# SOCIAL MEASUREMENT AND CAUSAL INFERENCE WITH TEXT

A Dissertation Presented

by

KATHERINE A. KEITH

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of

DOCTOR OF PHILOSOPHY

September 2021

College of Information and Computer Sciences

# SOCIAL MEASUREMENT AND CAUSAL INFERENCE WITH TEXT

A Dissertation Presented

by

KATHERINE A. KEITH

Approved as to style and content by:

Brendan O'Connor, Chair

David Jensen, Member

Mohit Iyyer, Member

Douglas Rice, Member

James Allan, Chair of the Faculty
College of Information and Computer Sciences

# ACKNOWLEDGMENTS

First, thank you to my advisor, Brendan O'Connor. His focus on high-quality research has expanded my technical abilities beyond what I imagined possible five years ago and his attention to detail has helped shape me into an assiduous scholar. Thank you to my committee members: David Jensen's wisdom and tangible excitement about causal inference have deeply shaped the way I do research; Doug Rice's support and insight into social science applications with text data have kept me motivated in the final stretches of this Ph.D. process; and Mohit Iyyer's presence and leadership within the NLP group at UMass have provided invaluable learning opportunities. Having my labmates—Su Lin Blodgett and Abe Handler—along for the entirety of this Ph.D. journey has been such a gift, particularly our conversations about the purpose of NLP as a field and its ensuing societal impact.

Thank you to my peer network at UMass CICS—particularly Sam Witty, Emma Strubell, Pat Verga, Neha Nayak, Kevin Winner, Conrad Holtsclaw, Justin Svegliato, Ian Gemp, Lucas Chaufournier, Rian Shambaugh, Sheikh Sarwar, Hia Ghosh, David Tench, Kalpesh Krishna, and Nader Akoury—and peers in the broader NLP community—particularly Andrew Halterman, Malihe Alikhani, Zach Wood-Doughty, and Lucy Li—who have both kept me hungry for better science and have also been extremely supportive of the balanced lifestyle I seek to live. Thank you to the CSWomen organization at UMass, whose meetings and empathy were invaluable in times of doubt. Thank you to the UMass CICS staff—particularly Leeanne Leclerc, Eileen Hamel, and Malaika Ross—for keeping me on track and ensuring so many logistics in our college run smoothly. Thank you to the Data Science and AI groups at Bloomberg—particularly my mentors Amanda Stent and Edgar Meij—who believed

in me enough to give me my first internships. I am indebted to my undergraduate mentors at Lewis & Clark College: Paul T. Allen, who encouraged me to apply to computer science Ph.D. programs even though I thought I was not qualified, and Cliff Bekar, whose research first introduced me to computational social science and lit a spark of curiosity that has lasted all these years.

Finally, I imagine I would not have accomplished nearly as much in my academic career without my family. To my sister, Ginny Keith, who was always there for the phone-calls and late night conversations when I encountered tough times. To my father, Kurt Keith, who taught me dedication, hard work, and the pursuit of excellence. To my mother and seventh grade math teacher, Anne Keith, who laid the foundation for me to believe in myself as a woman capable of doing mathematical and computational work. I am forever grateful to my parents who—from the moment they chose the name for their daughter that would sound best with a "Dr." in front of it—have taught me to dream big and provided me with unwavering support to turn those dreams into reality.

# ABSTRACT

## SOCIAL MEASUREMENT AND CAUSAL INFERENCE WITH TEXT

SEPTEMBER 2021

KATHERINE A. KEITH

B.A., LEWIS & CLARK COLLEGE

M.S., UNIVERSITY OF MASSACHUSETTS AMHERST

Ph.D., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor Brendan O'Connor

The digital age has dramatically increased access to large-scale collections of digitized text documents. These corpora include, for example, digital traces from social media, decades of archived news reports, and transcripts of spoken interactions in political, legal, and economic spheres. For social scientists, this new widespread data availability has potential for improved quantitative analysis of relationships between language use and human thought, actions, and societal structure. However, the large-scale nature of these collections means that traditional manual approaches to analyzing content are extremely costly and do not scale. Furthermore, incorporating unstructured text data into quantitative analysis is difficult due to texts' high-dimensional nature and linguistic complexity.

This thesis blends (a) the computational strengths of natural language processing (NLP) and machine learning to automate and scale-up quantitative text analysis with

(b) two themes central to social scientific studies but often under-addressed in NLP: measurement—creating quantifiable summaries of empirical phenomena—and causal inference—estimating the effects of interventions. First, we address measuring class prevalence in document collections; we contribute a generative probabilistic modeling approach to prevalence estimation and show empirically that our model is more robust to shifts in class priors between training and inference. Second, we examine cross-document entity-event measurement; we contribute an empirical pipeline and a novel latent disjunction model to identify the names of civilians killed by police from our corpus of web-scraped news reports. Third, we gather and categorize applications that use text to reduce confounding from causal estimates and contribute a list of open problems as well as guidance about data processing and evaluation decisions in this area. Finally, we contribute a new causal research design to estimate the natural indirect and direct effects of social group signals (e.g. race or gender) on conversational outcomes with separate aspects of language as causal mediators; this chapter is motivated by a theoretical case study of U.S. Supreme Court oral arguments and the effect of an advocate's gender on interruptions from justices. We conclude by discussing the relationship between measurement and causal inference with text and future work at this intersection.

# TABLE OF CONTENTS

# LIST OF TABLES

**Table** **Page**

# LIST OF FIGURES

**Figure** **Page**

# CHAPTER 1

# INTRODUCTION

Language is an inherently social process that underlies most human interactions. As such, analysis of written language can provide insight into relationships between language use and human thought, actions, and societal structure. For instance, politics relies on language—candidates debate policy, representatives write legislation, nations negotiate peace treaties, and media outlets report on international relations [Grimmer and Stewart, 2013]. In economics, product reviews can provide insight into consumer decision making, public company filings insight into asset price movements, and media sentiment insight into the stock market [Gentzkow et al., 2019]. In sociology, communication within and between groups underlies collective action, social relationships, and social roles [Evans and Aceves, 2016]. This importance of language in unpacking human thought, behavior, and society has led to decades of manual analysis of text by social scientists and numerous academic guidebooks on the subject, e.g. Neuendorf [2017], Krippendorff [2018].

Digital collections of text and other social data have dramatically increased in the last few decades. Social data now includes large-scale business and government records of digital traces—byproducts of humans' everyday actions that are stored digitally [Sandvig and Hargittai, 2015, Salganik, 2017, Olteanu et al., 2019]. Advances in technology such as the digitization of historical documents via optical character recognition [Mori et al., 1999] and digital systems (e.g. social media) that record user-generated language [Sandvig and Hargittai, 2015] have greatly increased the amount of text to which researchers have access. This explosion of data has been one of the

catalysts of the academic field of *computational social science* [Lazer et al., 2009], and as Watts [2011] speculates,

> Rather, just as the invention of the telescope revolutionized the study of the heavens, so too by rendering the unmeasurable measurable, the technological revolution in mobile, Web, and Internet communications has the potential to revolutionize our understanding of ourselves and how we interact.

Yet, in the study of the portion of "the heavens" that is language, a "telescope" of *manual* content analysis simply does not scale. Instead, many have turned to computational methods from natural language processing (NLP) [Jurafsky and Martin, 2019, Eisenstein, 2019] to automate and scale-up analysis of text. For this set of "text-as-data" methods, statistical models of language are built and deployed, typically via computer programs [Grimmer and Stewart, 2013, Grimmer et al., 2021]. In one of the earliest text-as-data applications, Mosteller and Wallace [1963] apply statistical text analysis to infer the unknown authorship of certain Federalist Papers. Since then, automated text-as-data methods have swept across the social sciences [O'Connor et al., 2011, Grimmer and Stewart, 2013, Evans and Aceves, 2016, Gentzkow et al., 2019, Nguyen et al., 2020]. These methods have been crucial in studies of large-scale collections of text including: studying the nature of online censorship in China with 11 million social media posts [King et al., 2013], studying racial disparities in police officers language with roughly 37,000 spoken utterances [Voigt et al., 2017], and studying drivers of newspapers' political slant with one year's worth of articles from over 400 newspapers [Gentzkow and Shapiro, 2010].

Despite this growing interest in text-as-data methods as the "telescope" that could provide insight into human behavior and society, text-as-data methods are often designed for a different purpose than they are used for by social scientists. As Antoniak and Mimno [2018] describe, many methods in NLP are "downsteam-centered" in which the end goal is improving predictive performance on a more complicated downstream task. In contrast, many social science applications are "corpus-centered" in

which the end goal is to use NLP methods to provide evidence about the nature of the author's thought, culture, or linguistic tendencies.

This thesis aims to help close this gap between how the text-as-data "telescope" is designed and used. In particular, this thesis focuses on two themes central to social scientific studies but often under-addressed in NLP—measurement (§1.1) and casual inference (§1.2). In the remainder of this introduction, we provide definitions and several challenges of measurement and causal inference with text. While these are themes that could span entire book chapters (e.g. Grimmer et al. [2021]), we describe this thesis's particular conceptual, methodological, and empirical contributions along these two themes. We wrap-up the introduction with a thesis statement that synthesizes these ideas (§1.3).

## 1.1 Measurement with text

**Definition.** Central to analysis of text data is *measurement*—creating quantifiable summaries of empirical phenomena. Measurement has a long history, dating back to Stevens [1946]: "measurement, in the broadest sense, is defined as assignment of numerals to objects or events according to rules." Dimensionality reduction is essential to measurement in the social sciences. Patty and Penn [2015] discuss how empirical analysis of social constructs require a "data reduction" of higher-dimensional data into lower-dimensional measures; and Grimmer et al. [2021] emphasize that for text specifically "measurement is fundamentally about compression" in which one throws away specific information to focus on a generalizable property.[1]

Accurate and valid measurement at scale is key in text-as-data studies. Revisiting the studies we previously highlighted, King et al. [2013] aim to understand the nature

[1]Measurement is a concept closely related to the *measurement modeling* literature in the social sciences. Measurement modeling consists of mapping observable data to theoretical constructs and emphasizes the importance of *validity* (is it right?) and *reliability* (can it be repeated?) [Loevinger, 1957, Messick, 1987, Quinn et al., 2010, Jacobs and Wallach, 2021].

of the content censored by the Chinese government and thus develop measures for the censorship magnitude of specific topic areas (e.g. call for collective action or critique of the state). Voigt et al. [2017] aim to understand how police officers speak differently to citizens of different races and thus develop linguistic measures of respect. Gentzkow and Shapiro [2010] aim to understand how economic market forces determine the political ideology of news outlets and thus develop a measure of ideological slant. Across these three examples, the measures in question—censorship topic areas, respect, and ideological slant—are complicated social constructs that require a rich knowledge of how language is created and social theory for why these measures are important.

