whether an explanation would have been rated as informative.[7] This annotation classifier achieves an accuracy of 67.2% and an AUC score of 0.726 on held-out explanations, indicating it is a useful but noisy stand-in for the clinician. Additional details are in §8.5.5.2.

Model	Score	Interval
Logistic	35	(31, 49)
Cosine	38	(32, 51)
CNN	42	(33, 52)
CAML	44	(33, 52)
DR-CAML	48	(34, 53)
Proxy	52	(34, 54)

Table 8-X. Binary plausibility evaluation using classifier annotations. We collapse 'Highly Informative' and 'Informative' from Mullenbach et al. (2018) to a single positive class. The Score column is out of 99; we threshold so the same total proportion of explanations are deemed plausible. The Interval column shows a 95% bootstrap interval from sampling 1000 labels from the classifier probabilities.

To conduct our plausibility evaluation, we first use or reproduce the baseline methods from Mullenbach et al. (2018). Each model, including the proxy, produces a 14-token explanation from the discharge summary by first finding the 4-gram with the largest *average feature importance* and then including five tokens on either side of the 4-gram. The logistic regression baseline is the same as in § 8.5.4, where feature importance is computed using the coefficients of the logistic model. The proxy model's explanations are computed in the same manner, finding the 4-gram with the largest average coefficient weights. For CAML, DR-CAML, and the CNN models, we use the code released by Mullenbach et al. (2018) to extract explanations. The CNN baseline primarily differs from CAML in that it does not use an attention mechanism. Finally, we reimplement their Cosine baseline which picks the 4-gram with the highest cosine similarity to the ICD-9 code description text.

[7] We collapse the "informative" and "highly informative" labels into a single positive class.

Model	Theirs	Ours	E1	E2	Full
Logistic	41	43	47	49	35
Cosine	48	48	41	40	38
CNN	36	46	51	47	42
CAML	46	54	47	43	44
DR-CAML	–	–	45	44	48

Table 8-XI. Plausibility evaluations and comparison to Mullenbach et al. (2018). The Theirs column shows the published numbers; Ours shows our best attempt at matching the clinical evaluation to the trained models. While the numbers change dramatically, the ordering only changes by two swaps. The clinical evaluation did not include DR-CAML. E1 and E2 show the results with predicted plausibility labels under the two evaluation settings described in 8.5.5.2. Full duplicates the results from Table 8-X for comparison.

We extract the model's explanations for the same[8] discharge summaries as were evaluated by Mullenbach et al. (2018). For each explanation, we use the annotation classifier described above to predict the probability that each explanation would have been labeled as informative. If we set the classifier threshold such that 45% of explanations are rated as informative (matching the proportion from the original annotations), we get the results in the Score column of Table 8-X. The proxy model produces the largest number of informative explanations according to our classifier; however, the classifier's inaccuracy introduces uncertainty. Rather than thresholding the outputs of the annotation classifier, we can use its probability outputs to sample a set of informative labels for each explanation. We sample 1000 such sets of labels and report the 95% confidence interval for each model's score in the Interval column of Table 8-X. Accounting for this uncertainty dramatically reduces the differences between the methods. Because 95% of all classified plausibility probabilities are between 24.1% and 58.1%, these intervals skew towards lower scores. Despite the inherent uncertainty involved in extrapolating plausible scores from a fixed set of clinical annotations, our evaluations suggest that the proxy model produces explanations that are at least as plausible as those of DR-CAML.

[8] Using the 99 (of 100) discharge summaries that could be uniquely identified. See §8.5.5.1 for details.

Tables 8-VII and 8-VIII show that for the three examples considered in Mullenbach et al. (2018), DR-CAML and our proxy model produce very similar explanations. This is perhaps surprising because DR-CAML extracts explanations using its attention mechanism, whereas the proxy model uses unigram feature importance values that do not vary between examples. For these examples, it appears that the proxy is faithful both in the predictions it makes and how it makes those predictions. Table 8-IX shows three examples where the proxy and DR-CAML diverge the most. These rare cases highlight two benefits of the proxy model. First, its feature importance weights are *global* across all predictions, providing an aggregate representation of the proxy's behavior. Second, the approach for extracting proxy explanation *n*-grams is transparent and simulatible; it is just the average of *n* feature weights. These factors may be particularly appealing in cases where explainability is paramount.

8.5.5.1 Reproducing plausibility scores

The clinical plausibility annotations provided to us by the authors of Mullenbach et al. (2018) contains the text explanations and their corresponding annotations, but is missing the crucial metadata of which models produced which explanations. The metadata also did not indicate from which specific discharge summary the texts were derived; while the text explanations were uniquely identifying for all but one of the 100 examples. For that one example, because some patients had multiple documents sometimes containing duplicated segments of text, there were three discharge summaries from which the explanations could have been drawn. We thus excluded this example from our analyses. To replicate their analysis as best as we could, we retrained or reimplemented their logistic regression, vanilla CNN, and cosine similarity methods. We then looked at the attention or feature importance weights for each trained model and the text explanations that had been annotated, and assigned each model the text explanation for which it provided the highest weight. This assignment did not perfectly align with past work: there were six cases (out of 99) where a text explanation was “chosen” by more models than times it appeared as an option. Ignoring that issue and then simply aggregating the Informative and Highly Informative clinician annotations, we obtained the plausibility scores in the Ours column of Table 8-XI. The Theirs column shows the published numbers from Mullenbach et al. (2018). While the numbers change substantially, the ordering is relatively stable with only two swaps: CAML and Cosine, and Logistic and CNN. The other columns of the table are described below.

8.5.5.2 Plausibility annotation classifier

To evaluate the plausibility of our proxy model’s explanations, we trained a classifier to predict whether an explanation would have been labeled as plausible by the clinical domain expert. We treat this as a binary classification task by grouping the "Informative" and "Highly Informative" annotations as a single "plausible" label. Conscious of the fact that we have only 99 examples with four text explanations each, we use two approaches with which to train and evaluate our classifier. The first used leave-one-out cross validation at the example level, such that the classifier was trained on 98 examples at a time and then evaluated on the remaining one. We refer to this evaluation as "E1" in Table 8-XI. The second also used leave-on-out cross validation but at the explanation level; we held out a single text explanation, trained on all other explanations across all examples, and then evaluated on the held-out explanation. When an explanation appeared more than once in a single example, we made sure to remove its duplicates from the training data for predicting that explanation. We refer to this evaluation as "E2" in Table 8-XI.

The trained model is a simple logistic regression classifier trained on a fastText embedding of both the explanation and the target ICD-9 code description. Using the BioWordVec embeddings released by Zhang et al. (2019), we embed each both the explanation and code description into a 200-dimensional vector, concatenate the two vectors, and pass it to the logistic regression. In the E1 evaluation, the model achieves an accuracy of 60.6% and an ROC AUC score of .640. In the E2 evaluation, that increases to an accuracy of 67.2% and an AUC score of .726, indicating that the additional within-example explanations substantially help the classifier.

When using these classifiers to label the explanations generated by each model instead of the plausibility scores derived in 8.5.5.1, we get the results shown in columns E1 and E2 of Table 8-XI.

Finally, we retrain our final classifier on all the explanations, leaving none held out. Rather than using our classifier to evaluate the explanations that were actually shown to the clinician, we instead use our (re-)implementation of the four models to extract an explanation from each of the 99 discharge summaries. These explanations thus may or may not appear in the training data for the classifier. For the Full evaluation we are not worried about the classifier overfitting, as the classifier functions as a direct replacement for the clinician who produced the training data. The results of this analysis are the numbers shown in Table 8-X in § 8.5.5, reproduced in Table 8-XI in the "Full" column. The Logistic model does much worse on the Full evaluation than in either E1 or E2. This may be because the explanations selected by the trained model were worse than those selected by the model which was used for the original clinical evaluation.

8.5.6 Limitations

The work in this section has several limitations that are left for future work. Though the task of medical code prediction has important implications and has been widely studied in interpretability research, we only consider this single task on a single English-language dataset. We believe this proxy model approach is generally applicable as a post hoc interpretability method for arbitrary models, but this must be further studied on new datasets and different trained models. It is possible that in some domains, trained models might be more difficult to mimic than DR-CAML. If so, the application may require a trade-off between a less restrictive proxy model class and a less faithful proxy.

Our evaluation is also limited in that it only considers a single form of explanation: n-grams extracted via feature importances or attention weights. Recent work has explored alternate formulations for a quality explanation (Barocas, Selbst, and Raghavan, 2020); some formulations may be more or less accommodating of our proxy model method. Our plausibility evaluations rely heavily on a single set of expert annotations from which we extrapolate using a classifier. To demonstrate that our method can reliably provide both plausible and faithful explanations, additional evaluations must collect new plausibility annotations or build off of existing resources (DeYoung et al., 2020).

8.6 Discussion

We have introduced a method for post hoc explanations that is designed to be interpretable, plausible and globally consistent while maintaining faithfulness to the trained model. By constraining the proxy to a class of models that is decomposible, simulatible, and algorithmically transparent, our optimization for faithfulness gives us a clear way to evaluate several dimensions of interpretability.

A key benefit of our method is its simplicity and wide applicability. Even for a proprietary trained model for which the learned parameters are unknown, a proxy can be trained as long as we have a dataset that includes the trained model's predictions. Our approach has the additional benefit of producing a standalone proxy model that can provide *global* feature explanations. Depending on the gap in predictive performance between the proxy and original model, a skeptic of post hoc methods (e.g. Rudin (2019)) might prefer to discard the original model altogether and simply use the proxy's predictions, for which its explanations are faithful by design. In Chapter 9, we will take this approach and conduct an analysis with both the original trained model(s) and with proxy models.

Our experimental evaluation on two datasets representing complex real-world phenomena demonstrates that our method can provide a holistic explanation of the trained model's behavior across an entire dataset. We have shown that our method can be several orders of magnitude faster than existing, local methods of explanation that use sampling or gradients. In both domains we considered, we found that a linear proxy with tens or tens of thousands of parameters can perform comparably to an existing model with many millions.

While local methods such as LIME provide inconsistent (local) explanations across instances, our approach provides a globally-consistent explanations. Once trained, the proxy model naturally aggregates the importance of features into a holistic and interpretable explanation. This is particularly true in our L_1-regularized case, where sparsity reduces the number of nonzero features. This provides a proxy that is simulatible and decomposible (Lipton, 2018) – a domain expert can 'decompose' the coefficients of the linear proxy and understand (simulate) how they are individually combine to produce the prediction. For an expert who knows the features and their real-world relevance, such a holistic explanation allows them to easily compare their own judgment against the proxy model's explanation.

Our results also suggest opportunities for future work in Explainable AI. A proxy model can provide fast, global explanations that could inform which examples require additional local explanation, by checking where the black-box model's predictions differ from those of the proxy. Proxy model importance scores could similarly inform LIME's sampling procedure by suggesting which features need to be perturbed to learn a reliable surrogate explanation. Future work should also explore how experts in different domains respond to explanations from a proxy model compared to those of other methods.

As the ML community continues to explore new directions for interpretable methods, definitions of desiderata may continue to evolve. Such criteria will always depend on the domain experts who turn to an ML method for decision support. Interpretable ML methods should clearly define how they expect to satisfy such criteria; by designing for plausibility and transparency and optimizing for faithfulness, our proposed method is broadly applicable.

Chapter 9

Measurement Error Correction of Proxy Models for Gender and Toxic Language

9.1 Introduction

In recent years, several analyses have repeatedly found that women are subject to a disproportionate amount of abuse on Twitter (Usher, Holcomb, and Littman, 2018; Rego, 2018; Akhtar and Morrison, 2019). Delisle et al. (2018) analyzed 157k tweets annotated for abusive language by a large crowdsourced effort involving 4.5k volunteers, again finding widespread abuse targeted at women politicians and journalists from the United States and United Kingdom. Such an extensive crowdsourcing project has great potential for helping understand how public figures are in particular targeted. But with upwards of 500M tweets posted every day, manual annotations cannot possibly scale to provide data on the ongoing phenomenon of abusive language on Twitter and how it intersects with gender. If we want to study this issue at the scale of millions of tweets, we need ML classifiers that can automatically infer gender and toxicity – but we also must be careful about what assumptions we are making when we combine imperfect classifiers into an analysis of complex issues. This chapter seeks to tie together several threads from this thesis to explore the relationship between gender and abusive language on

Twitter.

We will formalize this research question by asking: are men and women equally likely to receive a toxic response to a tweet they post? This seemingly simple question requires defining gender and toxic language. If we had an infinite supply of people willing to annotate tweets, we could follow Delisle et al. (2018) by carefully writing down definitions and guidelines and measuring the inherent uncertainty of these subjective categories by how reliably the annotators agree with one another. In the absence of such resources, we will turn to automated ML classifiers, like the ones introduced in Chapter 4. We will use our `Demographer` classifier to label users as men, women, or organizations. We will use the `Perspective` classifier, which we discuss in §9.3, to label tweets for whether they contain toxic language.

Naively combining the predicted outputs of `Demographer` and `Perspective` to label the same tweets is implicitly making the assumption that the two classifiers are independent of each other. While we do not know what data Perspective was trained on, such data is likely influenced by the same relationship of gender and toxicity that we wish to study. We would prefer to consider a joint model of gender and toxicity given the same set of features, to understand how they interact. Unfortunately, the trained model and data of Perspective were not released. We will instead draw on our proxy model methodology from Chapter 8 to produce a joint model that we can use for our analysis. Finally, we will use the measurement error framework from Chapter 5 and sensitivity analyses from Chapter 6. While these analyses are still preliminary work, combining these different approaches will allow us to account for the assumptions we are making in our approach and help us understand the uncertainty of our analyses. Before we discuss the details of these classifiers and the setup of our joint proxy model, we will introduce our dataset.

9.2 Data Collection

The goal of this analysis is to study the widespread phenomenon of toxic language on Twitter and its intersection with gender. We chose to collect data from the Twitter 1% (“Spritzer”) feed, which contains 1% of all tweets, sampled deterministically. While this feed has known biases and is not representative of the full Twitter population (Morstatter et al., 2013), it is an easy and often the best way to cheaply collect millions of tweets from a diverse set of users.

To understand the behavior of how Twitter users receive toxic replies, we first need a dataset that matches tweets to their responses. The Twitter API does not make it possible to search for all replies to a given tweet, but the tweet metadata of replies does provide the ID of the replied-to tweet. To leverage this, we first find replies in the 1% feed and then scrape the tweets to which those replies reply. Twitter allows for two ways to respond to other users – replying and quote retweeting; we will use graphical terminology and refer to both replies and quote retweets as ”children” and the tweets to which they respond as “parents.” A single parent can have many children, but almost[1] all children have only a single parent.

We begin our data collection by finding all 24.4M children in the 1% feed from midnight December 1, 2020 to midnight December 15, 2020. We then find the IDs of 18.0M unique parents responded to by the children, excluding 200k parents that were already collected as children. We then turn to the Twitter API to scrape all parent tweets, of which we could retrieve 13.3M (73.9%). Those 13.3M parents had 18.3M children, of which we could scrape 14.9M (81.4%). Tweets are unable to be scraped if they are deleted or if the author of the tweet has made their account private. The missingness introduced deleted tweets is a common problem

[1] Twitter allows you to reply to one tweet with a quote retweet of another tweet, which makes it possible for a child to have at most two parents. There are only 150k instances of this in our entire dataset.

in computational social science research (Wu, Rizoiu, and Xie, 2020). Future work has the opportunity to address this issue with a missing data framework that can account for the fact that tweets are certainly not deleted completely at random (Malinsky, Shpitser, and Tchetgen, 2021).

We now take several filtering steps to prepare our data for analysis and to strengthen the validity of our results. Because Twitter is to some extent dominated by "power users" who have the most influence (Kwak et al., 2010), we first filter out parent tweets with a large number of children. From across all (before scraping) 18M parents and 24.4M children, while 89% of parents have only one child, the extremely skewed distribution (one parent has a maximum of 18k children) brings the mean up to 1.35. To focus our analysis on the replies received by "everyday" users, we remove the 18k (0.01%) parents with more than 36 children and the 1.8M total children that replied to those tweets.

To ensure that `Perspective` could classify the children as toxic or not, we removed all children that had no text or had text in a language the classifier does not support (see §9.3). We then split our data into train and test sets, using the halfway cutoff of midnight, December 8. To enable a fair experimental setup for evaluating our demographic classifier proxy model (see §9.6), we identify all users who authored parent tweets in both the train and test sets and remove all their tweets from the test set.

After these filtering steps, we match the remaining parents to the remaining children, adding a pair to our final dataset only if both tweets have met all inclusion criteria above. This results in a dataset of 4.5M parent-child pairs, with 3.38M and 1.17 pairs in the train and test sets, respectively. To supplement this dataset with labels on gender and toxic language, we will introduce our automated classifiers.

9.3 Demographer and Perspective Classifiers

In §4.4 we introduced several classifiers that have been released as part of the `Demographer` package. For this chapter, we stack the organization and gender classifiers from Wood-Doughty, Mahajan, and Dredze (2018) and Wood-Doughty et al. (2018) to filter out users labeled as organizations, and label the remainder as either men or women. As discussed in §4.2, treating gender as binary is a significant limitation, but one we cannot overcome with our available tools. This approach also assumes away the existence of partially or fully automated accounts such as spambots or malicious political actors (Davis et al., 2016; Broniatowski et al., 2018). Both of these classifiers use a neural network to learn a feature representation of a Twitter user's name; the organization classifier also uses profile features such as the number of followers and the text of the profile description.

Perspective is a classifier that labels text as 'toxic' or not, using the definition from Borkan et al. (2019) of anything "rude, disrespectful, or unreasonable that would make someone want to leave a conversation." It has been in widespread use for several years on websites including the New York Times (Etim, 2017). Perspective is not Twitter-specific; it only uses the text of the tweet and not any metadata. In addition to English, it supports Spanish, Portuguese, French, German, Italian, and Russian. As noted in §9.2, we discard any tweets that either have no text or text not labeled as Twitter as containing one of these languages. Unlike our Demographer package, Perspective's code and training data have not been released publicly, so we do not know the specifics of how it was trained or how those details may contribute to its potential issues of how it labels toxicity, such as a bias against African-American English or a vulnerability to adversarial attacks (Sap et al., 2019; Hosseini et al., 2017).

We run both classifiers over all 4.5M parent-child tweet pairs in our dataset, inferring organization and gender labels for the parents and toxicity labels for the children. If we simply count up the classifications of these models, we find that 54.2% of train parent authors and 53.4% of test parent authors are labeled as men, and 10.9% of train children and 11.2% of test children are labeled as toxic. The discrepancy in the gender classifications is large enough to suggest it may be a result of our preprocessing steps, but we leave this concern for future work. The simplest measure of whether parent author gender is correlated with the toxicity of children is to compute the difference in conditional probabilities. We refer to predicted toxicity as T^* with $T^* = 1$ indicating a comment was labeled as toxic; $G^* = 1$ similarly indicates a user was labeled as a man. Then our estimate of the correlation between gender and toxicity can be computed as:

$$\tau = p\left(T^* = 1 | G^* = 0\right) - p\left(T^* = 1 | G^* = 1\right) \tag{9.1}$$

In words, how much more likely is a woman to receive a toxic comment than a man? Counting our inferred labels across the 1.17M tweet pairs in the test set, we get an estimate of $\tau = -0.0013$, or essentially[2] no difference. However, as we have argued throughout this thesis, treating the outputs of these noisy classifiers as representative of ground truth concepts of gender and toxicity implicitly makes assumptions that are unlikely to be justified. To relax these assumptions and adjust for the classifiers' errors in our measurement error framework, we first need to calculate their accuracy on held-out datasets with ground truth labels for gender and toxic language.

[2] This is in fact significant at $p = 0.024$ according to a Fisher Exact test, but at large sample sizes p-values are often significant even for trivial differences (Lantz, 2013).

9.4 Validation Datasets for Classifier Accuracy

9.4.1 Demographics Validation Dataset

To validate the accuracy of our Demographer gender classifier, we use a dataset released by Preoţiuc-Pietro and Ungar (2018) of users who explicitly report their gender identity in a survey. We used this same dataset in §4.4.3 as our ground-truth dataset for predicting race and ethnicity. The original dataset contains 4.1k users; we were able to scrape tweets for 3.1k users. Because our organization classifier appears to have extremely high accuracy, it simplifies our analyses to treat its predictions as ground truth. If we sought to explicitly explore trends in how *organizations* receive toxic replies on Twitter, we would want to model and correct for the errors of this classifier as well. Of the 3.0k users who were labeled as individuals by our organization classifier, 33.8% identify as men compared to 39.3% that are labeled as men by Demographer. This gives an estimate of the model's accuracy at 83.3%.

9.4.2 Toxicity Validation Dataset

To validate the accuracy of Perspective, we use a dataset released by Founta et al. (2018). The dataset was constructed by selecting tweets from the Twitter 1% feed and then using an iterative process to collect and refine annotations from workers on the CrowdFlower platform. Crowdsourcing annotations provides a powerful way to label large datasets for NLP tasks (Snow et al., 2008), but require carefully defining subjective terms and clearly communicating with the workers conducting the annotations (Hansson et al., 2016). There are also widespread concerns about the labor rights of crowdworkers, as they have few protections from exploitative practices of both researchers and the platforms that employ them (Irani and Silberman, 2013; Silberman et al., 2018).

The original dataset provided over 80k tweets, of which we could scrape 42k.

To create a paired parent-child test set that we could compare against our collected dataset, we found all tweets that were “children” (either replies or quote retweets) and tried to download the corresponding parents. We were able to match and scrape 7.8k such pairs, for which 6.1k had the parent labeled as an individual by our organization classifier. Of those, 62% of parents were labeled as men. For the unpaired 42k tweets, 18% were toxic according to the ground truth annotations and 20% were labeled as toxic by Perspective, for an accuracy of 93.6%. For the 6.1k tweets for our final paired test set, 23% had ground-truth toxic labels and 26% were predicted as toxic by Perspective, for an accuracy of 91.5%. Removing pairs for which the parent was classified as an organization did not affect the test accuracy.

9.4.3 Construct Validity and Validation Datasets

Our measurement error approach, which we introduced in §2.2.2 and have used throughout this thesis, allows us to avoid making a naive assumption that the predictions of a ML classifier are perfect. While we can maintain *internal validity* with this framework, we still must make assumptions about the construct validity of our methodology. In particular, while our validation datasets give us a means to measure the accuracy of the Demographer and Perspective classifiers, we cannot evaluate how well the crowdsourced or self-reported labels in these datasets correspond to the higher-order concepts of gender and toxicity.

Both gender and toxic language are complex socially-constructed concepts that depend on sociocultural context. While an evaluation of the construct validity of our measurements for these concepts is outside the scope of this chapter, we can contextualize our analyses with similar research. Hate speech in online environments has been studied in terms of its measures and construct validity (Saha, Chandrasekharan, and De Choudhury, 2019; Samory et al., 2021). Researchers have also studied the construct validity of different measures of self-reported and perceived gender

(Mensinger, Bonifazi, and LaRosa, 2007; Kachel, Steffens, and Niedlich, 2016), though not in an social media context. While these investigations of validity cannot directly inform our analysis of gender and toxic language, they provide guidance for how our study could be strengthened. In particular, we could follow Broniatowski and Tucker (2017) and use different measures of toxicity (e.g. lexicons for offensive language (Arco et al., 2019)) to assess convergent validity of the crowdsourced labels in our validation dataset. As Demographer predicts perceived gender whereas our validation dataset contains self-reported gender, we could similarly evaluate convergent validity by investigating other measures of perceived gender (Mensinger, Bonifazi, and LaRosa, 2007). We could also evaluate our toxicity classifier or crowdsourced labels for discriminative validity by considering the topic of individual posts. While we do not explicitly control for post content, it is likely that tweets concerning certain topics (e.g. politics) are more likely to receive toxic replies – establishing discriminant validity would require showing that Perspective is not simply learning to detect topic and instead focuses on features indicative of toxic language.

9.5 Measurement Error Correction

Our test set evaluation demonstrates that the labels predicted by our classifiers are imperfect. We can formalize these predicted gender and toxicity as noisy proxies for the labels that would have been generated if we could have crowdsourced and surveyed ground-truth labels. This allows us to use our measurement error framework from Chapter 5 to correct for the disparity between predicted labels and ground truth and produce an estimate of the true relationship between gender and toxicity. We first must extend the derivations we showed in Chapter 2 to our setting with two variables that are mismeasured. This derivation is not for a causal question per se, but rather the joint distribution if our classifiers had been perfect.

Non-differential error assumptions, discussed in §6.6.1, are important to this analysis. This assumption states that a classifier's error rate is independent of all other variables in the analysis. For the toxicity classifier, this assumption implies that Perspective is equally accurate at classifying tweets that are posted in response to men and women. There is reason to believe this assumption is violated in practice, based on the way that toxicity annotations are collected and models are trained (Binns et al., 2017). For the gender classifier, however, we may be more willing to assume non-differential errors, as the ability to classify a given user likely does not depend on the replies received by that user.

We will begin by deriving the measurement error correction for the gender classifier, making the assumption of non-differential error. For convenience of notation, define the two gender error rates as:

$$\delta_g = p\left(G^* = g' \middle| G = g\right) \tag{9.2}$$

$$\epsilon_g = p\left(G^* = g \middle| G = g'\right) \tag{9.3}$$

Where g and g' are two different labels for gender. Note that $\delta_{g=0} = \epsilon_{g=1}$; there are only two error rates for gender under the non-differential assumption. We can correct for gender measurement error:

$$p\left(G = g, T^* = t\right) = \frac{(1-\delta_g)p\left(G^* = g, T^* = t\right) + \delta_g p\left(G^* = g', T^* = t\right)}{1 - \delta_g - \epsilon_g} \tag{9.4}$$

We can correct for toxicity measurement error in one of two ways.This assumption can be formalized as:

$$\delta_t = p\left(T^* = t' \middle| T = t\right) = p\left(T^* = t' \middle| T = t, G = g\right) \tag{9.5}$$

$$\epsilon_t = p\left(T^* = t \middle| T = t'\right) = p\left(T^* = t \middle| T = t', G = g\right) \tag{9.6}$$

Where t and t' are two different labels for toxicity, and g is any label for gender. Having computed $p\left(G = g, T^* = t\right)$ in (9.4), we can now correct for toxicity measurement

Sensitivity Analysis	Differential Error	τ Estimate(s)		
		2.5th	Mean	97.5th
None	No		-0.002	
	Yes		0.002	
Bootstrap	No	-0.002	-0.002	-0.002
	Yes	-0.022	-0.011	0.002
Clopper	No	-0.003	-0.003	-0.002
	Yes	-0.030	-0.027	-0.008

Table 9-I. Estimates and intervals for τ (9.1) using `Demographer` and `Perspective` predictions. All estimates use measurement error correction, but use different sensitivity analyses and make different assumptions regarding differential error. Negative values indicate that men are more likely to receive toxic replies.

error:

$$p\left(G=g,T=t\right)=\frac{(1-\delta_t)p\left(G=g,T^*=t\right)+\delta_t p\left(G=g,T^*=t'\right)}{1-\delta_t-\epsilon_t} \tag{9.7}$$

If we are unwilling to make a non-differential error assumption, we need to define error rates for the toxicity classifier that are conditional on the predicted gender of the parent tweet. We can do so as:

$$\delta_{t|g}=p\left(T^*=t'\middle|T=t,G=g\right) \tag{9.8}$$

$$\epsilon_{t|g}=p\left(T^*=t\middle|T=t',G=g\right) \tag{9.9}$$

No longer assuming that $\delta_t = \delta t \mid g$ doubles the number of error rates we need to estimate. While $\delta_{t=1|g=0}=\epsilon_{t=0|g=0}=p\left(T^*=0|T=1,G=0\right)$, we must estimate $\delta_{1|g}$ and $\epsilon_{1|g}$ for $g\in[0,1]$. Then, correcting for toxicity measurement error becomes:

$$p\left(G=g,T=t\right)=\frac{(1-\delta_{t|g})p\left(G=g,T^*=t\right)+\delta_{t|g}p\left(G=g,T^*=t'\right)}{1-\delta_{t|g}-\epsilon_{t|g}} \tag{9.10}$$

The implication of whether or not we make a non-differential error assumption is whether or not we can compute a conditional error rate such as $\delta_{t|g}$. If we had not scraped the parents of the tweets labeled for ground-truth toxicity in the Founta et al. (2018) dataset, we would be unable to estimate this error rate.