We formally define *measurement* of these types of social constructs from text as

$$U = g(X) \tag{1.1}$$

where $g$ is the measurement function that maps text, $X$, to the concept of interest, $U$. Egami et al. [2018] call this $g$-function the "codebook function" and describe how it can generically map text to any lower-dimensional representation. Using NLP methods, $g$ could take many forms including rule-based dictionary look-ups, supervised classifiers, unsupervised learners (e.g. topic models or word embeddings), or a combination of these methods.

**Challenges.** To preface the contributions of this thesis, we highlight several settings in which it is challenging to adapt existing NLP methods to *measurement* for text:

1. *Aggregate corpus-level measurement.* Many tasks in NLP—especially recent popular benchmarks for general-purpose "natural language understanding," e.g. Wang et al. [2018, 2019]—focus on settings for which $X$ is a single sentence. Other work in NLP addresses $X$ as a document (e.g. Iyyer et al. [2015], Yang et al.

[2016]). Yet, it is rare to see NLP tasks for which $X$ is an entire corpus. However, in the social sciences, corpus-level measurement is abundant: some aim to measure a corpus's proportion of categories—for example, the proportion of constituent mail in specific policy areas [Hopkins and King, 2010]—or corpus-level counts of events induced by particular actors—for example, counts of the different kinds of police intervention in ethnic conflict [Wilkinson, 2006]. Yet, simple aggregations of NLP predictions at the sentence or document level do not necessarily result in corpus-level accuracy.

2. *Distributional shifts.* Challenge #1 is further exacerbated if social scientists aim to characterize temporal or domain changes. For instance, $g$ (from Equation 1.1) could be a trained classifier that is used to infer the construct of interest for a collection of documents at each time step: $U_t = g(X_t)$ for $t = 1, 2, \ldots, n$. However, most ML and NLP models assume the data is *independent and identically distributed (i.i.d.)* in which the training and test sets are drawn from the same distribution. However, in most settings the data is not i.i.d., in which case $g$ can often biased towards the class prevalence at training time. While distribution shifts are a longstanding research area in ML and NLP [Hand, 2006, Blitzer et al., 2007, Daumé III, 2007] and recent efforts have gathered and characterized empirical examples of these distribution shifts [Koh et al., 2021], this is still a difficult and open problem in porting "off-the-shelf" methods from NLP to social-science measurement.

3. *Linguistic complexity, ambiguity, and diversity.* Compared to other mediums of data, $g$ is often difficult to construct because language has complex structure which sometimes leads to inherent ambiguity and often results in multiple constructions having the same semantic meaning [Bender, 2013, Bender and Lascarides, 2019]. Language is more than just a "bag-of-words." Semantics, the meaning of language, is built from syntax, the structure of language. For

instance, "Mary likes John" takes a different semantic meaning than "John likes Mary" even though both examples contain the same words. Language can also be ambiguous; in the example "She saw the man with the telescope" the prepositional phrase "with the telescope" could attach to either "saw" or "man," giving the sentence two distinct but plausible meanings. While these linguistic challenges are a fundamental focus of NLP research, they are particularly important in the gap between NLP methods and social scientific measurement because the cultural concepts central to social science studies often have a very complex linguistic structure—e.g. the measures of censorship topic areas, respect, and ideological slant we previously mentioned. Furthermore, social scientists may also want to quantify *uncertainty* resulting from any inherent ambiguity in language.

4. *Small annotation budgets.* For supervised learning settings, it is typically assumed that the more data one has, the better the accuracy of one's model [Halevy et al., 2009]. Many NLP benchmark datasets require enormous amounts of time and money to construct—for example, the creation of the Penn Treebank took eight years and thousands of annotation hours [Marcus et al., 1993]. Yet, for many social science applications, the annotations of interest are complex social constructs that often require domain experts to annotate, which could result in high costs. Thus, it is of particular importance to text-based social measurement to focus on regimes with small amounts of labeled data.

**Thesis contributions.** In Chapter 2, we address challenges #1 and #2 within the context of *prevalence estimation*—the task of inferring the relative frequency of classes of unlabeled examples in a group; for example, the proportion of a document collection with positive sentiment. We contribute (1) a generative probabilistic modeling approach to prevalence estimation and (2) the construction and evaluation of prevalence confidence intervals in order to reflect uncertainty over the predicted preva-

lence from imperfect classifiers. We show that an off-the-shelf discriminitive classifier can be given a generative re-interpretation by backing out an implicit individual-level likleihood function. Empirically, we demonstrate our approach provides better confidence interval coverage than alternatives, and is more robust to shifts in the class prior between training and inference (challenge #2).

In Chapter 3, we address challenges #3 and #4 within the context of *entity-event measurement*—measuring entities who are actors or recipients of certain events—and the specific application of extracting names of persons who have been killed by police from a corpus of news documents. We propose police fatalities are a useful test case for text measurement and event extraction research because fatalities are a well defined event type with clear semantics. Overall, we contribute a novel police fatality corpus and present a model to solve this application with no annotated data (challenge #4) by using EM-based distant supervision—inducing labels by aligning relation-entity entries from a gold standard database to their mentions in a corpus—with logistic regression and convolutional neural network classifiers. Our model outperforms two off-the-shelf event extractor systems, and it can suggest candidate victim names in some cases faster than one of the major manually-collected police fatality databases. We address linguistic ambiguity in difficult sentences (challenge #3) by using a method that samples from the full joint distribution of dependency parse trees to communicate ambiguity in language syntax and demonstrate this approach has improved empirical results for our police fatality pipeline.

## 1.2 Causal inference with text

**Definition.** Beyond measurement, social scientists are often interested in *causal* questions [Morgan and Winship, 2015, Grimmer, 2015]. In contrast to descriptive or predictive tasks, *causal inference* aims to understand how *intervening* on one variable affects another variable [Holland, 1986, Morgan and Winship, 2015, Pearl, 2009b].

Morgan and Winship [2015] describe various modes of causal inquiry in the social sciences: associational analysis between observed treatments and outcomes, targeted analysis of the effect of one or more focal causes, and finally all-cause structural analysis. Similarly, Pearl [2019] proposes a three-level causal hierarchy: association—purely statistical relationships defined by the naked data; intervention—researchers' manipulations of the data; and counterfactuals—retrospective reasoning.

In this thesis, we focus on estimating causal effects for questions that take the general form, "What is the effect of a *treatment* variable on an *outcome* variable?" For example,

1. What is the effect of alcohol use (treatment) on academic success (outcome) [Kiciman et al., 2018] (Chapter 4)?
2. What is the effect of lawyers' signalled gender (treatment) on whether U.S. Supreme Court justices interrupt them during oral arguments (outcome) (Chapter 5)?

Let $T$ be the treatment variable and $Y$ be the outcome variable. Formally, the causal questions presented above are inquiries of the (binary) treatment effect,

$$Pr(Y|do(T=1)) - Pr(Y|do(T=0)) \tag{1.2}$$

in which $do(X=x)$ represents a researcher's intervention that sets variable $X$ to the value $x$ [Pearl, 2009b], and $T \in \{0,1\}$ are specific values of treatment (e.g. male and female lawyers for Example 2).

However, researcher intervention on treatment variables is often infeasible or unethical in the social sciences. In Example 1 above, it would be unethical to assign participants to abuse alcohol due to potential health consequences. In Example 2, researchers cannot disrupt the proceedings of high-stakes U.S. Supreme Court oral arguments with controlled interventions. In these cases, researchers often turn to *observational* (non-experimental) data. In the observational setting, researchers will

often need to account for confounders ($C$)—variables that cause both $T$ and $Y$—and mediators ($M$)—variables on the causal path $T \rightarrow M \rightarrow Y$—in order to have unbiased estimates of the causal effects of interest.

In this thesis, we focus on the settings for which text is a proxy for confounders (Chapter 4) and mediators (Chapter 5) when estimating causal effects from observational data. In Example 1, researchers can use text from particpants' social media posts (e.g. Twitter messages) as a proxy for demographic variables to which they do not have access. In Example 2, researchers can use the language of U.S. Supreme Court lawyers as a mediating variable between gender signal and interruption.

We formally define a structural causal model (SCM) [Pearl, 2009b] for these settings of text as a proxy for confounders or mediators. Let $V$ be a set of endogenous variables and $F$ be a set of nonparametric functions that assigns each variable in $V$ a value based on the values of other variables in the model.[2] Then our SCM is

$$V = \{T, Y, C, M, X_1, X_2\} \tag{1.3}$$

$$F = \{f_C, f_M, f_Y\} \tag{1.4}$$

$$C = f_C(X_1) \tag{1.5}$$

$$M = f_M(X_2) \tag{1.6}$$

$$Y = f_Y(T, C, M) \tag{1.7}$$

where $T, Y, M$, and $C$ are the treatment, outcome, mediator, and confounder variables respectively; $X_1$ is the text that encodes the confounders; and $X_2$ is the text that encodes the mediators. This model is accompanied by the causal graph in Figure 1.1. Note, the nonparmetric functions for the confounder and mediator, $C = f_C(X_1)$ and $M = f_M(X_2)$, are equivalent to the text measurement functions in Equation 1.1.

[2]For simplicity, we exclude the set of exogenous variables here.

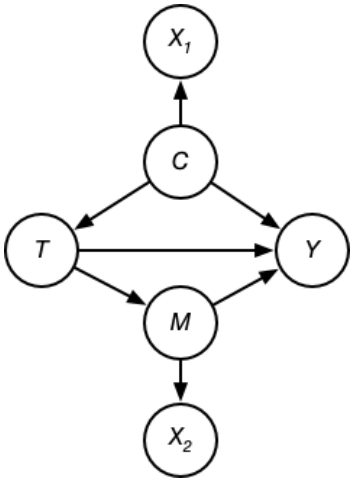



Figure 1.1: Causal diagram in which nodes are causal variables and arrows represent causal dependencies. Here, we include the causal variables for treatment ($T$), outcome ($Y$), confounder ($C$), mediator ($M$), and text ($X_1$ and $X_2$).

**Challenges.** Unlike text measurement—which has been explored extensively in the past few decades—causal inference with text is still a relatively new research area. As such, this thesis highlights several epistemological challenges with text-based casual inference:

1. *Methods and applications are scattered across different communities.* Causal inference methods have been reinvented and iterated on within the fields of statistics [Holland, 1986], epidemiology [Hernán and Robins, 2020], economics [Angrist and Pischke, 2008], computer science [Pearl, 2009b] and the broader social sciences [Morgan and Winship, 2015]. Yet, researchers have not reached consensus on causal formalisms, terminology, methods, and tasks. Furthermore, in the emerging subfield of text-based causal inference, applications are scattered across many different academic disciplines and publication venues, making it difficult to see gaps between desired applications and existing methods.
2. *Text-specific causal identification assumptions.* One of the major difficulties of causal inference is that estimation is contingent on often untestable *causal identification assumptions.* For instance, a researcher often must assume *un-*

*confoundedness*—that all latent confounders are accounted for—and *overlap* (also known as *positivity*)—that any unit has a non-zero probability of assignment to each treatment condition for all possible values of the confounder set [Morgan and Winship, 2015]. A major challenge of text-based causal inference is determining when these assumptions hold for causal estimates that include high-dimensional text data and when additional assumptions must be made.