For both the non-differential and differential measurement error corrections, we produce new estimates of our τ defined in (9.1). Under the non-differential assumption, we estimate $\tau = -0.0018$; under differential error, we estimate $\tau = 0.0022$. Both these estimates suggest there is no difference in rate of toxic comments received by men or women – but these estimates provide no assessment of the uncertainty in our methods. We need the sensitivity analyses introduced in Chapter 6 to understand how robust these estimates of no effect are.

We will use the Bootstrap and Clopper-Pearson sensitivity analyses from §6.3 to understand the uncertainty of our estimates. For the bootstrap approach, we will replace each δ and ϵ error rate in our measurement error corrections above with 256 new error rates created by resampling our test sets with replacement. For the Clopper-Pearson interval approach, we also produce 256 new error rate estimates, using $\alpha = 0.95$, $\gamma = 1$, and $k = 16$ (see 6.3.3 for details on these hyperparameters). For both of these sensitivity analyses, whether we make the non-differential error assumption changes the complexity of our approach. With that assumption, we consider 256 values of four different error rates: $\delta_{g=0}$, $\epsilon_{g=0}$, $\delta_{t=0}$, and $\epsilon_{t=0}$. Without that assumption, we have six error rates because we need $\delta_{g=0}$ and $\epsilon_{g=0}$ as well as $\delta_{t=0|g}$ and $\epsilon_{t=0|g}$ for $g \in [0, 1]$. Additionally, we have less data with which we can estimate each error rate, so we would expect larger intervals when we allow for differential error, as we have seen in §6.6.1.

We see these results in Table 9-I. As expected, we see much larger intervals when we allow for differential error. While these differences are still small, they suggest that users labeled as men may actually be up to two or three percent *more* likely to receive toxic replies than women. Before we discuss the full implications of these results and whether to trust them, we will first address another assumption we have made in our analyses so far.

Label	Model	ρ	τ	APS	AUC	Acc
Organization	Independent	0.557	0.429	0.992	0.911	0.944
	Chained	0.557	0.429	0.992	0.911	0.944
Gender	Independent	0.680	0.485	0.892	0.880	0.793
	Chained	0.682	0.486	0.893	0.881	0.794
Toxicity	Independent	0.390	0.270	0.556	0.814	0.910
	Chained	0.391	0.270	0.554	0.814	0.910

Table 9-II. Faithfulness evaluation of Independent and Chained proxies across the 1.17M pairs of tweets in the test set. Metrics are Spearman ρ, Kendall τ, Average Precision Score (APS), ROC Area Under the Curve (AUC), and Accuracy (Acc). Methods are essentially identical.

9.6 Proxy Models for Gender and Toxicity

A possible weakness of our analysis so far is that we do not know the precise details of how `Perspective` makes predictions. For our dataset in which we are interested in the toxicity of children tweets, we also have no way to modify the model to add in the context of the parent tweet. `Demographer` and `Perspective` allow us to infer $p\left(G^*|X_1\right)$ and $p\left(T^*|X_2\right)$, where X_1 and X_2 are disjoint sets of features. We do not have any way of modeling $p\left(T^*, G^*|X\right)$ without assuming that $G^* \perp T^*$. To explore a joint distribution of our two predicted variables, we turn to our proxy model methodology from Chapter 8.

We will train proxy models to predict organizations, gender, and toxicity given the same feature sets across the train set introduced in §9.2. We will train our proxies to predict the continuous-valued prediction of both `Demographer` and `Perspective` on the entire train set, and then evaluate its faithfulness on the test set and its accuracy on the datasets with ground-truth labels for gender and toxicity. We will train our gender proxy model to predict gender from features of the parent tweet. We will consider two types of proxy models to predict toxicity. The first will predict toxicity only from features in the child tweet independently of the gender proxy,

		Faithfulness to Classifier					Ground Truth		
Label	Model	ρ	τ	APS	AUC	Acc	APS	AUC	Acc
Gender	`Demographer`	1	1	1	1	1	0.691	0.853	0.824
	Independent	0.721	0.514	0.900	0.924	0.837	0.762	0.856	0.770
	Chained	0.723	0.515	0.901	0.925	0.835	0.765	0.858	0.766
Toxicity	`Perspective`	1	1	1	1	1	0.901	0.956	0.917
	Independent	0.656	0.473	0.885	0.906	0.880	0.814	0.906	0.882
	Chained	0.653	0.470	0.881	0.923	0.880	0.811	0.904	0.882

Table 9-III. Faithfulness evaluation of Independent and Chained proxies across the gender and toxicity ground-truth validation sets given by Preoțiuc-Pietro and Ungar (2018) and Founta et al. (2018). Metrics are defined in Table 9-II. Independent and Chained proxies are identical on all but a few comparisons.

i.e. $p\left(T^*|X\right)$. The second will predict toxicity using the gender proxy's prediction as a feature; i.e. $p\left(T^*|G^*, X\right)$. We refer to the first setting as 'Independent' and the second as 'Chained.' We treat the prediction of organizations as independent in all cases, and use it to initially filter out users labeled as organizations.

We extract bag-of-word features from the tweet text, restricting our vocabulary to the 56k tokens that appear at least 100 times in the parent tweets train set. We also extract features from the Twitter user's name and profile, following the methodology we introduced in Wood-Doughty, Mahajan, and Dredze (2018). We restrict the profile feature vocabulary in the same way, giving us 140k such features. We train our proxy models as `SGDRegressor` models from the `sklearn` library (Pedregosa et al., 2011), using L_1 regularization as in §8.4.3. We train on the 3.38M pairs of tweets in our train set and then evaluate the faithfulness of our methods by how closely their predictions match the classifiers' test set predictions.

Tables 9-II and 9-III show our faithfulness evaluations for both the Independent and Chained proxies on our test set and the ground-truth validation sets. We see that our proxies are quite faithful to the original `Demographer` and `Perspective` models in both the held-out test set and the two ground-truth validation sets. On

Model	Sensitivity Analysis	D.E.	τ Estimate(s)		
			2.5th	Mean	97.5th
Independent	None	No		−0.003	
		Yes		−0.006	
	Bootstrap	No	−0.003	−0.003	−0.003
		Yes	−0.018	−0.015	−0.006
	Clopper	No	−0.005	−0.005	−0.003
		Yes	−0.068	−0.060	−0.024
Chained	None	Yes		−0.006	
	Bootstrap	Yes	−0.019	−0.015	−0.006
	Clopper	Yes	−0.073	−0.065	−0.025

Table 9-IV. Estimates and intervals for τ (9.1) using predictions from our Independent and Chained proxies. All estimates use measurement error correction, but use different sensitivity analyses and make different assumptions regarding differential error (D.E.). Negative values indicate that men are more likely to receive toxic replies.

those validation sets, the Independent and Chained have solid but slightly lower performance compared to the original trained models. The Chained proxy varies more from the Independent proxy on the validation sets, but both are still nearly identical across the different metrics.

Table 9-IV shows the results of our τ estimates when using our proxy models to predict gender and toxicity. As in Table 9-I, we vary whether we make a differential error assumption and consider two sensitivity analyses. A first takeaway from these results is that the difference between the Independent and Chained proxies remains negligible. The difference between the non-differential and differential error settings, however, is comparatively enormous. When allowing for differential error, the Clopper-Pearson method greatly increases the interval width compared to Bootstrap. This result is consistent with our findings in Figure 6-4, where Clopper-Pearson provides the widest intervals on synthetic data.

Table 9-IV reflects Table 9-I in that our estimates of τ are negative with few exceptions, though we see much larger magnitudes of differences in the analyses using our proxy models. These results indicate that in our data, users labeled as men are slightly more likely to receive toxic replies than users labeled as women. While this may not match our intuition given previous work (Akhtar and Morrison, 2019; Delisle et al., 2018), we discuss in §9.7.2 whether the types of Twitter interactions captured in our data differ from those previously studied.

9.7 Conclusions

9.7.1 Contributions

In this chapter, we have tied together several threads of work throughout the thesis. We have collected a large dataset to allow for the analysis of how two different ML classifiers comport when used to infer different labels on the same data. We use two ground-truth validation sets to understand the accuracy of these classifiers and to set up a measurement error framework in which we can correct for the bias their misclassifications may introduce. We explore the role of the non-differential error assumption and show that without it, we see much more uncertainty in our estimates across all methods. Finally, we use our proxy model approach to replace the black-box `Perspective` model with an interpretable model that we can combine into a joint classifier for toxicity and gender.

9.7.2 Future Work

While these analyses are still preliminary, they raise interesting questions. Our results indicate that users labeled by `Demographer` as men may be slightly more likely to receive toxic replies. However, correlation is not causation, and this slight difference might reflect the aggregation over a wide variety of confounding factors. Two immediately-relevant forms of confounding are the relationship between the parent and child users and the topic or toxicity of the parent tweet. Two friends may use offensive language in a way that is not meant to offend the other; if this behavior is more common among men than women, that could influence our results. Similarly, if topics such as politics or sports are more likely to receive toxic replies, this could also provide an important confounding factor which would help contextualize our results. A simple extension of our work would be to use `Perspective` to infer the toxicity of parent tweets and extend our analysis to the distribution of $p\left(T_c \middle| G_c, T_p, G_p\right)$ for a parent p and child c.

A limitation of our work thus far is that while `Perspective` is marketed to classify toxicity in the seven languages in our data, we only consider a validation set of English-language tweets. Furthermore, `Demographer` was trained on a dataset of Twitter users who over-represent an English-language context, and may be much less accurate on Twitter users from outside the United States. If our classifiers error rates vary by language or user geolocation, stratifying our data could provide important context. A more robust solution would be to find toxicity validation datasets for each language we consider and validate `Demographer` on users from a more global context.

Preprocessing decisions also influence which tweets are included in our dataset, and what our analyses can tell us about the world. We constructed our dataset with tweets from December 2020, but did not scrape them using the Twitter API until May 2021, leading to 26% of parents and 18% of children being unable to be scraped. It's further likely that certain *kinds* of tweets or users are more likely to go missing from our dataset over time. If particularly toxic tweets are deleted by Twitter's moderators, our analysis may not be representative of how most users experience toxicity. Reproducing our analyses on data from the immediate past may shed some light on data degradation over time.

Finally, there are many assumptions that we cannot test. A particular concern in this analysis, following our discussion in Chapter 4, is selection bias. The data on which our classifiers were trained and the data we use for validation are unlikely to be representative of the entire Twitter distribution. There are methods which could help us recover from possible selection bias, but may require more knowledge about the data on which these models are trained (Bareinboim and Tian, 2015). Similarly, there are known issues of selection bias within the Twitter 1% sample which influence the ability of our results to generalize to the entire population (Morstatter et al., 2013), but cannot be easily corrected (Blank, 2017). Despite these limitations, our analyses provide a broader look at the dynamics between gender and toxicity, and explore the assumptions necessary to enable these and future studies.

Chapter 10

Conclusions

10.1 Contributions

In this thesis, we have explored the challenges of incorporating ML and NLP models into causal analyses, and developed methods and frameworks for enabling such analyses. In Chapters 2 and 3, we review the traditional assumptions of both causal inference and NLP, and discuss the intersections that pose challenges and opportunities. Chapter 4 recaps our past work on predicting demographics of social media users, which informs our later analyses and provides an example of how predictive models have been used to contextualize real-world analyses.

Chapter 5 introduces our measurement error framework for incorporating predictive models into causal analyses, and evaluates a theoretically-unbiased estimator on toy data. Chapter 6 extends these results by highlighting a key limitation of this estimator which we solve with a sensitivity analysis approach that incorporates the uncertainty of an ML classifier into the output of our causal analyses. We use this method for an analysis of Twitter posts about vaccine opinions. Chapter 7 introduces a general framework for evaluating methods that use text data in causal analyses. By using recent advances in text generation, we demonstrate that our measurement error approach improves over competing methods that make more assumptions.

Chapter 8 switches to focus on interpretability, introducing a proxy model approach that allows us to explain the predictions of two black-box models. Finally, Chapter 9 ties together work from Chapters 4, 5, 6, and 8 to use two NLP classifiers to explore the joint distribution of gender and to explore the joint distribution of gender and toxic language across of millions of tweets.

10.2 Limitations and Future Work

In Chapters 5 through 9, we have discussed the limitations specific to each of our contributions. While we will not recap these individually here, we will discuss the commonalities between them and how they may inform future work.

10.2.1 Selection Bias

As discussed in §3.3.1, 4.5, and 9.7.2, selection bias is a widespread and important concern in both predictive modeling and causal inference. Every method we have discussed relies on collecting data from a larger distribution and makes implicit assumptions that analyses on that data are generally applicable to the full distribution. Selection bias can violate these assumptions, and is almost certainly present to some extent in all social media datasets. There are causal inference methods that focus on recovering from selection bias (Bareinboim and Tian, 2015), but these have not been widely applied to improving ML or NLP classifiers and are not trivially incorporated into new causal analyses.

If public health officials turn to social media to survey population behaviors and opinions regarding health, they will rely on automated methods to analyze billions of tweets in real time. The state-of-the-art models for these tasks use pre-trained language models trained on massive datasets, which often reflect the sociocultural biases that dominate the websites from which these datasets were scraped (Bender et al., 2021). This leads to several questions that will rely on causal reasoning. Can we understand the selection bias introduced by scraping data from websites with specific cultural norms? Can we model how word embeddings may be confounded by historical biases? Can we correct the selection bias introduced by training demographic prediction models on data that is not representative of the entire Twitter population? If we are to use these social media analyses to make policy recommendations to domain experts, we need to understand how such biases in the data and methodologies may influence our conclusions.

10.2.2 Missing and Coarsened Data

The Twitter datasets we explore in Chapters 4, 6, and 9 all suffer from missing data, as tweets are deleted and users deactivate their accounts. Our demographic prediction models from Chapter 4 are also limited in that their predictions do not match the complexity of the underlying concept they seek to measure (Broniatowski and Tucker, 2017); there are more than four races or ethnicities and more than two genders (Keyes, May, and Carrell, 2021). Both of these issues may be framed as missing or coarsened data problems. Modeling the missingness process of how tweets are deleted or the way in which demographics subgroups are often not given explicit recognition in surveys.

10.2.3 Real-World Applications

Finally, the methods we have introduced in this thesis are motivated by the specific examples we introduced in §1.1, but we have not undertaken the additional domain-specific work necessary to actually apply our methodologies in a way that could actually affect expert decision-making. Such work requires working with domain experts to understand what assumptions are plausible, what predictive models are available, and how much uncertainty is acceptable. In the healthcare domain in particular, experts typically cannot share patient data across organizations. We need causal methods that enable longitudinal studies on internal datasets of clinical notes that can scale to diverse clinical settings, be as easily adopted as a new R or Python package, and keep causal assumptions transparent to the domain experts.

Suppose an outbreak of a novel disease has forced clinicians to make treatment decisions based on limited knowledge of the disease's biological mechanisms. If this disease has previously existed undiagnosed in the population, we may be able to train an ML model to identify patients who likely had the disease before it was identified. This could enable a retrospective observational study that could save lives but would be impossible with our current technology.

10.2.4 Future Directions

Over the past several years, there has been a steady increase in research into the intersection of causal inference and NLP (Keith, Jensen, and O'Connor, 2020; Veitch, Sridhar, and Blei, 2020; Yao et al., 2019; Feder et al., 2021; Pryzant et al., 2021). This line of work has led to an inaugural workshop co-located with EMNLP 2021, which I am helping organize. While this workshop will provide an exciting opportunity to discuss recent advances in this nascent research area, there are several longer-term future directions that deserve future study.

The challenges of selection bias and missing data are broadly applicable to any analysis of text data which must always be collected or annotated in a manner subject to selection bias or missingness. More generally, however, we need new frameworks for understanding how natural language fits into theories of causality. The basic assumptions on which causal inference depend, such as consistency (discussed in Chapter 2), are less obviously plausible when we start imagining hypothetical interventions on the *way in which we generate language*. Our measurement error framework is applicable to any setting in which we can identify a specific unmeasured variable for which we can predict a proxy, but there are many cases in which the construct in which we are interested cannot be easily predicted.

Thus far, while there has been wide interest in these questions, there has not been enough time for published applications to demonstrate whether causal inferences drawn from text data hold up under repeated scrutiny. The methods from causal inference research continue to gain wider acceptance in clinical and social science domains, in part because causal relations evidenced by observational studies can be supported by randomized control trials or further study. Wide-scale randomized studies that rely on natural language are likely infeasible, and repeated studies may have greater variance due to the inherently fluid nature of human language. Real-world observational studies that build on the research from this thesis and contemporary work will allow us to begin to question or validate the assumptions on which these methods rely. By working with experts in various domains and through trial and error, we can continue to develop more robust methods for causal inference with text.

References

Abat, Cédric and Didier Raoult (2018). “Human papillomavirus vaccine: Urgent need to promote gender parity.” In: *European journal of epidemiology* 33.3, pp. 259–261.

Afzal, Naveed, Vishnu Priya Mallipeddi, Sunghwan Sohn, Hongfang Liu, Rajeev Chaudhry, Christopher G Scott, Iftikhar J Kullo, and Adelaide M Arruda-Olson (2018). “Natural language processing of clinical notes for identification of critical limb ischemia.” In: *International journal of medical informatics* 111, pp. 83–89.

Akhtar, Shazia and Catriona M Morrison (2019). “The prevalence and impact of online trolling of UK members of parliament.” In: *Computers in Human Behavior* 99, pp. 322–327.

Al Baghal, Tarek, Luke Sloan, Curtis Jessop, Matthew L Williams, and Pete Burnap (2020). “Linking Twitter and survey data: The impact of survey mode and demographics on consent rates across three UK studies.” In: *Social Science Computer Review* 38.5, pp. 517–532.

Al Zamal, Faiyaz, Wendy Liu, and Derek Ruths (2012). “Homophily and Latent Attribute Inference: Inferring Latent Attributes of Twitter Users from Neighbors.” In: *ICWSM* 270.

Alber, Maximilian, Sebastian Lapuschkin, Philipp Seegerer, Miriam Hägele, Kristof T. Schütt, Grégoire Montavon, Wojciech Samek, Klaus-Robert Müller, Sven Dähne, and Pieter-Jan Kindermans (2019). “iNNvestigate Neural Networks!” In: *Journal of Machine Learning Research* 20.93, pp. 1–8.

Amir, Silvio, Mark Dredze, and John W. Ayers (2019). “Population Level Mental Health Surveillance over Social Media with Digital Cohorts.” In: *CLPsych*.

Amiri, Parisa, Golaleh Asghari, Hoda Sadrosadat, Mehrdad Karimi, Atieh Amouzegar, Parvin Mirmiran, and Fereidoun Azizi (2017). “Psychometric properties of a developed questionnaire to assess knowledge, attitude and practice regarding vitamin D (D-KAP-38).” In: *Nutrients* 9.5, p. 471.

Anderson, Michael L and Jeremy Magruder (2017). *Split-sample strategies for avoiding false discoveries*. Tech. rep. National Bureau of Economic Research.

Andrus, McKane, Elena Spitzer, Jeffrey Brown, and Alice Xiang (2021). “What We Can’t Measure, We Can’t Understand: Challenges to Demographic Data Procurement in the Pursuit of Fairness.” In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. FAccT ’21. Virtual Event, Canada: Association for Computing Machinery, pp. 249–260. DOI: 10.1145/3442188.3445888.

Angwin, Julia and Terry Parris Jr (2016). “Facebook lets advertisers exclude users by race.” In: *ProPublica blog* 28.

Arco, Flor Miriam Plaza del, M Dolores Molina-González, M Teresa Martín-Valdivia, and L Alfonso Urena Lopez (2019). “SINAI at SemEval-2019 Task 6: Incorporating lexicon knowledge into SVM learning to identify and categorize offensive language in social media." In: *Proceedings of the 13th International Workshop on Semantic Evaluation*, pp. 735–738.

Ardehaly, Ehsan Mohammady and Aron Culotta (2017). "Co-training for demographic classification using deep learning from label proportions." In: *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*. IEEE, pp. 1017–1024.

Arrieta, Alejandro Barredo, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. (2020). "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI." In: *Information Fusion* 58, pp. 82–115.

Atanasova, Pepa, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein (2020). "A Diagnostic Study of Explainability Techniques for Text Classification." In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 3256–3274.

Ayers, John W, Benjamin M Althouse, and Mark Dredze (2014). "Could behavioral medicine lead the web data revolution?" In: *Jama* 311.14, pp. 1399–1400.

Bach, Sebastian, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek (2015). "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation." In: *PloS one* 10.7, e0130140.

Balke, Alexander and Judea Pearl (1997). "Bounds on treatment effects from studies with imperfect compliance." In: *Journal of the American Statistical Association* 92.439, pp. 1171–1176.

Bamman, David, Jacob Eisenstein, and Tyler Schnoebelen (2014). "Gender identity and lexical variation in social media." In: *Journal of Sociolinguistics* 18.2, pp. 135–160.

Bang, Heejung and James M Robins (2005). "Doubly robust estimation in missing data and causal inference models." In: *Biometrics* 61.4, pp. 962–973.

Bareinboim, Elias and Judea Pearl (2012). "Controlling selection bias in causal inference." In: *Artificial Intelligence and Statistics*, pp. 100–108.

Bareinboim, Elias and Jin Tian (2015). "Recovering causal effects from selection bias." In: *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Barocas, Solon, Andrew D Selbst, and Manish Raghavan (2020). "The hidden assumptions behind counterfactual explanations and principal reasons." In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 80–89.

Beger, Andreas, Cassy L Dorff, and Michael D Ward (2016). "Irregular leadership changes in 2014: Forecasts using ensemble, split-population duration models." In: *International Journal of Forecasting* 32.1, pp. 98–111.

Belinkov, Yonatan and Yonatan Bisk (2018). "Synthetic and Natural Noise Both Break Neural Machine Translation." In: *International Conference on Learning Representations*.

Beller, Charley, Rebecca Knowles, Craig Harman, Shane Bergsma, Margaret Mitchell, and Benjamin Van Durme (2014). "I'm a Belieber: Social Roles via Self-identification and Conceptual Attributes." In: *ACL*, pp. 181–186.

Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen (2014). "Inference on treatment effects after selection among high-dimensional controls." In: *The Review of Economic Studies* 81.2, pp. 608–650.

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell (2021). "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" In: *Proceedings of FAccT*.

Benner, Katie, Glenn Thrush, and Mike Isaac (2019). “Facebook Engages in Housing Discrimination With Its Ad Practices, US Says.” In: *The New York Times* 28, p. 2019.

Benton, Adrian, Glen Coppersmith, and Mark Dredze (2017). “Ethical Research Protocols for Social Media Health Research.” In: *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*. Valencia, Spain: Association for Computational Linguistics, pp. 94–102. DOI: 10.18653/v1/W17-1612.

Bergsma, Shane, Mark Dredze, Benjamin Van Durme, Theresa Wilson, and David Yarowsky (2013). “Broadly Improving User Classification via Communication-Based Name and Location Clustering on Twitter.” In: *HLT-NAACL*.

Berk, Richard, Lawrence Brown, Andreas Buja, Kai Zhang, Linda Zhao, et al. (2013). “Valid post-selection inference.” In: *The Annals of Statistics* 41.2, pp. 802–837.

Binns, Reuben, Michael Veale, Max Van Kleek, and Nigel Shadbolt (2017). “Like trainer, like bot? Inheritance of bias in algorithmic content moderation.” In: *International conference on social informatics*. Springer, pp. 405–415.

Blank, Grant (2017). “The digital divide among Twitter users and its implications for social research.” In: *Social Science Computer Review* 35.6, pp. 679–697.

Blei, David M and Jon D McAuliffe (2007). “Supervised topic models.” In: *Proceedings of the 20th International Conference on Neural Information Processing Systems*, pp. 121–128.

Blei, David M, Andrew Y Ng, and Michael I Jordan (2003). “Latent dirichlet allocation.” In: *Journal of machine Learning research* 3.Jan, pp. 993–1022.

Blitzer, John, Mark Dredze, and Fernando Pereira (2007). “Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification.” In: *Proceedings of the 45th annual meeting of the association of computational linguistics*, pp. 440–447.

Blyth, Colin R (1972). “On Simpson’s paradox and the sure-thing principle.” In: *Journal of the American Statistical Association* 67.338, pp. 364–366.

Boag, Willie, Tristan Naumann, and Peter Szolovits (2016). “Towards the Creation of a Large Corpus of Synthetically-Identified Clinical Notes.” In: *Machine Learning for Health Workshop at NeurIPS*.

Bodnar, Lisa M, Hyagriv N Simhan, Janet M Catov, James M Roberts, Robert W Platt, Jill C Diesel, and Mark A Klebanoff (2014). “Maternal vitamin D status and the risk of mild and severe preeclampsia.” In: *Epidemiology (Cambridge, Mass.)* 25.2, p. 207.

Bollen, Kenneth A, Kathleen M Gates, and Zachary Fisher (2018). “Robustness conditions for MIIV-2SLS when the latent variable or measurement model is structurally misspecified.” In: *Structural equation modeling: a multidisciplinary journal* 25.6, pp. 848–859.

Bolukbasi, Tolga, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai (2016). “Man is to computer programmer as woman is to homemaker? debiasing word embeddings.” In: *Advances in neural information processing systems* 29, pp. 4349–4357.

Bordia, Shikha and Samuel Bowman (2019). “Identifying and Reducing Gender Bias in Word-Level Language Models.” In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pp. 7–15.

Borkan, Daniel, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman (2019). “Nuanced metrics for measuring unintended bias with real data for text classification.” In: *Companion proceedings of the 2019 world wide web conference*, pp. 491–500.

Broniatowski, David A (2021). “Psychological Foundations of Explainability and Interpretability in Artificial Intelligence.” In.

Broniatowski, David A, Amelia M Jamison, SiHua Qi, Lulwah AlKulaib, Tao Chen, Adrian Benton, Sandra C Quinn, and Mark Dredze (2018). "Weaponized health communication: Twitter bots and Russian trolls amplify the vaccine debate." In: *AJPH* 108.10, pp. 1378–1384.

Broniatowski, David A and Conrad Tucker (2017). "Assessing causal claims about complex engineered systems with quantitative data: internal, external, and construct validity." In: *Systems Engineering* 20.6, pp. 483–496.

Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Gretchen Herbert Ariel and-Voss Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei (2020). "Language models are few-shot learners." In: *Advances in Neural Information Processing Systems*.

Buczak, Anna L., Benjamin Baugher, Christine Martin, Meg Keiley-Listermann, James Howard II, Nathan Parrish, Anton Stalick, Daniel Berman, and Mark Dredze (2021). "Crystal Cube: Forecasting Disruptive Events." In: *Manuscript submitted for publication*.

Burger, John D, John Henderson, George Kim, and Guido Zarrella (2011). "Discriminating gender on Twitter." In: *Proceedings of the conference on empirical methods in natural language processing*. Association for Computational Linguistics, pp. 1301–1309.

Burnap, Pete and Matthew L Williams (2016). "Us and them: identifying cyber hate on Twitter across multiple protected characteristics." In: *EPJ Data Science* 5.1, p. 11.

Butler, Declan (2013). "When Google got flu wrong." In: *Nature news* 494.7436, p. 155.

Butler, Joseph S, Richard V Burkhauser, Jean M Mitchell, and Theodore P Pincus (1987). "Measurement error in self-reported health variables." In: *Review of Economics and Statistics* 69.4, pp. 644–650.

Butler, Judith (1988). "Performative acts and gender constitution: An essay in phenomenology and feminist theory." In: *Theatre journal* 40.4, pp. 519–531.

Campbell, Donald T and Julian C Stanley (1963). *Experimental and quasi-experimental designs for research*. Houghton Mifflin Company.

Cao, Yang Trista and Hal Daumé III (2020). "Toward Gender-Inclusive Coreference Resolution." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4568–4595.

Carroll, Raymond J, David Ruppert, Ciprian M Crainiceanu, and Leonard A Stefanski (2006). *Measurement error in nonlinear models: a modern perspective*. Chapman and Hall/CRC.

Caruana, Rich, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad (2015). "Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission." In: *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1721–1730.

Cessie, Saskia le, Jan Debeij, Frits R Rosendaal, Suzanne C Cannegieter, and Jan P Vandenbrouckea (2012). "Quantification of bias in direct effects estimates due to different types of measurement error in the mediator." In: *Epidemiology*, pp. 551–560.