3. *Lack of ground-truth for causal evaluation.* Unlike prediction, in which we can evaluate methods via predictive performance (e.g. accuracy or mean-squared error) on a held-out test set, causal evaluation is difficult because the true causal effects for real-world problems are typically unknown. Thus, incorporating text-specific causal assumptions (challenge #2) into a causal system that is already difficult to evaluate presents an even greater challenge.
4. *Causal estimates with multiple text measurements or measurement error.* There are often multiple, valid options of how to measure text and these options will often have varying levels of accuracy. Incorporating these noisy measurements into causal inference is a potential problem, and characterizing the extent of the problem is even more difficult without ground-truth causal evaluations (challenge #3). These issues are further complicated when text simultaneously encodes multiple causal variables (e.g. confounders and colliders) and one must separate measures of these variables.
5. *Relatively few causal designs specific to text-as-data.* Many text-as-data social science applications are asking causal questions, but either the causal question or causal assumptions are undeclared. Although this is slowly changing, we posit that developing more causal designs explicitly focused on text-as-data could help expand the subfield and number of potential causal applications with text in the social sciences.

**Thesis contributions.** Because text-based causal inference is a newly emerging subfield, Chapters 4 and 5 of this thesis focus on establishing the conceptual foundations (as opposed to empirical methods and results) for text-based causal inference while addressing the challenges presented above.

In Chapter 4, we focus on the specific setting in which text data encodes latent confounders and one wants to use this text to reduce confounding from causal estimates with observational data. Since methods and applications are scattered across different communities (challenge #1), we systematically gather and categorize examples of text as a proxy for confounders and contribute a guide to data processing decisions. We discuss text-specific causal assumptions (challenge #2), potential sensitivity of causal estimates to different representations and choices of imperfect measurements of text (challenge #4), and potential avenues forward for causal evaluation with text (challenge #3).

In Chapter 5, we focus explicitly on challenge #5 and contribute a new causal research design for observational (non-experimental) data to estimate the natural indirect and direct effects of social group signals (e.g. race or gender) on conversational outcomes with separate aspects of language as causal mediators. We illustrate the promises and challenges of this framework via a theoretical case study of the effect of an advocate's gender on interruptions from justices in U.S. Supreme Court oral arguments. We also discuss challenges conceptualizing and operationalizing causal variables such as gender and language that comprise of many components, and highlight issues when there are multiple potential operationalizations of causal variables using NLP methods (challenge #4). We also articulate potential open challenges in this research design including temporal dependence between mediators in conversations, causal dependence between multiple language mediators, and dependence between social group perception and language perception.

## 1.3 Thesis statement

This thesis contributes conceptual and empirical advances in quantitative analysis of text for the social sciences by blending: (a) the computational strengths of natural language processing (NLP) and machine learning to automate and scale-up text analysis with (b) two themes central to social scientific studies but often underaddressed in NLP: measurement—creating quantifiable summaries of empirical phenomena—and causal inference—estimating the effects of interventions.

In Chapter 6, we conclude with reflections on the relationship between text-based measurement and causal inference and future research directions along these two themes and at their intersection.

# CHAPTER 2

# MEASURING CLASS PREVALENCE IN DOCUMENTS

This chapter was originally published as Keith and O'Connor [2018].

## 2.1 Introduction

The goal of *prevalence estimation* is to infer the relative frequency of classes $y_i$ associated with unlabeled examples (e.g. documents) from a group, $x_i \in \mathcal{D}$. For example, one might want to estimate the proportion of blogs with a positive sentiment towards a political candidate [Hopkins and King, 2010], sentiment of responses to natural disasters on social media [Mandel et al., 2012], or prevalence of car types in street photos to infer neighborhood demographics [Gebru et al., 2017]. Often, an analyst wants to compare prevalence between multiple groups, such as inferring prevalence variation over time (e.g., changes to online abuse content [Bissias et al., 2016]), or across other covariates (e.g., changes in police officers' "respect" when speaking to minorities [Voigt et al., 2017]). This problem has been re-introduced in many different fields: as "quantification" in data mining [Forman, 2005, 2008], "prevalence estimation" in statistics and epidemiology [Gart and Buck, 1966], and "class prior estimation" in machine learning [Vucetic and Obradovic, 2001, Saerens et al., 2002]. In NLP, SemEval 2016 and 2017 included Twitter sentiment class prevalence tasks [Nakov et al., 2016, Rosenthal et al., 2017].

Prevalence estimation assumes access to a (potentially small) set of labeled examples to train a classifier; but unlike the task of individual classification, the goal is to estimate the proportion of a class among examples in a group. If a perfectly ac-

curate classifier is available, it is trivial to construct a perfect prevalence estimate by counting the classification decisions (§2.3.1). In fact, most application papers in the previous paragraph use this or a similar aggregation rule to conduct their prevalence estimates. However, classifiers often exhibit errors from different sources, including:

- Shifts in the class distribution from training to testing ($P_{train}(y) \neq P_{test}(y)$). A classifier may be biased toward predicting $P_{train}(y)$.
- Difficult classification tasks (such as predicting sentiment or sarcasm) that result in low accuracy classifiers; this can be exacerbated by limited training data, as is common in social science or industry settings that require manual human annotation for labels.

It is typically assumed (and sometimes confirmed) that when an individual classifier has less than 100% accuracy, it can still give reasonable prevalence estimates.[1] However, there is relatively little understanding to what extent the quality of the document-level model impacts prevalence estimates. Imperfect classifier accuracy ought to be reflected in uncertainty over the predicted prevalence.

In this work, we tackle both of these challenges simultaneously, using a generative probabilistic modeling approach to prevalence estimation. This model directly parameterizes and conducts inference for the unknown prevalence, naturally accommodating shifts between training and testing, and also allows us to infer confidence intervals for the prevalence. We show that our best model can be seen as an *implicit likelihood* generative re-interpretation of an off-the-shelf discriminative classifier (§2.4.2); this unifies it with previous work, and also is easy for a practitioner to apply.

We additionally review several types of class prevalence estimators from the literature (§2.3), and conduct a robust empirical evaluation on sentiment analysis over

[1] For example, Bissias et al. find a relative mean absolute error of less than 0.01 when the individual classifier has ROC AUC of 0.91.

hundreds of document groups, illustrating the methods' biases and robustness to class prior shift between training and testing. Our method provides better confidence interval coverage and is more robust to class prior shift than previous methods, and is substantially more accurate than an algorithm in widespread use in political science.

## 2.2 Problem definition

We consider two prevalence estimation problems: (1) point prediction and (2) confidence interval prediction. In this work, we are most interested in supervised learning for discrete-valued document labels, with access to a small to moderate number (e.g. around 1000) of labeled documents with text $x$ and label $y$: $(x_i, y_i) \in \mathcal{D}^{train}$. We restrict attention to binary-valued labels $y \in \{0, 1\}$. At test time, there are one or more groups of unlabeled test documents, $\mathcal{D}^{(1)}, \cdots, \mathcal{D}^{(G)}$; for example, one group might be a set of tweets sent during a certain month, or a set of online reviews associated with a particular product. For each group $\mathcal{D}$, let $\theta^* \equiv (1/n) \sum_i^n y_i$ be the true proportion of positive labels (where $n = |\mathcal{D}|$).

The *prevalence point prediction* problem is to take an unlabeled document group $\mathcal{D}$ as input and infer an estimated $\hat{\theta} \in [0, 1]$. Ideally, this point estimate should be close to the true prevalence $\theta^*$; we evaluate this by mean absolute error.

In this work, we are the first (that we know of) to introduce the question of *uncertainty* in prevalence estimation. Since document classifiers are typically far from perfectly accurate, we should expect substantial error in prevalence prediction, and inference methods should quantify such uncertainty. We formalize this as a *prevalence confidence interval* (CI) inference, which takes as input a desired nominal coverage level $(1 - \alpha)$, and predicts a real-valued interval $[\hat{\theta}_{lo}, \hat{\theta}_{hi}] \subseteq [0, 1]$. Ideally, a CI prediction algorithm should have frequentist coverage semantics: over a large

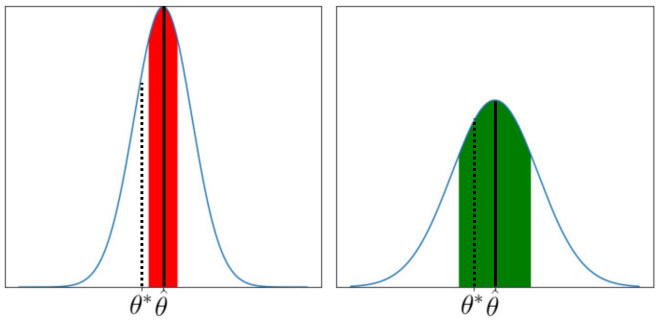



Figure 2.1: Example posterior distributions with MAP prevalence estimates, $\hat{\theta}$ (solid line) and the true prevalence, $\theta^*$ (dashed line). A desirable property is that confidence intervals, technically Bayesian credible intervals, (shaded regions) will be wider for more uncertain models. For example, the wider CI on the right (green) contains $\theta^*$ whereas the narrower CI interval on the left (red) does not.

number of test groups,[2] $(1-\alpha)\%$ of the predicted intervals ought to contain the true value $\theta^*$. If the problem is hard—for example, the relationship between document features and the label is not captured well by the model—the CI should be wide. We empirically evaluate coverage of CI-aware prevalence inference models. See Fig. 2.1 for an intuitive example.

## 2.3 Review and baselines: Discriminative individual classification aggregation

The most straightforward baseline approach to prevalence estimation is to build on discriminative, supervised learning for individual-level labels, such as binary logistic regression with bag-of-words features, randomized feature hashing [Weinberger et al., 2009], or neural networks [Goldberg, 2016]. Such a model defines an individual document's label probability $p_i \equiv p_\beta(y_i = 1 \mid x_i)$ where parameters $\beta$ are fit by maximizing regularized likelihood on the labeled training data.

[2]Or in fact, across many experiments in which the model or algorithm is applied [Wasserman, 2011].

### 2.3.1 Classify and count (CC)

For prevalence point estimation, Forman [2005] defines the "classify and count" (CC) method as simply averaging the most-likely individual label predictions,

$$\hat{\theta}^{CC} = \frac{1}{n} \sum_i 1\{p_i > 0.5\}. \tag{2.1}$$

This is the most obvious approach for practitioners, but it has at least two weaknesses, which have been addressed in different groups of prior work. First, the class proportions may change between training and test groups, which the Adjusted CC and ReadMe algorithms attempt to fix (§2.3.2–2.3.3). Second, it discards probabilistic information, which is remedied by the Probabilistic CC method, and an extension we propose (§2.3.4–2.3.5).