Chancellor, Stevie and Munmun De Choudhury (2020). "Methods in predictive techniques for mental health status on social media: a critical review." In: *NPJ digital medicine* 3.1, pp. 1–11.

Chandrasekaran, Ranganathan, Vikalp Mehta, Tejali Valkunde, and Evangelos Moustakas (2020). "Topics, trends, and sentiments of tweets about the COVID-19 pandemic: Temporal infoveillance study." In: *Journal of medical Internet research* 22.10, e22624.

Chang, Jonathan, Itamar Rosenn, Lars Backstrom, and Cameron Marlow (2010). "epluribus: Ethnicity on social networks." In: *ICWSM*.

Chen, Jonathan H and Steven M Asch (2017). "Machine learning and prediction in medicine – beyond the peak of inflated expectations." In: *NEJM* 376.26, p. 2507.

Chen, Tao and Mark Dredze (2018). "Vaccine images on Twitter: Analysis of what images are shared." In: *JMIR* 20.4.

Chen, Xin, Yu Wang, Eugene Agichtein, and Fusheng Wang (2015). "A Comparative Study of Demographic Attribute Inference in Twitter." In: *ICWSM* 15, pp. 590–593.

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, and Whitney K Newey (2016). *Double machine learning for treatment and causal parameters*. Tech. rep. cemmap working paper, Centre for Microdata Methods and Practice.

Choi, Edward, Siddharth Biswal, Bradley Malin, Jon Duke, Walter F Stewart, and Jimeng Sun (2017). "Generating multi-label discrete patient records using generative adversarial networks." In: *Machine learning for healthcare conference*. PMLR, pp. 286–305.

Clopper, Charles J and Egon S Pearson (1934). "The use of confidence or fiducial limits illustrated in the case of the binomial." In: *Biometrika*, pp. 404–413.

Coldman, Andrew J, Terry Braun, and Richard P Gallagher (1988). "The classification of ethnic status using name information." In: *Journal of Epidemiology and Community Health* 42.4, pp. 390–395.

Comstock, R Dawn, Edward M Castillo, and Suzanne P Lindsay (2004). "Four-year review of the use of race and ethnicity in epidemiologic and public health research." In: *American journal of epidemiology* 159.6, pp. 611–619.

Conover, Michael, Jacob Ratkiewicz, Matthew R Francisco, Bruno Gonçalves, Filippo Menczer, and Alessandro Flammini (2011). "Political Polarization on Twitter." In: *ICWSM* 133, pp. 89–96.

Cook, Samantha, Corrie Conrad, Ashley L Fowlkes, and Matthew H Mohebbi (2011). "Assessing Google flu trends performance in the United States during the 2009 influenza virus A (H1N1) pandemic." In: *PloS one* 6.8, e23610.

Cook, Thomas D, Donald Thomas Campbell, and William Shadish (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin Boston, MA.

Crawford, Kate (2017). "The trouble with bias." In: *Conference on Neural Information Processing Systems, invited speaker*.

Culley, Lorraine (2006). "Transcending transculturalism? Race, ethnicity and health-care." In: *Nursing Inquiry* 13.2, pp. 144–153.

Culotta, Aron, Nirmal Kumar, and Jennifer Cutler (2015). "Predicting the demographics of twitter users from website traffic data." In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 29.

Culotta, Aron, Nirmal Ravi Kumar, and Jennifer Cutler (2016). "Predicting Twitter User Demographics using Distant Supervision from Website Traffic Data." In: *J. Artif. Intell. Res.(JAIR)* 55, pp. 389–408.

Dathathri, Sumanth, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu (2019). "Plug and Play Language Models: A Simple Approach to Controlled Text Generation." In: *International Conference on Learning Representations*.

Davis, Clayton Allen, Onur Varol, Emilio Ferrara, Alessandro Flammini, and Filippo Menczer (2016). "Botornot: A system to evaluate social bots." In: *Proceedings of the 25th international conference companion on world wide web*, pp. 273–274.

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

Delisle, Laure, Alfredo Kalaitzis, Krzysztof Majewski, Archy de Berker, Milena Marin, and Julien Cornebise (2018). "A large-scale crowdsourced analysis of abuse against women journalists and politicians on Twitter." In: *Proceedings of the 32nd International Conference on Neural Information Processing Systems*.

Deng, Jia, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei (2009). "Imagenet: A large-scale hierarchical image database." In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee, pp. 248–255.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2018). "Bert: Pre-training of deep bidirectional transformers for language understanding." In: *arXiv*.

DeYoung, Jay, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C Wallace (2020). "ERASER: A Benchmark to Evaluate Rationalized NLP Models." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4443–4458.

Dorie, Vincent, Jennifer Hill, Uri Shalit, Marc Scott, Dan Cervone, et al. (2019). "Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition." In: *Statistical Science* 34.1, pp. 43–68.

Doshi-Velez, Finale and Been Kim (2017). "Towards a rigorous science of interpretable machine learning." In: *arXiv preprint arXiv:1702.08608*.

Dredze, Mark, David A Broniatowski, Michael C Smith, and Karen M Hilyard (2016). "Understanding vaccine refusal: why we need social media now." In: *American journal of preventive medicine* 50.4, pp. 550–552.

Dredze, Mark, Miles Osborne, and Prabhanjan Kambadur (2016). "Geolocation for Twitter: Timing Matters." In: *NAACL*.

Dredze, Mark, Michael J Paul, Shane Bergsma, and Hieu Tran (2013). "Carmen: A twitter geolocation system with applications to public health." In: *Workshops at the Twenty-Seventh AAAI Conference on Artificial Intelligence*.

Dressler, William W, Kathryn S Oths, and Clarence C Gravlee (2005). "Race and ethnicity in public health research: models to explain health disparities." In: *Annu. Rev. Anthropol.* 34, pp. 231–252.

Drton, M, B Sturmfels, and S Sullivant (2009). *Lectures on Algebraic Statistics. Oberwohlfach Mathematical Seminars, vol. 39*.

Drton, Mathias and Marloes H Maathuis (2017). "Structure learning in graphical modeling." In: *Annual Review of Statistics and Its Application* 4, pp. 365–393.

Du, Mengnan, Ninghao Liu, and Xia Hu (2019). "Techniques for interpretable machine learning." In: *Communications of the ACM* 63.1, pp. 68–77.

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

Eisenstein, Jacob, Brendan O'Connor, Noah A Smith, and Eric P Xing (2014). "Diffusion of lexical change in social media." In: *PloS one* 9.11, e113114.

Eliashberg, Jehoshua and John R Hauser (1985). "A measurement error approach for modeling consumer risk preference." In: *Management Science* 31.1, pp. 1–25.

Elliott, Marc N, Allen Fremont, Peter A Morrison, Philip Pantoja, and Nicole Lurie (2008). "A new method for estimating race/ethnicity and associated disparities where administrative records lack self-reported race/ethnicity." In: *Health services research* 43.5p1, pp. 1722–1736.

Elliott, Marc N, Peter A Morrison, Allen Fremont, Daniel F McCaffrey, Philip Pantoja, and Nicole Lurie (2009). "Using the Census Bureau's surname list to improve estimates of race/ethnicity and associated disparities." In: *Health Services and Outcomes Research Methodology* 9.2, pp. 69–83.

Elman, Jeffrey L (1990). "Finding structure in time." In: *Cognitive science* 14.2, pp. 179–211.

Etim, Bassey (June 2017). *The Times Sharply Increases Articles Open for Comments, Using Google's Technology*.

Evans, Robin J (2016). "Graphs for margins of Bayesian networks." In: *Scandinavian Journal of Statistics* 43.3, pp. 625–648.

Eyres, Sophie, Amy Carey, Gill Gilworth, Vera Neumann, and Alan Tennant (2005). "Construct validity and reliability of the Rivermead post-concussion symptoms questionnaire." In: *Clinical rehabilitation* 19.8, pp. 878–887.

Farrell, Max H, Tengyuan Liang, and Sanjog Misra (2021). "Deep neural networks for estimation and inference." In: *Econometrica* 89.1, pp. 181–213.

Feder, Amir, Nadav Oved, Uri Shalit, and Roi Reichart (2021). "CausaLM: Causal model explanation through counterfactual language models." In: *Computational Linguistics* 47.2, pp. 333–386.

Fedus, William, Barret Zoph, and Noam Shazeer (2021). "Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity." In: *arXiv preprint arXiv:2101.03961*.

Feng, Shi, Eric Wallace, Alvin Grissom II, Mohit Iyyer, Pedro Rodriguez, and Jordan Boyd-Graber (2018). "Pathologies of Neural Models Make Interpretations Difficult." In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 3719–3728.

Fiesler, Casey and Nicholas Proferes (2018). "'Participant' Perceptions of Twitter Research Ethics." In: *Social Media+ Society* 4.1, p. 2056305118763366.

Finkelstein, Noam, Roy Adams, Suchi Saria, and Ilya Shpitser (2020). "Partial Identifiability in Discrete Data With Measurement Error." In: *arXiv preprint arXiv:2012.12449*.

Fiscella, Kevin and Allen M Fremont (2006). "Use of geocoding and surname analysis to estimate race and ethnicity." In: *Health services research* 41.4p1, pp. 1482–1500.

Fleiss, JL and PE Shrout (1977). "The effects of measurement errors on some multivariate procedures." In: *American journal of public health* 67.12, pp. 1188–1191.

Flekova, Lucie, Jordan Carpenter, Salvatore Giorgi, Lyle Ungar, and Daniel Preoţiuc-Pietro (2016). "Analyzing Biases in Human Perception of User Age and Gender from Text." In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics, pp. 843–854. DOI: 10.18653/v1/P16-1080.

Founta, Antigoni Maria, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis (2018). "Large scale crowdsourcing and characterization of twitter abusive behavior." In: *Twelfth International AAAI Conference on Web and Social Media*.

Frohard-Dourlent, Hélène, Sarah Dobson, Beth A Clark, Marion Doull, and Elizabeth M Saewyc (2017). ""I would have preferred more options": accounting for non-binary youth in health research." In: *Nursing inquiry* 24.1, e12150.

Garg, Nikhil, Londa Schiebinger, Dan Jurafsky, and James Zou (2018). "Word embeddings quantify 100 years of gender and ethnic stereotypes." In: *Proceedings of the National Academy of Sciences* 115.16, E3635–E3644.

Gentzel, Amanda, Dan Garant, and David Jensen (2019). "The case for evaluating causal models using interventional measures and empirical data." In: *Advances in Neural Information Processing Systems*, pp. 11722–11732.

Gilpin, Leilani H, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal (2018). "Explaining explanations: An overview of interpretability of machine learning." In: *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*. IEEE, pp. 80–89.

Ginsberg, Jeremy, Matthew H Mohebbi, Rajan S Patel, Lynnette Brammer, Mark S Smolinski, and Larry Brilliant (2009). "Detecting influenza epidemics using search engine query data." In: *Nature* 457.7232, pp. 1012–1014.

Girju, Roxana (2003). "Automatic detection of causal relations for question answering." In: *Proceedings of the ACL 2003 workshop on Multilingual summarization and question answering-Volume 12*. Association for Computational Linguistics, pp. 76–83.

Glorot, Xavier, Antoine Bordes, and Yoshua Bengio (2011). "Domain adaptation for large-scale sentiment classification: A deep learning approach." In: *Proceedings of the 28th international conference on machine learning (ICML-11)*, pp. 513–520.

Greenland, Sander (2003). "Quantifying biases in causal models: classical confounding vs collider-stratification bias." In: *Epidemiology*, pp. 300–306.

Grewal, Rajdeep, Joseph A Cote, and Hans Baumgartner (2004). "Multicollinearity and measurement error in structural equation models: Implications for theory testing." In: *Marketing science* 23.4, pp. 519–529.

Grimmer, Justin (2015). "We are all social scientists now: How big data, machine learning, and causal inference work together." In: *PS: Political Science & Politics* 48.1, pp. 80–83.

Grimsley, Christopher, Elijah Mayfield, and Julia RS Bursten (2020). "Why Attention is Not Explanation: Surgical Intervention and Causal Reasoning about Neural Models." In: *Proceedings of The 12th Language Resources and Evaluation Conference*, pp. 1780–1790.

Guidotti, Riccardo, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi (2018). "A survey of methods for explaining black box models." In: *ACM computing surveys (CSUR)* 51.5, pp. 1–42.

Hahn, P Richard, Vincent Dorie, and Jared S Murray (2019). "Atlantic causal inference conference (acic) data analysis challenge 2017." In: *arXiv preprint arXiv:1905.09515*.

Hall, Peter (1988). "Theoretical comparison of bootstrap confidence intervals." In: *The Annals of Statistics*, pp. 927–953.

Hamidi, Foad, Morgan Klaus Scheuerman, and Stacy M Branham (2018). "Gender recognition or gender reductionism?: The social implications of embedded gender recognition systems." In: *CHI*. ACM, p. 8.

Hannan, Edward L (2008). "Randomized clinical trials and observational studies: guidelines for assessing respective strengths and limitations." In: *JACC: Cardiovascular Interventions* 1.3, pp. 211–217.

Hansson, Karin, Michael Muller, Tanja Aitamurto, Lilly Irani, Athanasios Mazarakis, Neha Gupta, and Thomas Ludwig (2016). "Crowd dynamics: Exploring conflicts and contradictions in crowdsourcing." In: *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, pp. 3604–3611.

Hashimoto, Tatsunori, Hugh Zhang, and Percy Liang (2019). "Unifying Human and Statistical Evaluation for Natural Language Generation." In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 1689–1701.

Hazlehurst, Brian, Allison Naleway, and John Mullooly (2009). "Detecting possible vaccine adverse events in clinical notes of the electronic medical record." In: *Vaccine* 27.14, pp. 2077–2083.

Herman, Bernease (2017). "he promise and peril of human evaluation for model interpretability." In: *Proceedings of the 31st Conference on Neural Information Processing Systems (NeurIPS 2017)*.

Hernán, Miguel A and Stephen R Cole (2009). "Invited commentary: causal diagrams and measurement bias." In: *American journal of epidemiology* 170.8, pp. 959–962.

Hill, Jennifer, Christopher Weiss, and Fuhua Zhai (2011). "Challenges with propensity score strategies in a high-dimensional setting and a potential alternative." In: *Multivariate Behavioral Research* 46.3, pp. 477–513.

Ho, Arnold K, Steven O Roberts, and Susan A Gelman (2015). "Essentialism and racial bias jointly contribute to the categorization of multiracial individuals." In: *Psychological Science* 26.10, pp. 1639–1645.

Hochreiter, Sepp and Jürgen Schmidhuber (1997). "Long short-term memory." In: *Neural computation* 9.8, pp. 1735–1780.

Hosseini, Hossein, Sreeram Kannan, Baosen Zhang, and Radha Poovendran (2017). "Deceiving google's perspective api built for detecting toxic comments." In: *arXiv preprint arXiv:1702.08138*.

Huang, Xiaolei and Michael J Paul (2019). "Neural User Factor Adaptation for Text Classification: Learning to Generalize Across Author Demographics." In: *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 136–146. DOI: 10.18653/v1/S19-1015.

Irani, Lilly C and M Six Silberman (2013). "Turkopticon: Interrupting worker invisibility in amazon mechanical turk." In: *Proceedings of the SIGCHI conference on human factors in computing systems*, pp. 611–620.

Jacovi, Alon 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*, pp. 4198–4205.

Jain, Sarthak and Byron C Wallace (2019). "Attention is not Explanation." In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 3543–3556.

Jain, Sarthak, Sarah Wiegreffe, Yuval Pinter, and Byron C. Wallace (July 2020). "Learning to Faithfully Rationalize by Construction." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 4459–4473. DOI: 10.18653/v1/2020.acl-main.409.

Jensen, David et al. (2019). "Comment: Strengthening Empirical Evaluation of Causal Inference Methods." In: *Statistical Science* 34.1, pp. 77–81.

Jiang, Heinrich, Been Kim, Melody Guan, and Maya Gupta (2018). "To trust or not to trust a classifier." In: *NeurIPS*, pp. 5541–5552.

Jiang, Jiachen and Soroush Vosoughi (2020). "Not Judging a User by Their Cover: Understanding Harm in Multi-Modal Processing within Social Media Research." In: *Proceedings of the 2nd International Workshop on Fairness, Accountability, Transparency and Ethics in Multimedia*, pp. 6–12.

Jiang, Jing and ChengXiang Zhai (2007). "Instance weighting for domain adaptation in NLP." In: *Proceedings of the 45th annual meeting of the association of computational linguistics*, pp. 264–271.

Jin, Qiao, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu (2019). "PubMedQA: A Dataset for Biomedical Research Question Answering." In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2567–2577.

Johansson, Fredrik, Uri Shalit, and David Sontag (2016). "Learning representations for counterfactual inference." In: *International conference on machine learning*, pp. 3020–3029.

Johnson, Alistair EW, Tom J Pollard, Lu Shen, H Lehman Li-Wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark (2016). "MIMIC-III, a freely accessible critical care database." In: *Scientific data* 3.1, pp. 1–9.

Jung, Soon-Gyo, Jisun An, Haewoon Kwak, Joni Salminen, and Bernard Jansen (2018). "Assessing the accuracy of four popular face recognition tools for inferring gender, age, and race." In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 12.

Jurgens, David (2013). "That's what friends are for: Inferring location in online social media platforms based on social relationships." In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 7. 1.

Jurgens, David, Tyler Finethy, James McCorriston, Yi Tian Xu, and Derek Ruths (2015). "Geolocation Prediction in Twitter Using Social Networks: A Critical Analysis and Review of Current Practice." In: *ICWSM*, pp. 188–197.

Kachel, Sven, Melanie C Steffens, and Claudia Niedlich (2016). "Traditional masculinity and femininity: Validation of a new scale assessing gender roles." In: *Frontiers in psychology* 7, p. 956.

Kalisch, Markus and Peter Bühlman (2007). "Estimating high-dimensional directed acyclic graphs with the PC-algorithm." In: *Journal of Machine Learning Research* 8.3.

Kandula, Sasikiran and Jeffrey Shaman (2019). "Reappraising the utility of Google flu trends." In: *PLoS computational biology* 15.8, e1007258.

Kaplan, Randy M and Genevieve Berry-Rogghe (1991). "Knowledge-based acquisition of causal relationships in text." In: *Knowledge Acquisition* 3.3, pp. 317–337.

Karimi, Fariba, Claudia Wagner, Florian Lemmerich, Mohsen Jadidi, and Markus Strohmaier (2016). "Inferring gender from names on the web: A comparative evaluation of gender detection methods." In: *Proceedings of the 25th International conference companion on World Wide Web*, pp. 53–54.

Kedzie, Chris and Kathleen McKeown (Nov. 2020). "Controllable Meaning Representation to Text Generation: Linearization and Data Augmentation Strategies." In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pp. 5160–5185. DOI: 10.18653/v1/2020.emnlp-main.419.

Keith, Katherine, David Jensen, and Brendan O'Connor (July 2020). "Text and Causal Inference: A Review of Using Text to Remove Confounding from Causal Estimates." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 5332–5344. DOI: 10.18653/v1/2020.acl-main.474.

Keskar, Nitish Shirish, Bryan McCann, Lav R Varshney, Caiming Xiong, and Richard Socher (2019). “Ctrl: A conditional transformer language model for controllable generation.” In: *arXiv preprint arXiv:1909.05858*.

Keyes, Os (2018). “The misgendering machines: Trans/HCI implications of automatic gender recognition.” In: *Proceedings of the ACM on Human-Computer Interaction* 2.CSCW, pp. 1–22.

Keyes, Os, Chandler May, and Annabelle Carrell (2021). “You Keep Using That Word: Ways of Thinking about Gender in Computing Research.” In: *Proceedings of the ACM on Human-Computer Interaction* 5.CSCW1, pp. 1–23.

Khayrallah, Huda and Philipp Koehn (2018). “On the Impact of Various Types of Noise on Neural Machine Translation.” In: *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pp. 74–83.

Khayrallah, Huda, Brian Thompson, Kevin Duh, and Philipp Koehn (2018). “Regularized training objective for continued training for domain adaptation in neural machine translation.” In: *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pp. 36–44.

Kim, Jungi and Patricia O’Neill-Brown (Aug. 2019). “Improving American Sign Language Recognition with Synthetic Data.” In: *Proceedings of Machine Translation Summit XVII Volume 1: Research Track*. Dublin, Ireland: European Association for Machine Translation, pp. 151–161.

Kim, Sang-Bum, Kyoung-Soo Han, Hae-Chang Rim, and Sung Hyon Myaeng (2006). “Some effective techniques for naive bayes text classification.” In: *IEEE transactions on knowledge and data engineering* 18.11, pp. 1457–1466.

Knol, Mirjam J and Tyler J VanderWeele (2012). “Recommendations for presenting analyses of effect modification and interaction.” In: *International journal of epidemiology* 41.2, pp. 514–520.

Knowles, Rebecca, Josh Carroll, and Mark Dredze (2016). “Demographer: Extremely simple name demographics.” In: *NLP+CSS*, pp. 108–113.

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E Hinton (2012). “Imagenet classification with deep convolutional neural networks.” In: *Advances in neural information processing systems*, pp. 1097–1105.

Kuroki, Manabu and Judea Pearl (2014). “Measurement bias and effect restoration in causal inference.” In: *Biometrika* 101.2, pp. 423–437.

Kwak, Haewoon, Changhyun Lee, Hosung Park, and Sue Moon (2010). “What is Twitter, a social network or a news media?” In: *Proceedings of the 19th international conference on World wide web*, pp. 591–600.

Lai, Siwei, Liheng Xu, Kang Liu, and Jun Zhao (2015). “Recurrent Convolutional Neural Networks for Text Classification.” In: *AAAI*. Vol. 333, pp. 2267–2273.

Landeiro, Virgile and Aron Culotta (2016). “Robust text classification in the presence of confounding bias.” In: *Thirtieth AAAI Conference on Artificial Intelligence*.

— (2017). “Controlling for Unobserved Confounds in Classification Using Correlational Constraints.” In: *arXiv preprint arXiv:1703.01671*.

Lantz, Björn (2013). “The large sample size fallacy.” In: *Scandinavian journal of caring sciences* 27.2, pp. 487–492.

Lazer, David, Ryan Kennedy, Gary King, and Alessandro Vespignani (2014). “The parable of Google Flu: traps in big data analysis.” In: *Science* 343.6176, pp. 1203–1205.

Lee, Brian K, Justin Lessler, and Elizabeth A Stuart (2011). "Weight trimming and propensity score weighting." In: *PloS one* 6.3.

Lee, Jason D, Dennis L Sun, Yuekai Sun, Jonathan E Taylor, et al. (2016). "Exact post-selection inference, with application to the lasso." In: *The Annals of Statistics* 44.3, pp. 907–927.

Lee, Jinhyuk, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang (2020). "BioBERT: a pre-trained biomedical language representation model for biomedical text mining." In: *Bioinformatics* 36.4, pp. 1234–1240.

Leetaru, Kalev and Philip A Schrodt (2013). "Gdelt: Global data on events, location, and tone, 1979–2012." In: *ISA annual convention*. Vol. 2. 4. Citeseer, pp. 1–49.

Lima, Luciano RS de, Alberto HF Laender, and Berthier A Ribeiro-Neto (1998). "A hierarchical approach to the automatic categorization of medical documents." In: *Proceedings of the seventh international conference on Information and knowledge management*, pp. 132–139.

Linden, Ilse van der, Hinda Haned, and Evangelos Kanoulas (2019). "Global aggregations of local explanations for black box models." In: *SIGIR Workshop on FACTS-IR*.

Lipton, Zachary C (2018). "The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery." In: *Queue* 16.3, pp. 31–57.

Lipton, Zachary C and Jacob Steinhardt (2019). "Troubling Trends in Machine Learning Scholarship: Some ML papers suffer from flaws that could mislead the public and stymie future research." In: *Queue* 17.1, pp. 45–77.

Liu, Jingshu, Zachariah Zhang, and Narges Razavian (2018). "Deep EHR: Chronic Disease Prediction Using Medical Notes." In: *Machine Learning for Healthcare Conference*, pp. 440–464.

Liu, Wendy and Derek Ruths (2013). "What's in a Name? Using First Names as Features for Gender Inference in Twitter." In: *AAAI spring symposium: Analyzing microtext*. Vol. 13, p. 01.

Lou, Yin, Rich Caruana, and Johannes Gehrke (2012). "Intelligible models for classification and regression." In: *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 150–158.

Loveys, Kate, Jonathan Torrez, Alex Fine, Glen Moriarty, and Glen Coppersmith (2018). "Cross-cultural differences in language markers of depression online." In: *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pp. 78–87.

Low, Yen Sia, Blanca Gallego, and Nigam Haresh Shah (2016). "Comparing high-dimensional confounder control methods for rapid cohort studies from electronic health records." In: *Journal of comparative effectiveness research* 5.2, pp. 179–192.

Madden, Mary (2012). "Privacy management on social media sites." In: *Pew Internet Report*, pp. 1–20.

Maerlender, A, L Flashman, A Kessler, S Kumbhani, R Greenwald, T Tosteson, and T McAllister (2010). "Examination of the construct validity of ImPACT computerized test, traditional, and experimental neuropsychological measures." In: *The Clinical Neuropsychologist* 24.8, pp. 1309–1325.

Malinsky, Daniel, Ilya Shpitser, and Eric J Tchetgen Tchetgen (2021). "Semiparametric inference for nonmonotone missing-not-at-random data: the no self-censoring model." In: *Journal of the American Statistical Association*, pp. 1–9.

Manski, Charles F (1990). "Nonparametric bounds on treatment effects." In: *The American Economic Review* 80.2, pp. 319–323.

Marwick, Alice E and danah boyd (2011). “I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience.” In: *New media and society* 13.1, pp. 114–133.

Matthay, Ellicott C and M Maria Glymour (2020). “A graphical catalog of threats to validity: Linking social science with epidemiology.” In: *Epidemiology (Cambridge, Mass.)* 31.3, p. 376.

McCaffrey, Daniel F, Greg Ridgeway, and Andrew R Morral (2004). “Propensity score estimation with boosted regression for evaluating causal effects in observational studies.” In: *Psychological methods* 9.4, p. 403.

McCorriston, James, David Jurgens, and Derek Ruths (2015). “Organizations Are Users Too: Characterizing and Detecting the Presence of Organizations on Twitter.” In: *ICWSM*, pp. 650–653.

McVeigh, Katharine H, Remle Newton-Dame, Pui Ying Chan, Lorna E Thorpe, Lauren Schreibstein, Kathleen S Tatem, Claudia Chernov, Elizabeth Lurie-Moroni, and Sharon E Perlman (2016). “Can electronic health records be used for population health surveillance? Validating population health metrics against established survey data.” In: *eGEMs* 4.1.

Meinshausen, Nicolai (2018). “Causality from a distributional robustness point of view.” In: *2018 IEEE Data Science Workshop (DSW)*. IEEE, pp. 6–10.

Melamud, Oren and Chaitanya Shivade (2019). “Towards Automatic Generation of Shareable Synthetic Clinical Notes Using Neural Language Models.” In: *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pp. 35–45.

Mensinger, Janell Lynn, Deanne Zotter Bonifazi, and Judith LaRosa (2007). “Perceived gender role prescriptions in schools, the superwoman ideal, and disordered eating among adolescent girls.” In: *Sex Roles* 57.7, pp. 557–568.

Meystre, Stéphane M, Guergana K Savova, Karin C Kipper-Schuler, and John F Hurdle (2008). “Extracting information from textual documents in the electronic health record: a review of recent research.” In: *Yearbook of medical informatics* 17.01, pp. 128–144.

Miao, Wang, Zhi Geng, and Eric J Tchetgen Tchetgen (2018). “Identifying causal effects with proxy variables of an unmeasured confounder.” In: *Biometrika* 105.4, pp. 987–993.

Michels, Karin B, Sander Greenland, and Bernard A Rosner (1998). “Does body mass index adequately capture the relation of body composition and body size to health outcomes?” In: *American Journal of Epidemiology* 147.2, pp. 167–172.

Mikal, Jude, Samantha Hurst, and Mike Conway (2016). “Ethical issues in using Twitter for population-level depression monitoring: a qualitative study.” In: *BMC medical ethics* 17.1, p. 22.

Mislove, Alan, Sune Lehmann, Yong-Yeol Ahn, Jukka-Pekka Onnela, and J Niels Rosenquist (2011). “Understanding the demographics of twitter users.” In: *ICWSM*.