### 2.3.2 Adjusted classify and count (ACC)

CC may encounter problems if the test class distribution is different than the training's. The "adjusted classify-and-count" method [Gart and Buck, 1966, Forman, 2005] treats the classifier output as a proxy variable, and estimates a separate confusion model of classifier output $\hat{y}_i \equiv 1\{p_i > 0.5\}$ conditional on the true label, $p(\hat{y} \mid y)$, from cross-validation within the training set. Assuming the confusion model extends to the test data, a moment-matching approach is then used to infer the true label proportions, by first observing $p_{test}(\hat{y}) = \sum_y p(\hat{y} \mid y) p_{test}(y)$ and solving the linear system for $p_{test}(y)$, the test-time expected class prevalence. Using empirical estimates for the true positive rate $\text{TPR} = p(\hat{y} = 1 \mid y = 1)$, and false positive rate $\text{FPR} = p(\hat{y} = 1 \mid y = 0)$, and $\hat{\theta}^{CC} = p(\hat{y} = 1)$, it has the closed form

$$\hat{\theta}^{ACC} = \frac{\hat{\theta}^{CC} - \text{FPR}}{\text{TPR} - \text{FPR}}. \tag{2.2}$$

By design, ACC is more robust to a new test-time prevalence, but it relies on the accuracy of its TPR and FPR estimates, and its lack of probabilistic semantics makes it unclear how to infer confidence intervals.

### 2.3.3 ReadMe algorithm

An interesting extension to ACC is to remove the need for a discriminative classifier, by directly modeling text conditional on the latent document class. The ReadMe algorithm, developed in political science [Hopkins and King, 2010], extends ACC's linear system for every term type in a (subsampled and augmented) term vocabulary $\mathcal{V}$, and calculates their class-conditional probabilities from the training data. Assuming these conditional models also hold in the test data, that implies $p_{test}(w) = \sum_y \hat{p}(w \mid y) p_{test}(y)$; the algorithm infers $p_{test}(y)$ by minimizing the squared error of predicted versus empirical term frequencies in the test set. The open-source ReadMe software package[3] has been used in numerous political science studies, including inferring proportions of types of censored Chinese news [King et al., 2013], credit claiming in Congressional press releases [Grimmer et al., 2012], and voter intentions among Twitter messages [Ceron et al., 2015].

ReadMe is theoretically appealing in that it infers latent class prevalences to explain the test group's textual evidence; but as a non-probabilistic model, it does not directly imply a method for confidence intervals (Hopkins and King use the bootstrap). Furthermore, our experiments (§2.5), contra the original paper, show its implementation exhibits poor performance.

### 2.3.4 Probabilistic classify and count (PCC)

Both the CC and ACC methods discard uncertainty information from the classification model. In a difficult classification setting, for example, we might expect many

[3] https://gking.harvard.edu/readme

probabilities to be near, say, 0.6, in which case the CC method may undercount the negative class. This suggests an alternative method, "probabilistic classify and count" (PCC):

$$\hat{\theta}^{PCC} = \frac{1}{n} \sum_i p_i \tag{2.3}$$

which is the expected prevalence, $(1/n) \sum_i y_i$, assuming each $y_i$ is distributed according to the original probabilistic classifier.

### 2.3.5 PCC Poisson-Binomial distribution (PB-PCC)

If we assume each $y_i$ is conditionally independent given text $x_i$ and model parameters $\beta$, this defines a fully probabilistic model for the class prevalence. Let the latent variable $S = \sum_i y_i$; its distribution is thus Poisson-Binomial [Chen and Liu, 1997]. The modeled prevalence distribution $p(\frac{S}{n} \mid \mathcal{D})$ can be exactly inferred by Monte Carlo inference: each iteration samples every $y_i$ and sums for an $S$ sample. The $S/n$ distribution over many iterations can be used to construct a Monte Carlo CDF $\hat{F}$, from which any $[\hat{F}(t), \hat{F}(t+1-\alpha)]$ is an $(1-\alpha)$-sized credible interval (where $0 \leq t \leq t + 1 - \alpha \leq 1$). This model has prevalence expectation $E[\frac{S}{n}] = \hat{\theta}^{PCC}$, and variance

$$\text{Var}\left[\frac{S}{n}\right] = \frac{1}{n^2} \sum_i p_i(1 - p_i). \tag{2.4}$$

To a certain degree, this model captures uncertainty in the classifier since per-document variance, $p_i(1 - p_i)$, is high when $p_i = 0.5$ and low when near 0 or 1. However, it also has a major weakness—the variance concentrates with a large test group size $n$, which is the wrong behavior when a classifier is truly noisy, for example, when a classifier is genuinely uncertain and predicts the same constant $p_i = q$ for each document. In this case, the correct behavior would be to maintain a flat, wide

posterior belief about $\theta$, which is better accomplished by the generative model we introduce in the subsequent section.

## 2.4 Our approach: generative probabilistic modeling

We turn to generative modeling, that seeks to to jointly model the probability of labels and text in both the training and test groups, by assuming a document's text is generated conditional on the document label. Language models have widespread use in natural language processing, and class-conditional models have been used for document classification (e.g. multinomial Naive Bayes; [McCallum and Nigam, 1998]). We use a similar generative setup to explicitly model a class prevalence for test group $g$, with a generative story for each (bag-of-words) document $i$ in the group:

$$\theta_g \sim \text{Dist}(\alpha) \tag{2.5}$$

$$y_{i,g} \sim \text{Bernoulli}(\theta_g) \tag{2.6}$$

$$x_{i,g} \sim \text{Multinomial}(\phi_{y_{i,g}}) \tag{2.7}$$

The test group is assumed to have a latent class prior $\theta_g$, which itself has a prior distribution (we assume $\text{Dist}(\alpha) = \text{Unif}(0, 1)$ in this work). For each class $k$, $\phi_k$ is a class-conditional unigram language model, which is learned from the training data but fixed at test time. We then perform inference to find $\theta_g$ that gives a high probability to text data $\{x_i \in \mathcal{D}^{(g)}\}$. Figure 2.2 shows the probabilistic graphical model.

### 2.4.1 MNB and loglinear language models

We experiment with two explicit language models in this generative framework: (1) multinomial Naive Bayes (**MNB**), using a training-time symmetric Dirichlet prior $\phi_y \sim \text{Dir}(\lambda/V)$ for vocabulary size $V$ and "pseudocount" $\lambda$, and (2) an additive log

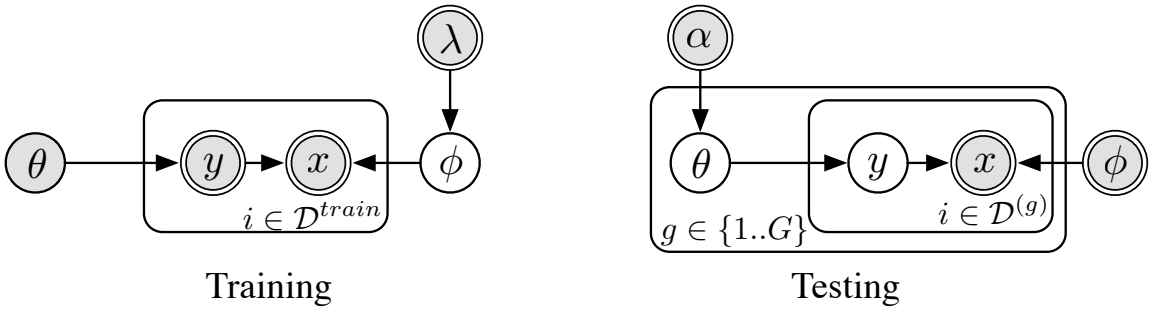



Figure 2.2: Our generative model for prevalence estimation. **Left:** Class-conditional language models ($\phi$) are learned at training time. **Right:** Test-time inference for multiple groups' latent prevalences ($\theta$).

linear model (**Loglin**, a.k.a. SAGE [Eisenstein et al., 2011]). Loglin estimates words' probabilities as deviations from a background log-probability $m$,

$$\eta_{y,w} \sim \text{Laplace}(\lambda) \tag{2.8}$$
$$\phi_{y,w} = \exp(m_w + \eta_{y,w}) / \sum_j \exp(m_j + \eta_{y,j})$$

where $m_w$ is the empirical log probability of a word $w$ among all training documents, and $\eta_{y,w}$ denotes class-specific deviations of the log-probability of a word $w$, MAP estimated under a sparsity-inducing L1 penalty. Such sparse additive models have been used in both supervised and unsupervised document modeling; for example, as a document-level posterior classifier it outperforms MNB [Eisenstein et al., 2011], or even discriminative models [Taddy, 2013], and its sparsity helps interpretability for analyzing political, literary, and legal texts [Monroe et al., 2008, Sim et al., 2013, Bamman et al., 2014, Wang et al., 2012].

### 2.4.2 Implicit likelihoods from discriminative classifiers (LR-Implicit)

This generative formulation has a major advantage over the discriminative, CC-style aggregation models because it sets up a likelihood and posterior distribution over $\theta$. But in terms of document modeling for classification purposes, the independence

assumptions of the generative model are typically too strong, and for document-level classification, discriminative models tend to outperform similarly parameterized generative ones, especially when the training set is sufficiently large [Ng and Jordan, 2002]. Thus, discriminative models may have information better suited to class prevalence inference. Also, since the most common practice for document classification is to use discriminative models, it would be helpful to more effectively use discriminative posteriors within our generative context.

In Naive Bayes-style generative document classification, the model defines $p_{gen}(x \mid y)$ and class prior $p(y)$, which are combined to calculate the posterior $p_{gen}(y \mid x) \propto p_{gen}(x \mid y)p(y)$. Discriminative models, by contrast, directly define a $p_{disc}(y \mid x)$. We can, however, expand this quantity via Bayes Rule:

$$p_{\text{disc}}(y \mid x) = p_{\text{implicit}'}(x \mid y)p_{\text{train}}(y)/p(x). \tag{2.9}$$

The "implicit document likelihood" $p_{\text{implicit}'}(x \mid y)$ is a likelihood function that, combined with a particular class prior $p(y)$, would have resulted in the same posterior predicted by the discriminative model. Given the discriminative posterior predictions and the training-time class prior $p_{\text{train}}(y) = \hat{\theta}_{\text{train}}$, an implicit likelihood function can be backed out for any particular document $x$; we define the "simple implicit" likelihood for document $x$ to be:

$$p_{\text{implicit}}(x \mid y) = p_{\text{disc}}(y \mid x)/\hat{\theta}_{\text{train}}. \tag{2.10}$$

This takes the form of a correction of the discriminative posterior, by dividing out the training-time class prevalence.[4]

[4]Technically, $p_{\text{implicit}'}$ is retrievable only up to a constant, and $p_{\text{implicit}}$ is one particular compatible implicit likelihood, since it can be multiplied by any constant and is still consistent with Eq. 2.9, and would give rise to the same document- and group-level posteriors.