Mittelstadt, Brent, Chris Russell, and Sandra Wachter (2019). “Explaining explanations in AI.” In: *Proceedings of the conference on fairness, accountability, and transparency*, pp. 279–288.

Mohri, Mehryar, Afshin Rostamizadeh, and Ameet Talwalkar (2018). *Foundations of machine learning*. MIT press.

Montavon, Grégoire, Alexander Binder, Sebastian Lapuschkin, Wojciech Samek, and Klaus-Robert Müller (2019). “Layer-wise relevance propagation: an overview.” In: *Explainable AI: interpreting, explaining and visualizing deep learning*, pp. 193–209.

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

Morstatter, Fred, Jürgen Pfeffer, Huan Liu, and Kathleen M Carley (2013). "Is the sample good enough? comparing data from twitter's streaming api with twitter's firehose." In: *ICWSM*.

Mozer, Reagan, Luke Miratrix, Aaron Russell Kaufman, and L Jason Anastasopoulos (2018). "Matching with text data: An experimental evaluation of methods for matching documents and of measuring match quality." In: *arXiv preprint arXiv:1801.00644*.

Muandet, Krikamol, David Balduzzi, and Bernhard Schölkopf (2013). "Domain generalization via invariant feature representation." In: *International Conference on Machine Learning*. PMLR, pp. 10–18.

Mueller, Aaron, Zach Wood-Doughty, Silvio Amir, Mark Dredze, and Alicia L Nobles (2021). "Demographic Representation and Collective Storytelling in the Me Too Twitter Hashtag Activism Movement." In: *Proceedings of the ACM on Human-Computer Interaction* 5.CSCW.

Mullenbach, James, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein (2018). "Explainable Prediction of Medical Codes from Clinical Text." In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 1101–1111.

Nabi, Razieh, Daniel Malinsky, and Ilya Shpitser (2019). "Optimal Training of Fair Predictive Models." In: *arXiv preprint arXiv:1910.04109*.

Nabi, Razieh, Todd McNutt, and Ilya Shpitser (2017). "Semiparametric Causal Sufficient Dimension Reduction Of High Dimensional Treatments." In: *arXiv preprint arXiv:1710.06727*.

Nabi, Razieh and Ilya Shpitser (2017). "Semi-Parametric Causal Sufficient Dimension Reduction Of High Dimensional Treatments." In: *arXiv preprint arXiv:1710.06727*.

— (2018). "Fair inference on outcomes." In: *Thirty-Second AAAI Conference on Artificial Intelligence*.

Nagata, Jason M, Isabel Hernández-Ramos, Anand Sivasankara Kurup, Daniel Albrecht, Claudia Vivas-Torrealba, and Carlos Franco-Paredes (2013). "Social determinants of health and seasonal influenza vaccination in adults over 65 years: a systematic review of qualitative and quantitative data." In: *BMC Public Health* 13.1, p. 388.

Nan, Xiaoli (2012). "Communicating to young adults about HPV vaccination: Consideration of message framing, motivation, and gender." In: *Health Communication* 27.1, pp. 10–18.

Neal, Brady, Chin-Wei Huang, and Sunand Raghupathi (2020). "RealCause: Realistic Causal Inference Benchmarking." In: *arXiv preprint arXiv:2011.15007*.

Neyman, Jerzy (1923). "Sur les applications de la thar des probabilities aux experiences Agaricales: Essay des principle. Excerpts reprinted (1990) in English." In: *Statistical Science* 5, pp. 463–472.

O'Connor, Brendan, Ramnath Balasubramanyan, Bryan R Routledge, and Noah A Smith (2010). "From tweets to polls: Linking text sentiment to public opinion time series." In: *ICWSM* 11.122-129, pp. 1–2.

Obermeyer, Ziad and Ezekiel J Emanuel (2016). "Predicting the future – data, machine learning, and clinical medicine." In: *NEJM* 375.13, p. 1216.

Oktay, Hüseyin, Akanksha Atrey, and David Jensen (2019). "Identifying When Effect Restoration Will Improve Estimates of Causal Effect." In: *Proceedings of the 2019 SIAM International Conference on Data Mining*. SIAM, pp. 190–198.

Olteanu, Alexandra, Carlos Castillo, Fernando Diaz, and Emre Kiciman (2019). "Social data: Biases, methodological pitfalls, and ethical boundaries." In: *Frontiers in Big Data* 2, p. 13.

Osborne, Miles, Sean Moran, Richard McCreadie, Alexander Von Lunen, Martin D Sykora, Elizabeth Cano, Neil Ireson, Craig Macdonald, Iadh Ounis, Yulan He, et al. (2014). "Real-time detection, tracking, and monitoring of automatically discovered events in social media." In: *ACL*.

Parrish, Nathan H, Anna L Buczak, Jared T Zook, James P Howard, Brian J Ellison, and Benjamin D Baugher (2018). "Crystal Cube: Multidisciplinary Approach to Disruptive Events Prediction." In: *International Conference on Applied Human Factors and Ergonomics*. Springer, pp. 571–581.

Patki, Neha, Roy Wedge, and Kalyan Veeramachaneni (2016). "The synthetic data vault." In: *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE, pp. 399–410.

Paul, Michael J (2017). "Feature Selection as Causal Inference: Experiments with Text Classification." In: *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pp. 163–172.

Paul, Michael J and Mark Dredze (2011). "You are what you tweet: Analyzing twitter for public health." In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 5.

Pavalanathan, Umashanthi and Jacob Eisenstein (2015). "Confounds and consequences in geotagged Twitter data." In: *arXiv:1506.02275*.

Pearl, Judea (1995). "Causal diagrams for empirical research." In: *Biometrika* 82.4, pp. 669–688.

— (2009). *Causality*. Cambridge university press.

— (2010). "On measurement bias in causal inference." In: *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*. AUAI Press, pp. 425–432.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. (2011). "Scikit-learn: Machine learning in Python." In: *Journal of machine learning research* 12.Oct, pp. 2825–2830.

Pennacchiotti, Marco and Ana-Maria Popescu (2011). "A machine learning approach to twitter user classification." In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 5.

Preoţiuc-Pietro, Daniel, Sharath Chandra Guntuku, and Lyle Ungar (Sept. 2017). "Controlling Human Perception of Basic User Traits." In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, pp. 2335–2341. DOI: 10.18653/v1/D17-1248.

Preoţiuc-Pietro, Daniel and Lyle Ungar (Aug. 2018). "User-Level Race and Ethnicity Predictors from Twitter Text." In: *Proceedings of the 27th International Conference on Computational Linguistics*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, pp. 1534–1545.

Pruthi, Danish, Mansi Gupta, Bhuwan Dhingra, Graham Neubig, and Zachary C. Lipton (July 2020). "Learning to Deceive with Attention-Based Explanations." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 4782–4793. DOI: 10.18653/v1/2020.acl-main.432.

Pryzant, Reid, 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*, pp. 4095–4109.

Quinonero-Candela, Joaquin, Masashi Sugiyama, Neil D Lawrence, and Anton Schwaighofer (2009). *Dataset shift in machine learning*. Mit Press.

Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever (2019). "Language models are unsupervised multitask learners." In: *OpenAI Blog* 1.8, p. 9.

Rajkomar, Alvin, Eyal Oren, Kai Chen, Andrew M Dai, Nissan Hajaj, Michaela Hardt, Peter J Liu, Xiaobing Liu, Jake Marcus, Mimi Sun, et al. (2018). "Scalable and accurate deep learning with electronic health records." In: *NPJ Digital Medicine* 1.1, p. 18.

Rajpurkar, Pranav, Jian Zhang, Konstantin Lopyrev, and Percy Liang (2016). "SQuAD: 100,000+ Questions for Machine Comprehension of Text." In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2383–2392.

Raleigh, Clionadh and Caitriona Dowd (June 2021). *Armed conflict location and event data project (ACLED)*.

Rao, Delip, David Yarowsky, Abhishek Shreevats, and Manaswi Gupta (2010). "Classifying latent user attributes in twitter." In: *Proceedings of the 2nd international workshop on Search and mining user-generated contents*. ACM, pp. 37–44.

Raschka, Sebastian (2014). "Naive bayes and text classification i-introduction and theory." In: *arXiv preprint arXiv:1410.5329*.

Rego, Richard (2018). "Changing forms and platforms of misogyny: Sexual harassment of women journalists on twitter." In: *Media Watch* 9.3, pp. 472–85.

Rehkopf, David H, M Maria Glymour, and Theresa L Osypuk (2016). "The consistency assumption for causal inference in social epidemiology: when a rose is not a rose." In: *Current epidemiology reports* 3.1, pp. 63–71.

Reid, Stephen, Jonathan Taylor, and Robert Tibshirani (2017). "Post-selection point and interval estimation of signal sizes in Gaussian samples." In: *Canadian Journal of Statistics* 45.2, pp. 128–148.

Ribeiro, Marco Tulio, 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*, pp. 1135–1144.

Roberts, Margaret E, Brandon M Stewart, and Richard A Nielsen (2018). "Adjusting for Confounding with Text Matching." In.

Robins, James (1986). "A new approach to causal inference in mortality studies with a sustained exposure period with application to control of the healthy worker survivor effect." In: *Mathematical modelling* 7.9-12, pp. 1393–1512.

Robins, James M, Andrea Rotnitzky, and Daniel O Scharfstein (2000). "Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models." In: *Statistical models in epidemiology, the environment, and clinical trials*. Springer, pp. 1–94.

Rosenbaum, Paul (1984). "The consequences of adjustment for a concomitant variable that has been affected by the treatment." In: *Journal of the Royal Statistical Society: Series A (General)* 147.5, pp. 656–666.

Rosenbaum, Paul and Donald Rubin (1983). "The central role of the propensity score in observational studies for causal effects." In: *Biometrika* 70.1, pp. 41–55.

Rosenbloom, S Trent, Joshua C Denny, Hua Xu, Nancy Lorenzi, William W Stead, and Kevin B Johnson (2011). "Data from clinical notes: a perspective on the tension between structure and flexible documentation." In: *Journal of the American Medical Informatics Association* 18.2, pp. 181–186.

Rout, Dominic, Kalina Bontcheva, Daniel Preoţiuc-Pietro, and Trevor Cohn (2013). "Where's@ wally?: a classification approach to geolocating users based on their social ties." In: *Proceedings of the 24th ACM Conference on Hypertext and Social Media*. ACM, pp. 11–20.

Rubin, Donald (1976). "Causal Inference and Missing Data (with discussion)." In: *Biometrika* 63, pp. 581–592.

Ruch, Patrick, Robert Baud, and Antoine Geissbühler (2003). "Using lexical disambiguation and named-entity recognition to improve spelling correction in the electronic patient record." In: *Artificial intelligence in medicine* 29.1-2, pp. 169–184.

Rudin, Cynthia (2019). "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead." In: *Nature Machine Intelligence* 1.5, pp. 206–215.

Saha, Koustuv, Eshwar Chandrasekharan, and Munmun De Choudhury (2019). "Prevalence and psychological effects of hateful speech in online college communities." In: *Proceedings of the 10th ACM conference on web science*, pp. 255–264.

Samory, Mattia, Indira Sen, Julian Kohne, Fabian Flöck, and Claudia Wagner (2021). "Call me sexist, but…: Revisiting sexism detection using psychological scales and adversarial samples." In: *Intl AAAI Conf. Web and Social Media*, pp. 573–584.

Sanh, Victor, Lysandre Debut, Julien Chaumond, and Thomas Wolf (2019). "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter." In: *arXiv preprint arXiv:1910.01108v4*.

Sani, Numair, Daniel Malinsky, and Ilya Shpitser (2020). "Explaining the behavior of black-box prediction algorithms with causal learning." In: *arXiv preprint arXiv:2006.02482*.

Sap, Maarten, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith (2019). "The risk of racial bias in hate speech detection." In: *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 1668–1678.

Satcher, David (2001). *Mental health: Culture, race, and ethnicity. Supplement to mental health: A report of the surgeon general*.

Savova, Guergana K, James J Masanz, Philip V Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C Kipper-Schuler, and Christopher G Chute (Sept. 2010). "Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications." In: *JAMIA* 17.5, pp. 507–513. DOI: 10.1136/jamia.2009.001560.

Scheurwegs, Elyne, Kim Luyckx, Léon Luyten, Walter Daelemans, and Tim Van den Bulcke (2016). "Data integration of structured and unstructured sources for assigning clinical codes to patient stays." In: *Journal of the American Medical Informatics Association* 23.e1, e11–e19.

Schöttker, Ben, Kai-Uwe Saum, Laura Perna, José Manuèl Ordóñez-Mena, Bernd Holleczek, and Hermann Brenner (2014). "Is vitamin D deficiency a cause of increased morbidity and mortality at older age or simply an indicator of poor health?" In: *European journal of epidemiology* 29.3, pp. 199–210.

Selbst, Andrew D, Danah Boyd, Sorelle A Friedler, Suresh Venkatasubramanian, and Janet Vertesi (2019). "Fairness and abstraction in sociotechnical systems." In: *Proceedings of the conference on fairness, accountability, and transparency*, pp. 59–68.

Shalev-Shwartz, Shai and Shai Ben-David (2014). *Understanding machine learning: From theory to algorithms*. Cambridge university press.

Shen, Dinggang, Guorong Wu, and Heung-Il Suk (2017). "Deep learning in medical image analysis." In: *Annual review of biomedical engineering* 19, pp. 221–248.

Sheng, Emily, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng (2019). "The Woman Worked as a Babysitter: On Biases in Language Generation." In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3398–3403.

Shi, Xu, Wang Miao, Jennifer C Nelson, and Eric J Tchetgen Tchetgen (2020). "Multiply robust causal inference with double-negative control adjustment for categorical unmeasured confounding." In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 82.2, pp. 521–540.

Shimoni, Yishai, Chen Yanover, Ehud Karavani, and Yaara Goldschmnidt (2018). "Benchmarking framework for performance-evaluation of causal inference analysis." In: *arXiv preprint arXiv:1802.05046*.

Shpitser, Ilya, Karthika Mohan, and Judea Pearl (2015). "Missing data as a causal and probabilistic problem." In: *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence*. AUAI Press, pp. 802–811.