Our **LR-Implicit** generative model uses the same class prevalence and document label generation setup as before, but to calculate the individual documents' $p(x \mid y)$ probabilities, it uses $p_{\text{implicit}}$ based on a logistic regression $p_{\text{disc}}$.[5]

This model is inspired by [Saerens et al., 2002]'s EM algorithm for adjusting a classifier for a test set's class prior; they derive it differently by applying the assumption $p_{\text{train}}(x \mid y) = p_{\text{test}}(x \mid y)$, expanding each side with Bayes' Rule, solving for $p_{\text{test}}(y \mid x)$, then estimating $p_{\text{test}}(y)$ via EM. This in fact optimizes the same marginal likelihood function in the next section under the implicit-discriminative generative model; our formulation broadens it as a fully Bayesian or likelihood-based model.

### 2.4.3 Inference

To estimate class prevalence, we use the marginal log likelihood over $\theta$ to obtain a posterior over $\theta$. For each each test group $g$, we have the marginal log probability of all document texts,

$$\begin{aligned} \text{MLL}_g(\theta) &\equiv \log p(\mathcal{D}^{(g)} \mid \theta) \\ &= \sum_{i \in \mathcal{D}^{(g)}} \log \sum_{y \in \{0,1\}} p(x_i, y_i = y \mid \theta) \\ &= \sum_{i \in \mathcal{D}^{(g)}} \log \Big( \theta L_i^+ + (1-\theta) L_i^- \Big), \end{aligned} \tag{2.11}$$

where we denote the class-conditional document text likelihoods $L_i^+ \equiv p(x_i \mid y_i = 1)$ and $L_i^- \equiv p(x_i \mid y_i = 0)$. The gradient for an individual document is $(L_i^+ - L_i^-)/(\theta L_i^+ + (1-\theta) L_i^-)$; intuitively, the sign of the numerator says that documents that are more likely under the positive than negative class encourage higher likelihood for larger values of $\theta$. When the model is uncertain about a document—that is, when $L_i^+ \approx L_i^-$—that document contributes a relatively flat likelihood curve, expressing

[5]The implicit likelihood still has the form of a logistic regression, adjusting its bias term: if $p_{\text{disc}}(y \mid x) = \sigma(\beta' x + \beta_0)$, then $p_{\text{implicit}}(x \mid y) = \sigma(\beta' x + \beta_0 - \log(\theta_{\text{train}}/(1-\theta_{\text{train}})))$.

little preference for likely values of $\theta$. If a model is more heavily regularized—for example, when the log-linear additive model is more dominated by the background language model—this condition tends to hold for the documents, leading to a flat, highly uncertain likelihood curve.

The marginal log likelihood is unimodal over $\theta \in [0, 1]$, since it is concave, being a sum of concave log-linear functions, and having negative curvature:

$$\frac{\partial^2 \text{MLL}_g}{\partial \theta^2} = -\sum_{i \in \mathcal{D}^{(g)}} \left( \frac{L_i^+ - L_i^-}{\theta L_i^+ + (1-\theta) L_i^-} \right)^2. \tag{2.12}$$

Since it is concave and there is only one parameter, a very wide variety of techniques could be used to reliably find a mode, including EM or first- or second-order methods. At least two approaches to inferring confidence intervals are possible. One is to use a central limit theorem-style approximation, assuming the sampling distribution is approximated by a normal with mean $\theta^{\text{MLE}}$ and variance $-[\partial^2 \text{MLL}_g / \partial \theta^2]^{-1}$. The second, which we focus on, is Bayesian estimation for $\log p(\theta_g \mid \mathcal{D}^{(g)}) \propto \log p(\theta_g) + \text{MLL}_g(\theta_g)$ by simply using a grid search over values $\theta \in \{0.001, 0.002, ...0.999\}$ to infer both the posterior mode $\theta^{\text{MAP}}$ as well as a 90% highest posterior density interval.[6] In small-scale experiments, this model had very similar results to the central limit theorem (with EM for $\theta^{\text{MLE}}$).

## 2.5 Experiments

### 2.5.1 Data

To compare document class prevalence estimators, we desire datasets that (1) have natural document groups that correspond to realistic, real-world applications, (2) have a large number of test groups (hundreds or more), and (3) are freely available

[6]Since we use a uniform prior, this is just the MLE. Technically, we used a prior of Beta(1.0001, 1.0001) to avoid certain issues with tie-breaking, but it was not necessary.

for academic research. It has been a challenge to fulfill these criteria in previous work. Nakov et al. [2016] conduct large-scale manual annotation of Twitter sentiment for SemEval 2016 Task 4, with topic-based test groups; unfortunately, redistribution is restricted to message IDs, making the original dataset difficult to reconstruct under Twitter's terms of service if messages have since been deleted. Bella et al. [2010] and Esuli and Sebastiani [2015] use large, pre-existing labeled document corpora, but they do not contain natural groups; evaluations utilize randomly sampled synthetic groups.

To better fulfill these criteria, we select the task of business review sentiment prevalence, where the goal is to estimate the proportion of reviews that are positive for one particular business; specifically, we use labeled data from the Yelp Dataset Challenge Round Nine[7] corpus, which consists of 4.1M reviews by 1M users for 144K businesses. We sample 500 businesses with at least 200 reviews each as the test groups. We treat the task as binary classification, and assign $y_i = 1$ to reviews with 3 or more stars. This task seems reasonably representative of real-world sentiment analysis problems, and this type of dataset can easily be collected and reproduced from Yelp or other widely available review data.

For training, we simulate a small-scale annotation project by sampling 2000 labeled documents from the rest of the corpus. This is a **natural** prevalence that on average is about the same as the test groups, though individual test groups may have a much different prevalence (ranging from 0.096 to 0.997, mean (stdev) 0.823 (0.136)). We also construct a **synthetic** training setting with a highly skewed class prior, selecting 2000 documents with a 0.1 class prevalence (i.e. 200 positive documents in the group). In each case, for every model, we re-run and average results over

[7]Downloaded June 2017 from https://www.yelp.com/dataset_challenge.

| | | Natural training prevalence ≈ 0.8 | | | | Synthetic training prevalence = 0.1 | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Point est. | | CIs | | Point est. | | CIs | |
| | | MAE | Bias | Cover. | Width | MAE | Bias | Cover. | Width |
| Const. | Pred. train mean | 0.114 | -0.045 | — | — | 0.723 | -0.723 | — | — |
| | Pred. 100% | 0.177 | 0.177 | — | — | 0.177 | 0.177 | — | — |
| | ReadMe | 0.233 | -0.222 | — | — | 0.383 | -0.382 | — | — |
| Disc. (LR) | CC | 0.048 | 0.042 | — | — | 0.503 | -0.503 | — | — |
| | ACC | 0.048 | -0.001 | — | — | 0.132 | -0.015 | — | — |
| | PB-PCC | 0.049 | -0.017 | 0.283 | 0.044 | 0.464 | -0.464 | 0.001 | 0.054 |
| Gen. (MLL) | MNB | 0.078 | 0.058 | 0.120 | 0.046 | 0.199 | -0.199 | 0.022 | 0.073 |
| | Loglin | 0.089 | -0.070 | 0.410 | 0.100 | 0.140 | -0.036 | 0.510 | 0.273 |
| | LR-Implicit | 0.050 | 0.001 | 0.454 | 0.074 | 0.069 | -0.051 | 0.439 | 0.082 |

Table 2.1: Mean absolute error (MAE), bias, nominal 90% confidence interval coverage, and average CI width for the 500 Yelp data test groups, averaged over 10 simulations of resampled training (2000 document) sets. We examine both the natural positive class training prevalence ($E[\theta_{train}] = 0.7783$), and a synthetic fixed prevalence of 0.1. Dashes indicate the methods that are not able to calculate confidence intervals.

10 different samples of the training set. For preprocessing, we tokenize with NLTK[8] and lowercase.

### 2.5.2 Model training

We use L1 regularization for logistic regression based on the vector of a documents' word counts, to be most directly comparable to the generative models; for each model, we select its hyperparameter (LR and Loglin's $\lambda$, or MNB's pseudocount) by minimizing cross-validated cross-entropy of individual document posteriors (within the labeled training set), over a grid search of powers of 2. The log-linear additive model is trained with OWL-QN [Andrew and Gao, 2007][9] and the logistic regression model

[8] http://www.nltk.org/

[9] Via github.com/larsmans/pylbfgs

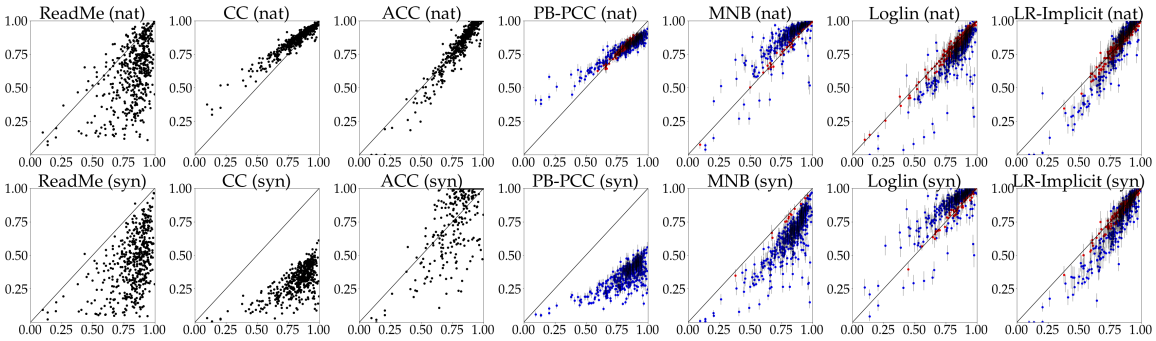



Figure 2.3: Gold prevalence $\theta^*$ (x-axis) versus predicted prevalence $\hat{\theta}$ (y-axis) for each of the 500 test groups with **natural** (nat) training prevalence (top row) and **synthetic** (syn) 0.1 training prevalence (bottom row). A black $y = x$ line is plotted for visualization. For the models that allow for confidence intervals, 90% CIs for each group are given by the faint grey lines. Blue dots indicate the CI does not contain $\theta^*$ and red dots indicate the CI does contain $\theta^*$. For each setting, we show the the model with median MAE across training resamplings.

is trained with the default implementation in scikit-learn [Pedregosa et al., 2011].[10] We used ReadMe with its default parameters.[11]

### 2.5.3 Results

For each of the 500 test groups, we calculate a prevalence point estimate $\hat{\theta}$ with each method, and evaluate by averaging across groups for mean absolute error $\sum_g |\hat{\theta}_g - \theta^*_g|$ and bias $\sum_g (\hat{\theta}_g - \theta^*_g)$.[12] For the models that allow for confidence interval prediction, we infer 90% intervals and calculate coverage, which is best if it is 0.90. We also report average CI width; a narrower interval indicates more confi-

[10] Version 0.18.2

[11] Version 0.99837 from https://gking.harvard.edu/readme, with default parameters features=15, n.subset=300, prob.wt=1. We bypass the ReadMe software's text preprocessing pipeline, and instead have it use nearly the same document-term matrices as the other models. Since it only handles binary document-term matrices, we transformed counts to indicators; with other models this change only made a minor difference in results.