Shpitser, Ilya and Judea Pearl (2008). "Complete identification methods for the causal hierarchy." In: *Journal of Machine Learning Research* 9.Sep, pp. 1941–1979.

Shrikumar, Avanti, Peyton Greenside, Anna Shcherbina, and Anshul Kundaje (2016). "Not just a black box: Learning important features through propagating activation differences." In: *arXiv preprint arXiv:1605.01713*.

Silberman, M Six, Bill Tomlinson, Rochelle LaPlante, Joel Ross, Lilly Irani, and Andrew Zaldivar (2018). "Responsible research with crowds: pay crowdworkers at least minimum wage." In: *Communications of the ACM* 61.3, pp. 39–41.

Silva, Lindsay M, Marianne Coolman, Eric AP Steegers, Vincent WV Jaddoe, Henriette A Moll, Albert Hofman, Johan P Mackenbach, and Hein Raat (2008). "Low socioeconomic status is a risk factor for preeclampsia: the Generation R Study." In: *Journal of hypertension* 26.6, pp. 1200–1208.

Simonyan, Karen, Andrea Vedaldi, and Andrew Zisserman (2014). "Deep inside convolutional networks: Visualising image classification models and saliency maps." In: *In Workshop at International Conference on Learning Representations*. Citeseer.

Simpson, Edward H (1951). "The interpretation of interaction in contingency tables." In: *Journal of the Royal Statistical Society: Series B (Methodological)* 13.2, pp. 238–241.

Sinha, Samiran and Yanyuan Ma (2014). "Semiparametric analysis of linear transformation models with covariate measurement errors." In: *Biometrics* 70.1, pp. 21–32.

Sinnenberg, Lauren, Alison M Buttenheim, Kevin Padrez, Christina Mancheno, Lyle Ungar, and Raina M Merchant (2017). "Twitter as a tool for health research: a systematic review." In: *American journal of public health* 107.1, e1–e8.

Slack, Dylan, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju (2020). "Fooling lime and shap: Adversarial attacks on post hoc explanation methods." In: *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pp. 180–186.

Sloan, Luke, Jeffrey Morgan, William Housley, Matthew Williams, Adam Edwards, Pete Burnap, and Omer Rana (2013). "Knowing the tweeters: Deriving sociologically relevant demographics from Twitter." In: *Sociological research online* 18.3, p. 7.

Smith, Michael C (2020). "Motivating the Inclusion of Construct Validity in Social Media Monitoring: How to Leverage Surveys to Measure Trust in Vaccines on Twitter." PhD thesis. The George Washington University.

Smith, Michael C, Thomas A Mazzuchi, and David A Broniatowski (2020). "Validating Social Media Monitoring: Statistical Pitfalls and Opportunities from Public Opinion." In:

International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation. Springer, pp. 65–74.

Snow, Rion, Brendan O'connor, Dan Jurafsky, and Andrew Y Ng (2008). "Cheap and fast–but is it good? evaluating non-expert annotations for natural language tasks." In: *Proceedings of the 2008 conference on empirical methods in natural language processing*, pp. 254–263.

Stefanski, Leonard A and Raymond J Carroll (1985). "Covariate measurement error in logistic regression." In: *The Annals of Statistics*, pp. 1335–1351.

Stern, Ariel Dora and W Nicholson Price (2019). "Regulatory oversight, causal inference, and safe and effective health care machine learning." In: *Biostatistics*.

Stewart, Ian, Diyi Yang, and Jacob Eisenstein (2020). "Characterizing collective attention via descriptor context: A case study of public discussions of crisis events." In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 14, pp. 650–660.

Stuart, Elizabeth A (2010). "Matching methods for causal inference: A review and a look forward." In: *Statistical science: a review journal of the Institute of Mathematical Statistics* 25.1, p. 1.

Subbaswamy, Adarsh, Peter Schulam, and Suchi Saria (2019). "Preventing Failures Due to Dataset Shift: Learning Predictive Models That Transport." In: *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 3118–3127.

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

Tonekaboni, Sana, Shalmali Joshi, Melissa D McCradden, and Anna Goldenberg (2019). "What clinicians want: contextualizing explainable machine learning for clinical end use." In: *Machine learning for healthcare conference*. PMLR, pp. 359–380.

Topaz, Maxim, Leah Shafran-Topaz, and Kathryn H Bowles (2013). "ICD-9 to ICD-10: evolution, revolution, and current debates in the United States." In: *Perspectives in health information management/AHIMA, American Health Information Management Association* 10.Spring.

Tsiatis, Anastasios (2007). *Semiparametric theory and missing data*. Springer Science and Business Media.

Usher, Nikki, Jesse Holcomb, and Justin Littman (2018). "Twitter makes it worse: Political journalists, gendered echo chambers, and the amplification of gender bias." In: *The international journal of press/politics* 23.3, pp. 324–344.

VanderWeele, Tyler J (2009). "Concerning the consistency assumption in causal inference." In: *Epidemiology* 20.6, pp. 880–883.

Vansteelandt, Stijn, Tyler J VanderWeele, Eric J Tchetgen Tchetgen, and James M Robins (2008). "Multiply robust inference for statistical interactions." In: *Journal of the American Statistical Association* 103.484, pp. 1693–1704.

Vargas, Nicholas and Kevin Stainback (2016). "Documenting contested racial identities among self-identified Latina/os, Asians, Blacks, and Whites." In: *American Behavioral Scientist* 60.4, pp. 442–464.

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

Vig, Jesse, 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: *NeurIPS*.

Volkova, Svitlana and Yoram Bachrach (2015). “On predicting sociodemographic traits and emotions from communications in social networks and their implications to online self-disclosure.” In: *Cyberpsychology, Behavior, and Social Networking* 18.12, pp. 726–736.

Volkova, Svitlana, Glen Coppersmith, and Benjamin Van Durme (June 2014). “Inferring User Political Preferences from Streaming Communications.” In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Baltimore, Maryland: Association for Computational Linguistics, pp. 186–196. DOI: 10.3115/v1/P14-1018.

Volkova, Svitlana, Theresa Wilson, and David Yarowsky (Oct. 2013). “Exploring Demographic Language Variations to Improve Multilingual Sentiment Analysis in Social Media.” In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Seattle, Washington, USA: Association for Computational Linguistics, pp. 1815–1827.

Wager, Stefan and Susan Athey (2017). “Estimation and inference of heterogeneous treatment effects using random forests.” In: *Journal of the American Statistical Association*.

Wallach, Hanna M (2006). “Topic modeling: beyond bag-of-words.” In: *ICML*, pp. 977–984.

Wang, Dingquan and Jason Eisner (Oct. 2018). “Synthetic Data Made to Order: The Case of Parsing.” In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, pp. 1325–1337. DOI: 10.18653/v1/D18-1163.

Wang, Xuan and Qihua Wang (2015). “Semiparametric linear transformation model with differential measurement error and validation sampling.” In: *Journal of Multivariate Analysis* 141, pp. 67–80.

Wang, Zhao and Aron Culotta (2019). “When Do Words Matter? Understanding the Impact of Lexical Choice on Audience Perception Using Individual Treatment Effect Estimation.” In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33, pp. 7233–7240.

Wang, Zijian, Scott A. Hale, David Ifeoluwa Adelani, Przemyslaw A. Grabowicz, Timo Hartmann, Fabian Flöck, and David Jurgens (2019). “Demographic Inference and Representative Population Estimates from Multilingual Social Media Data.” In: *The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*. Ed. by Ling Liu, Ryen W. White, Amin Mantrach, Fabrizio Silvestri, Julian J. McAuley, Ricardo Baeza-Yates, and Leila Zia. ACM, pp. 2056–2067. DOI: 10.1145/3308558.3313684.

Weld, Galen, Peter West, Maria Glenski, David Arbour, Ryan Rossi, and Tim Althoff (2020). “Adjusting for Confounders with Text: Challenges and an Empirical Evaluation Framework for Causal Inference.” In: *arXiv preprint arXiv:2009.09961*.

Wendling, T, K Jung, A Callahan, A Schuler, NH Shah, and B Gallego (2018). “Comparing methods for estimation of heterogeneous treatment effects using observational data from health care databases.” In: *Statistics in medicine* 37.23, pp. 3309–3324.

Wiegreffe, Sarah, Edward Choi, Sherry Yan, Jimeng Sun, and Jacob Eisenstein (2019). “Clinical Concept Extraction for Document-Level Coding.” In: *Proceedings of the 18th BioNLP Workshop and Shared Task*, pp. 261–272.

Wiegreffe, Sarah and Yuval Pinter (Nov. 2019). “Attention is not not Explanation.” In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, pp. 11–20. DOI: 10.18653/v1/D19-1002.

Willett, Walter (1989). “An overview of issues related to the correction of non-differential exposure measurement error in epidemiologic studies.” In: *Statistics in Medicine* 8.9, pp. 1031–1040.

Winata, Genta Indra, Andrea Madotto, Chien-Sheng Wu, and Pascale Fung (Nov. 2019). “Code-Switched Language Models Using Neural Based Synthetic Data from Parallel Sentences.” In: *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*. Hong Kong, China: Association for Computational Linguistics, pp. 271–280. DOI: 10.18653/v1/K19-1026.

Wood-Doughty, Zach, Nicholas Andrews, Rebecca Marvin, and Mark Dredze (2018). “Predicting Twitter User Demographics from Names Alone.” In: *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, pp. 105–111.

Wood-Doughty, Zach, Isabel Cachola, and Mark Dredze (2021). “Faithful and Plausible Explanations of Medical Code Predictions.” In: *arXiv preprint arXiv:2104.07894*.

Wood-Doughty, Zach, Praateek Mahajan, and Mark Dredze (2018). “Johns Hopkins or johnny-hopkins: Classifying Individuals versus Organizations on Twitter.” In: *PEOPLES*, pp. 56–61.

Wood-Doughty, Zach, Ilya Shpitser, and Mark Dredze (2018). “Challenges of Using Text Classifiers for Causal Inference.” In: *EMNLP*, pp. 4586–4598.

— (2020). “Sensitivity Analyses for Incorporating Machine Learning Predictions into Causal Estimates.” In: *NeurIPS 2020 Workshop on Causal Discovery and Causality-Inspired Machine Learning*.

— (2021). “Generating Synthetic Text Data to Evaluate Causal Inference Methods.” In: *arXiv preprint arXiv:2102.05638*.

Wood-Doughty, Zach, Michael Smith, David Broniatowski, and Mark Dredze (2017). “How Does Twitter User Behavior Vary Across Demographic Groups?” In: *Proceedings of the Second Workshop on NLP and Computational Social Science*, pp. 83–89.

Wood-Doughty, Zach, Paiheng Xu, Xiao Liu, and Mark Dredze (2021). “Using Noisy Self-Reports to Predict Twitter User Demographics.” In: *SocialNLP 2021*, p. 123.

World Development Indicators (2021). DataBank.

Wu, Chia-Yi, Chin-Kuo Chang, Debbie Robson, Richard Jackson, Shaw-Ji Chen, Richard D Hayes, and Robert Stewart (2013). “Evaluation of smoking status identification using electronic health records and open-text information in a large mental health case register.” In: *PloS one* 8.9.

Wu, Siqi, Marian-Andrei Rizoiu, and Lexing Xie (2020). “Variation across scales: Measurement fidelity under twitter data sampling.” In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 14, pp. 715–725.

Xu, Peng, Mostofa Patwary, Mohammad Shoeybi, Raul Puri, Pascale Fung, Animashree Anandkumar, and Bryan Catanzaro (2020). “Controllable Story Generation with External Knowledge Using Large-Scale Language Models.” In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2831–2845.

Yang, Mochen, Gediminas Adomavicius, Gordon Burtch, and Yuqing Ren (2018). “Mind the gap: Accounting for measurement error and misclassification in variables generated via data mining.” In: *Information Systems Research* 29.1, pp. 4–24.

Yang, Mochen, Edward McFowland, Gordon Burtch, and Gediminas Adomavicius (2019). “Achieving Reliable Causal Inference with Data-Mined Variables: A Random Forest

Approach to the Measurement Error Problem." In: *Kelley School of Business Research Paper* 19-20.

Yang, Shu and Peng Ding (2019). "Combining multiple observational data sources to estimate causal effects." In: *Journal of the American Statistical Association*, pp. 1–33.

— (2020). "Combining multiple observational data sources to estimate causal effects." In: *Journal of the American Statistical Association* 115.531, pp. 1540–1554.

Yao, Liuyi, Sheng Li, Yaliang Li, Hongfei Xue, Jing Gao, and Aidong Zhang (2019). "On the estimation of treatment effect with text covariates." In: *Proceedings of the 28th International Joint Conference on Artificial Intelligence*. AAAI Press, pp. 4106–4113.

Zhang, Danchen, Daqing He, Sanqiang Zhao, and Lei Li (2017). "Enhancing automatic ICD-9-CM code assignment for medical texts with PubMed." In: *BioNLP 2017*, pp. 263–271.

Zhang, Xiang, Junbo Zhao, and Yann LeCun (2015). "Character-level convolutional networks for text classification." In: *Advances in neural information processing systems*, pp. 649–657.

Zhang, Yijia, Qingyu Chen, Zhihao Yang, Hongfei Lin, and Zhiyong Lu (2019). "BioWordVec, improving biomedical word embeddings with subword information and MeSH." In: *Scientific data* 6.1, pp. 1–9.

Zheng, Kai, David A Hanauer, Rema Padman, Michael P Johnson, Anwar A Hussain, Wen Ye, Xiaomu Zhou, and Herbert S Diamond (2011). "Handling anticipated exceptions in clinical care: investigating clinician use of 'exit strategies' in an electronic health records system." In: *Journal of the American Medical Informatics Association* 18.6, pp. 883–889.

Zhong, Ruiqi, Steven Shao, and Kathleen McKeown (2019). "Fine-grained sentiment analysis with faithful attention." In: *arXiv preprint arXiv:1908.06870*.

Zhuang, Fuzhen, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He (2020). "A comprehensive survey on transfer learning." In: *Proceedings of the IEEE* 109.1, pp. 43–76.

Vita

Zach Wood-Doughty received a B.A. degree in Computer Science and Mathematics from Carleton College in 2014 and a M.S.E. degree from Johns Hopkins University in 2017 after enrolling the Ph.D. program in 2016. Zach's research explores how statistical models of text can be incorporated into causal analyses to draw insights from large datasets such as social media or clinical notes.