[12] For the generative (MLL) models, $\hat{\theta}$ is the MAP estimate; the posterior mean gives similar results.

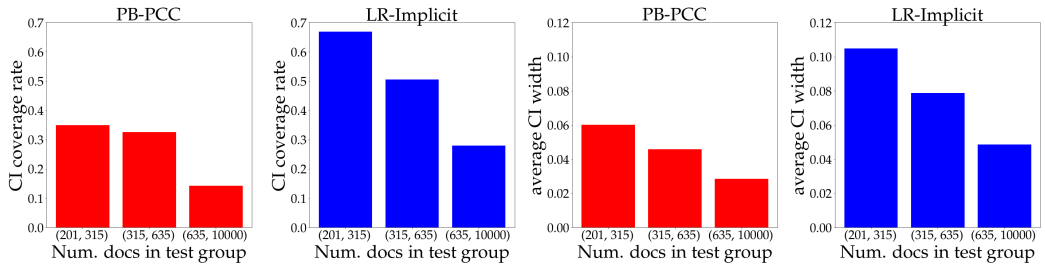



Figure 2.4: CI coverage rate (left two graphs) and average CI width (right two graphs) for three bins of the test groups, binned by number of documents.

dence (even if misplaced). Results are in Table 3.7; every result is averaged over 10 resamplings of the training set.

The ReadMe software did not have competitive performance; we hope in follow-up work to understand why Hopkins and King found it had considerably stronger performance than SVM-based CC.

For the natural training class prevalence setting (first column, Table 3.7), the discriminative-based models (CC, PCC and the adjusted variants ACC and LR-Implicit) all have very similar point estimate performance, outperforming the purely generative models (MNB and Loglin). For CI coverage, the log-linear and LR-Implicit generative models have significantly better coverage than the discriminative model (PB-PCC) or MNB. Future work is required to improve coverage to be closer to the nominal ideal of 90%.

By contrast, when the class prevalences are mismatched (second column, Table 3.7), the non-adjusted CC and PCC methods give extremely poor and biased point estimates, and PB-PCC has incredibly poor CI coverage. ACC and the generative models do much better, presumably because their models directly allow for variability in the test class prior. While Loglin has somewhat higher coverage in this setting, overall, LR-Implicit has consistently strong performance in both training settings, and for both point estimation and (relatively, at least) confidence intervals.

Figure 2.3 shows $\theta^*$ versus $\hat{\theta}$ for each of the 500 test groups for each of the models, including predicted CIs. CC's and PCC's erroneous assumptions are directly viewable: in the natural prevalence setting, the slope shallower than 1, indicating a persistent under-sensitivity to the true class prevalence—unlike ACC and the generative models. In the synthetic training case, CC and PCC wildly underpredict, presumably because they are biased by the low training-time prevalence $\theta_{\text{train}} = 0.1$.

### 2.5.4 Comparison of PB-PCC and LR-Implicit

Since PB-PCC and LR-Implicit represent the strongest members of non-adjusted classification aggregation and generative modeling, respectively, we further compare their results. When varying synthetic training prevalence across 0.1 to 0.9 (Figure 2.5a), LR-Implicit has much better MAE in all settings except near the natural prevalence (the test groups have, on average, 0.82 positive prevalence), and consistently stronger CI coverage.

Figure 2.5b shows results for natural class prevalence when varying the training set size. Unfortunately, LR-Implicit is disadvantaged at very small test sizes—its MAE is higher when there are only a few hundred training documents ($\leq 2^8 = 256$), though performance converges after that. We suspect this may occur because, when textual evidence is weak, the classifier learns to more heavily rely on its bias term, which can be a useful form of bias when the training class prevalence matches the test groups (on average). However, at all levels, LR-Implicit's coverage is better.

Since we hypothesized that PB-PCC may be overconfident for large test groups (§2.3.5), we test this by binning test groups by the number of documents per group. Figure 2.4 confirms that PB-PCC exhibits overconfidence for larger groups (smaller CI width alongside lower CI coverage), but LR-Implicit suffers from the same problem as well.

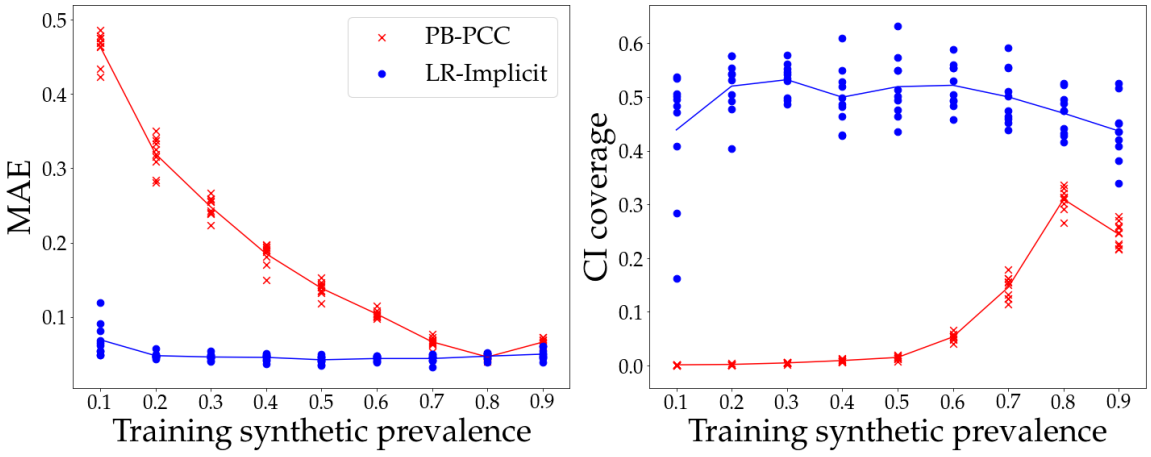
## 2.6 Additional related work

González et al. [2017a] reviews the class prevalence estimation literature, and we note a few threads of work here. Bella et al. [2010] propose a probabilistic variant of ACC, and Esuli and Sebastiani [2015] compare many methods on news article topics (RCV1) and medical record subject heading (OHSUMED-S) class prevalence tasks, finding varying results among CC, ACC, and PCC. A number of other empirical evaluations were conducted in two SemEval Twitter sentiment prevalence shared tasks, with varying results among these and other methods with a range of classifiers [Nakov et al., 2016, Rosenthal et al., 2017]; Nakov et al. note that CC was often one of the strongest methods. Esuli and Sebastiani as well as Xue and Weiss [2009] present semi-supervised loss-augmented classifier training methods to improve prevalence estimation. Tasche [2017] presents theoretical results for ACC and Saerens et al.'s EM method (what we call the LR-Implicit MLE), arguing they correctly predict $\theta^*$ under class prior shift; we confirm that those two methods are indeed better than many alternatives in our empirical evaluation. While we focus on inference of the test-time class prior as a class prevalence estimate, Saerens et al. [2002] also show their method can improve individual-level classification accuracy, which Sulc and Matas [2019] use for image classification. (From the viewpoint of individual classification, this phenomenon is known as prior probability shift [Moreno-Torres et al., 2012].) González et al. [2017b] and Card and Smith [2018], similarly to our results, find that CC is much poorer than ACC under class shift. Card and Smith also show that PCC can be sensitive to properties of the classifier, finding that well-calibrated classifiers can give strong performance. They argue that discriminative aggregation models are appropriate for tasks where humans respond to text. Jerzak et al. [2019] analyze issues in class prevalence estimation and propose the ReadMe2 algorithm, which adds external word embeddings, optimization-based dimension reduction, and similarity matching to ReadMe's moment-matching framework.

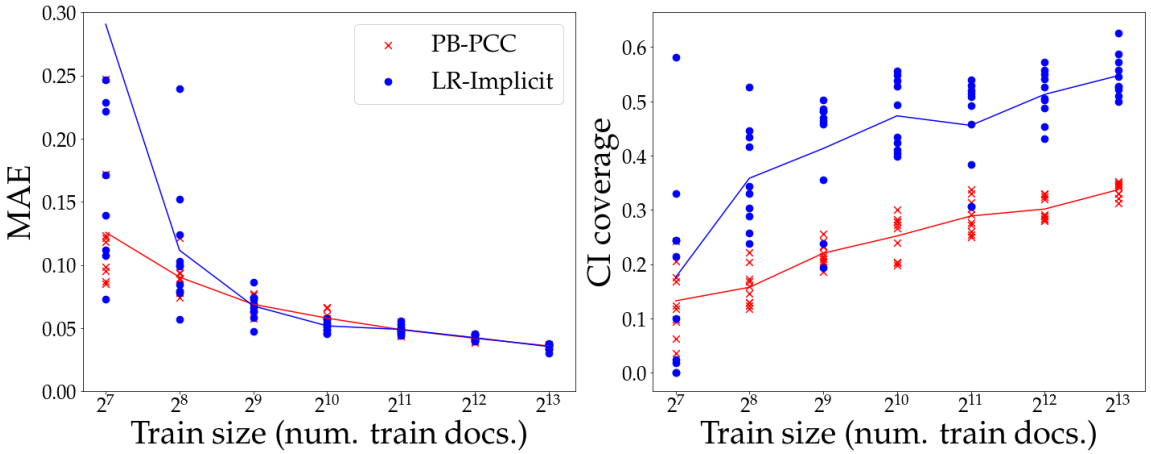
## 2.7 Conclusion

Document class prevalence estimation is a widespread and understudied task. We show that simple and obvious classifier aggregation methods display consistent biases, especially under class prior shift. Given how widely some of the less effective methods are used, machine learning and natural language processing research could have real impact in this space.

We also call attention to the need for *uncertainty aware* inference—methods that give confidence intervals to summarize their uncertainty. While our method is a first step, future work is necessary to better understand the problem and develop methods with improved coverage. Also, our framework can accommodate a wide array of document and language models—while we focus on bag-of-words models, recent advances in sequence, neural, and attention-based document models could be added directly to our generative model, or used as a discriminative-implicit component. The overall framework could also be extended to multiclass, and potentially, structured prediction settings.




(a) Varying training prevalence




(b) Varying training size

Figure 2.5: MAE and 90% CI coverage for PB-PCC while varying **(a)** training prevalence (the proportion of the 2000 training documents with positive reviews) and **(b)** training size (number of documents in the training data) with natural prevalence. Lines are the averages over 10 resamplings of training sets and points represent one resampling.

# CHAPTER 3

# ENTITY-EVENT MEASUREMENT FOR POLICE FATALITIES

We define *entity-event measurement* as measuring entities who are actors or recipients of certain events, and focus on a specific application of entity-event measurement—extracting the names of civilians killed by police from a collection of news reports. The remainder of this chapter consists of work originally published in Keith et al. [2017] and Keith et al. [2018].

## 3.1 Measuring police fatalities

### 3.1.1 Introduction

The United States government does not keep systematic records of when police kill civilians, despite a clear need for this information to serve the public interest and support social scientific analysis. Federal records rely on incomplete cooperation from local police departments, and human rights statisticians assess that they fail to document thousands of fatalities [Lum and Ball, 2015].

News articles have emerged as a valuable alternative data source. Organizations including The Guardian, The Washington Post, Mapping Police Violence, and Fatal Encounters have started to build such databases of U.S. police killings by manually reading millions of news articles[1] and extracting victim names and event details. This approach was recently validated by a Bureau of Justice Statistics study [Banks et al.,

[1]Fatal Encounters director D. Brian Burghart estimates he and colleagues have read 2 million news headlines and ledes to assemble its fatality records that date back to January, 2000 (pers. comm.); we find FE to be the most comprehensive publicly available database.

| Text | Person killed by police? |
|---|---|
| **Alton Sterling** was killed by police. | True |
| Officers shot and killed **Philando Castile**. | True |
| Officer **Andrew Hanson** was shot. | False |
| Police report **Megan Short** was fatally shot in apparent murder-suicide. | False |

Table 3.1: Toy examples (with entities in bold) illustrating the problem of extracting from text names of persons who have been killed by police.

2016], which augmented traditional police-maintained records with media reports, finding twice as many deaths compared to past government analyses. This suggests textual news data has enormous, real value, though *manual* news analysis remains extremely laborious.

We propose to help automate this process by extracting the names of persons killed by police from event descriptions in news articles (Table 3.1). This can be formulated as either of two cross-document entity-event extraction tasks:

1. Populating an entity-event database: From a corpus of news articles $\mathcal{D}^{(test)}$ over timespan $T$, extract the names of persons killed by police during that same timespan ($\mathcal{E}^{(pred)}$).

2. Updating an entity-event database: In addition to $\mathcal{D}^{(test)}$, assume access to both a historical database of killings $\mathcal{E}^{(train)}$ and a historical news corpus $\mathcal{D}^{(train)}$ for events that occurred before $T$. This setting often occurs in practice, and is the focus of this paper; it allows for the use of distantly supervised learning methods.[2]

The task itself has important social value, but the NLP research community may be interested in a scientific justification as well. We propose that police fatalities are

[2][Konovalov et al., 2017] studies the database update task where edits to Wikipedia infoboxes constitute events.

a useful test case for event extraction research. Fatalities are a well defined type of event with clear semantics for coreference, avoiding some of the more complex issues in this area [Hovy et al., 2013]. The task also builds on a considerable information extraction literature on knowledge base population (e.g. [Craven et al., 1998]). Finally, we posit that the field of natural language processing should, when possible, advance applications of important public interest. Previous work established the value of textual news for this problem, but computational methods could alleviate the scale of manual labor needed to use it.

To introduce this problem, we:

- Define the task of identifying persons killed by police, which is an instance of cross-document entity-event extraction (§3.2.1).
- Present a new dataset of web news articles collected throughout 2016 that describe possible fatal encounters with police officers (§3.2.2).
- Introduce, for the database update setting, a distant supervision model (§3.5) that incorporates feature-based logistic regression and convolutional neural network classifiers under a latent disjunction model.
- Demonstrate the approach's potential usefulness for practitioners: it outperforms two off-the-shelf event extractors (§3.3) and finds 39 persons not included in the Guardian's "The Counted" database of police fatalities as of January 1, 2017 (§3.6). This constitutes a promising first step, though performance needs to be improved for real-world usage.

### 3.1.2 Related work

This task combines elements of information extraction, including: *event extraction* (a.k.a. *semantic parsing*), identifying descriptions of events and their arguments from text, and cross-document *relation extraction*, predicting semantic relations over

entities. A fatality event indicates the killing of a particular person; we wish to specifically identify the names of fatality victims mentioned in text. Thus our task could be viewed as unary relation extraction: for a given person mentioned in a corpus, were they killed by a police officer?

Prior work in NLP has produced a number of event extraction systems, trained on text data hand-labeled with a pre-specified ontology, including ones that identify instances of killings [Li and Ji, 2014, Das et al., 2014]. Unfortunately, they perform poorly on our task (§3.3), so we develop a new method.

Since we do not have access to text specifically annotated for police killing events, we instead turn to *distant supervision*—inducing labels by aligning relation-entity entries from a gold standard database to their mentions in a corpus [Craven and Kumlien, 1999, Mintz et al., 2009, Bunescu and Mooney, 2007, Riedel et al., 2010]. Similar to this work, Reschke et al. [2014] apply distant supervision to multi-slot, template-based event extraction for airplane crashes; we focus on a simpler unary extraction setting with joint learning of a probabilistic model. Other related work in the cross-document setting has examined joint inference for relations, entities, and events [Yao et al., 2010, Lee et al., 2012, Yang et al., 2015].

Finally, other natural language processing efforts have sought to extract social behavioral event databases from news, such as instances of protests [Hanna, 2017], gun violence [Pavlick et al., 2016], and international relations [Schrodt and Gerner, 1994, Schrodt, 2012, Boschee et al., 2013, O'Connor et al., 2013, Gerrish, 2013]. They can also be viewed as event database population tasks, with differing levels of semantic specificity in the definition of "event."

| **Knowledge base** | **Historical** | **Test** |
|---|---|---|
| FE incident dates | Jan 2000 – Aug 2016 | Sep 2016 – Dec 2016 |
| FE gold entities ($\mathcal{G}$) | 17,219 | 452 |
| **News dataset** | **Train** | **Test** |
| doc. dates | Jan 2016 – Aug 2016 | Sep 2016 – Dec 2016 |
| total docs. ($\mathcal{D}$) | 866,199 | 347,160 |
| total ments. ($\mathcal{M}$) | 132,833 | 68,925 |
| pos. ments. ($\mathcal{M}^+$) | 11,274 | 6,132 |
| total entities ($\mathcal{E}$) | 49,203 | 24,550 |
| pos. entities ($\mathcal{E}^+$) | 916 | 258 |

Table 3.2: Data statistics for Fatal Encounters (FE) and scraped news documents. $\mathcal{M}$ and $\mathcal{E}$ result from NER processing, while $\mathcal{E}^+$ results from matching textual named entities against the gold-standard database ($\mathcal{G}$).

## 3.2 Task and data

### 3.2.1 Cross-document entity-event extraction for police fatalties

From a corpus of documents $\mathcal{D}$, the task is to extract a list of candidate person names, $\mathcal{E}$, and for each $e \in \mathcal{E}$ find

$$P(y_e = 1 \mid x_{\mathcal{M}(e)}). \tag{3.1}$$

Here $y \in \{0, 1\}$ is the entity-level label where $y_e = 1$ means a person (entity) $e$ was killed by police; $x_{\mathcal{M}(e)}$ are the sentences containing mentions $\mathcal{M}(e)$ of that person. A mention $i \in \mathcal{M}(e)$ is a token span in the corpus. Most entities have multiple mentions; a single sentence can contain multiple mentions of different entities.

### 3.2.2 News documents

We download a collection of web news articles by continually querying Google News[3] throughout 2016 with lists of police keywords (i.e police, officer, cop etc.) and fatality-related keywords (i.e. kill, shot, murder etc.). The keyword lists were constructed semi-automatically from cosine similarity lookups from the *word2vec* pre-trained word embeddings[4] in order to select a high-recall, broad set of keywords. The search is restricted to what Google News defines as a "regional edition" of "United States (English)" which seems to roughly restrict to U.S. news though we anecdotally observed instances of news about events in the U.K. and other countries. We apply a pipeline of text extraction, cleaning, and sentence de-duplication described in the appendix.

### 3.2.3 Entity and mention extraction

We process all documents with the open source *spaCy* NLP package[5] to segment sentences, and extract entity mentions. Mentions are token spans that (1) were identified as "persons" by spaCy's named entity recognizer, and (2) have a (firstname, lastname) pair as analyzed by the HAPNIS rule-based name parser,[6] which extracts, for example, (*John*, *Doe*) from the string *Mr. John A. Doe Jr.*.[7]

To prepare sentence text for modeling, our preprocessor collapses the candidate mention span to a special TARGET symbol. To prevent overfitting, other person

---

[3]https://news.google.com/

[4]https://code.google.com/archive/p/word2vec/

[5]Version 0.101.0, https://spacy.io/

[6]http://www.umiacs.umd.edu/~hal/HAPNIS/

[7]For both training and testing, we use a name matching assumption that a (firstname, lastname) match indicates coreference between mentions, and between a mention and a fatality database entity. This limitation does affect a small number of instances—the test set database contains the unique names of 453 persons but only 451 unique (firstname, lastname) tuples—but relaxing it raises complex issues for future work, such as how to evaluate whether a system correctly predicted two different fatality victims with the same name.

| | Rule | Prec. | Recall | F1 |
|---|---|---|---|---|
| SEMAFOR | R1 | 0.011 | 0.436 | 0.022 |
| | R2 | 0.031 | 0.162 | 0.051 |
| | R3 | 0.098 | 0.009 | 0.016 |
| RPI-JIE | R1 | 0.016 | 0.447 | 0.030 |
| | R2 | 0.044 | 0.327 | 0.078 |
| | R3 | 0.172 | 0.168 | **0.170** |
| Data upper bound (§3.5.6) | | 1.0 | 0.57 | 0.73 |

Table 3.3: Precision, recall, and F1 scores for test data using event extractors SEMAFOR and RPI-JIE and rules R1-R3 described below.

names are mapped to a different PERSON symbol; e.g. "TARGET was killed in an encounter with police officer PERSON."

There were initially 18,966,757 and 6,061,717 extracted mentions for the train and test periods respectively. To improve precision and computational efficiency, we filtered to sentences that contained at least one police keyword and one fatality keyword. This filter reduced positive entity recall a moderate amount (from 0.68 to 0.57), but removed 99% of the mentions, resulting in the $|\mathcal{M}|$ counts in Table 3.2.[8]

Other preprocessing steps included heuristics for extraction and name cleanups and are detailed in the appendix.

## 3.3 Off-the-shelf event extraction baselines

From a practitioner's perspective, a natural first approach to this task would be to run the corpus of police fatality documents through pre-trained, "off-the-shelf" event extractor systems that could identify killing events. In modern NLP research, a major paradigm for event extraction is to formulate a hand-crafted ontology of event classes, annotate a small corpus, and craft supervised learning systems to predict event parses of documents.

[8]In preliminary experiments, training and testing an n-gram classifier (§3.5.4) on the full mention dataset without keyword filtering resulted in a worse AUPRC than after the filter.

We evaluate two freely available, off-the-shelf event extractors that were developed under this paradigm: SEMAFOR [Das et al., 2014], and the RPI Joint Information Extraction System (RPI-JIE) [Li and Ji, 2014], which output semantic structures following the FrameNet [Fillmore et al., 2003] and ACE [Doddington et al., 2004] event ontologies, respectively.[9] [Pavlick et al., 2016] use RPI-JIE to identify instances of gun violence.

For each mention $i \in \mathcal{M}$ we use SEMAFOR and RPI-JIE to extract event tuples of the form $t_i =$ (event type, agent, patient) from the sentence $x_i$. We want the system to detect (1) killing events, where (2) the killed person is the target mention $i$, and (3) the person who killed them is a police officer. We implement a small progression of these neo-Davidsonian [Parsons, 1990] conjuncts with rules to classify $z_i = 1$ if:[10]

- **(R1)** the event type is 'kill.'
- **(R2)** R1 holds and the patient token span contains $e_i$.
- **(R3)** R2 holds and the agent token span contains a police keyword.

As in §3.5.1 (Eq. 3.3), we aggregate mention-level $z_i$ predictions to obtain entity-level predictions with a deterministic OR of $z_{\mathcal{M}(e)}$.

RPI-JIE under the full R3 system performs best, though all results are relatively poor (Table 3.3). Part of this is due to inherent difficulty of the task, though our

[9]Many other annotated datasets encode similar event structures in text, but with lighter ontologies where event classes directly correspond with lexical items—including PropBank, Prague Treebank, DELPHI-IN MRS, and Abstract Meaning Representation [Kingsbury and Palmer, 2002, Hajic et al., 2012, Oepen et al., 2014, Banarescu et al., 2013]. We assume such systems are too narrow for our purposes, since we need an extraction system to handle different trigger constructions like "killed" versus "shot dead."

[10]For SEMAFOR, we use the FrameNet 'Killing' frame with frame elements 'Victim' and 'Killer'. For RPI-JIE, we use the ACE 'life/die' event type/subtype with roles 'victim' and 'agent'. SEMAFOR defines a token span for every argument; RPI-JIE/ACE defines two spans, both a head word and entity extent; we use the entity extent. SEMAFOR only predicts spans as event arguments, while RPI-JIE also predicts entities as event arguments, where each entity has a within-text coreference chain over one or more mentions; since we only use single sentences, these chains tend to be small, though they do sometimes resolve pronouns. For determining R2 and R3, we allow a match on any of an entity's extents from any of its mentions.

| entity ($e$) | ment.($i$) prob. | ment. text ($x_i$) |
|---|---|---|
| **Keith Scott** (true pos) | 0.98 | Charlotte protests Charlotte's Mayor Jennifer Roberts speaks to reporters the morning after protests against the police shooting of **Keith Scott**, in Charlotte, North Carolina . |
| **Terence Crutcher** (true pos) | 0.96 | Tulsa Police Department released video footage Monday, Sept. 19, 2016, showing white Tulsa police officer Betty Shelby fatally shooting **Terence Crutcher,** 40, a black man police later determined was unarmed. |
| **Mark Duggan** (false pos) | 0.97 | The fatal shooting of **Mark Duggan** by police led to some of the worst riots in England's recent history. |
| **Logan Clarke** (false pos) | 0.92 | **Logan Clarke** was shot by a campus police officer after waving kitchen knives at fellow students outside the cafeteria at Hug High School in Reno, Nevada, on December 7. |

Table 3.4: Example of highly ranked entities, with selected mention predictions and text.

task-specific model still outperforms (Table 3.7). We suspect a major issue is that these systems heavily rely on their annotated training sets and may have significant performance loss on new domains, or messy text extracted from web news, suggesting domain transfer for future work.

## 3.4 Probabilistic rule-based IE with dependency parses

### 3.4.1 Summary of Monte Carlo syntax marginals and dependency path prediction

Dependency parses (e.g. [Chen and Manning, 2014]) are often used in downstream applications, such as the entity-event measurement discussed in this chapter. One commonly used parse substructure is the *dependency path* between two words, which is widely used in unsupervised lexical semantics [Lin and Pantel, 2001], distantly supervised lexical semantics [Snow et al., 2005], relation learning [Riedel et al., 2013], and supervised semantic role labeling [Hacioglu, 2004, Das et al., 2014], as well as

applications in economics [Ghose et al., 2007], political science [O'Connor et al., 2013], biology [Fundel et al., 2006], and the humanities [Bamman et al., 2013, 2014].

Keith et al. [2018] present a *Monte Carlo syntax marginal* inference method which exploits information across samples of the entire parse forest. It achieves higher accuracy predictions than a traditional greedy parsing algorithm, and allows tradeoffs between precision and recall. Keith et al. [2018] define a *dependency path* to be a set of edges from the dependency parse; for example, a length-2 path $p = \{\text{nsubj}(3, 1), \text{dobj}(3, 4)\}$ connects tokens 1 and 4. They define *dependency path prediction* as the task of predicting a set of dependency paths for a sentence; the paths do not necessarily have to come from the same tree, nor even be consistent with a single syntactic analysis. They approach this task with their Monte Carlo syntax marginal method, by predicting paths from the transition sampling parser. They treat each possible path as a structure query and return all paths whose marginal probabilities are at least threshold $t$. Varying $t$ trades off precision and recall.

### 3.4.2 Police killings victim extraction

Supervised learning typically gives the most accurate information extraction or semantic parsing systems, but for many applications where training data is scarce, Chiticariu et al. [2013] argue that rule-based systems are useful and widespread in practice, despite their neglect in contemporary NLP research. Syntactic dependencies are a useful abstraction with which to write rule-based extractors, but they can be brittle due to errors in the parser. We propose to integrate over parse samples to infer a *marginal* probability of a rule match, increasing robustness and allowing for precision-recall tradeoffs.

We examine the task of extracting the list of names of persons killed by police from a test set of web news articles in Sept–Dec 2016. We use the dataset released by Keith et al. [2017], consisting of 24,550 named entities $e \in \mathcal{E}$ and sentences from noisy

web news text extractions (that can be difficult to parse), each of which contains at least one $e$ (on average, 2.8 sentences/name) as well as keywords for both police and killing/shooting. The task is to classify whether a given name is a person who was killed by police, given 258 gold-standard names that have been verified by journalists.

### 3.4.3 Dependency rule extractor

In Section 3.3, we present a baseline rule-based method that uses Li and Ji [2014]'s off-the-shelf RPI-JIE ACE event parser to extract (event type, agent, patient) tuples from sentences, and assigns $f_{\text{JIE}}(x_i, e) = 1$ iff the event type was a killing, the agent's span included a police keyword, and the patient was the candidate entity $e$. An entity is classified as a victim if at least one sentence is classified as true, resulting in a 0.17 F1 score (as reported in previous work).[11]

We define a similar syntactic dependency rule system using a dependency parse as input: our extractor $f(x, e, y)$ returns 1 iff the sentence has a killing keyword $k$,[12] which both

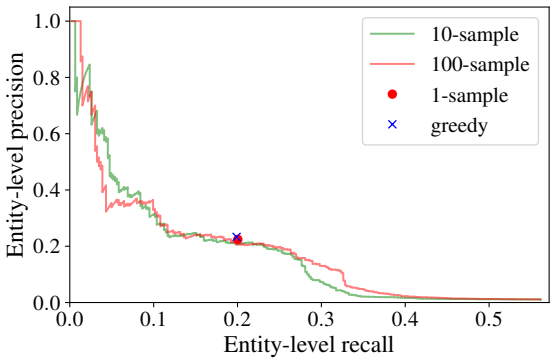
1. has an agent token $a$ (defined as, governed by *nsubj* or *nmod*) which is a police keyword, or $a$ has a (*amod* or *compound*) modifier that is a police keyword; and,
2. has a patient token $p$ (defined as, governed by *nsubjpass* or *dobj*) contained in the candidate name $e$'s span.

Applying this $f(x, e, y)$ classifier to greedy parser output, it performs better than the RPI-JIE-based rules (Figure 3.1, right), perhaps because it is better customized for the particular task.

Treating $f$ as a structure query, we then use our Monte Carlo marginal inference (§3.5) method to calculate the probability of a rule match for each sentence—that

[11] This measures recall of the entire gold-standard victim database, though the corpus only includes 57% of the victims.

[12] Police and killing/shooting keywords are from Keith et al.'s publicly released software.




| Method | F1 |
|---|---|
| RPI-JIE | 0.170 |
| Greedy | 0.215 |
| 1 samp. | 0.212 |
| 10 samp. | 0.219 |
| 100 samp. | **0.222** |

Figure 3.1: **Left:** Rule-based entity precision and recall for police fatality victims, with greedy parsing and Monte Carlo inference. **Right:** F1 scores for RPI-JIE, Greedy, and 1-sample methods, and maximum F1 on PR curve for probabilistic (multiple sample) inference.

is, the fraction of parse samples where $f(x, e, y^{(s)})$ is true—and infer the entity's probability with the *noisy-or* formula [Craven and Kumlien, 1999, Keith et al., 2017]. This gives soft classifications for entities.

### 3.4.4 Results

The Monte Carlo method achieves slightly higher F1 scores once there are at least 10 samples (Fig. 3.1, right). More interestingly, the soft entity-level classifications also allow for precision-recall tradeoffs (Fig. 3.1, left), which could be used to prioritize the time of human reviewers updating the victim database (filter to higher precision), or help ensure victims are not missed (with higher recall). We found the sampling method retrieved several true-positive entities where only a single sentence had a non-zero rule prediction at probability 0.01—that is, the rule was only matched in one of 100 sampled parses. Since current practitioners are already manually reviewing millions of news articles to create police fatality victim databases, the ability to filter to high recall—even with low precision—may be useful to help ensure victims are not missed.

| | $x$ | $z$ | $y$ |
|---|---|---|---|
| "Hard" training | observed | fixed (distantly labeled) | observed |
| "Soft" (EM) training | observed | latent | observed |
| Testing | observed | latent | latent |

Table 3.5: Training and testing settings for mention sentences $x$, mention labels $z$, and entity labels $y$.

**Supervised learning.** Sampling also slightly improves supervised learning for this problem. We modify Keith et al.'s logistic regression model based on a dependency path feature vector $f(x_i, y)$, instead creating feature vectors that average over multiple parse samples ($E_{\tilde{p}(y)}[f(x_i, y)]$) at both train and test time. With the greedy parser, the model results in 0.229 F1; using 100 samples slightly improves performance to 0.234 F1.

## 3.5 Additional models

Our goal is to classify entities as to whether they have been killed by police (§3.5.1). Since we do not have gold-standard labels to train our model, we turn to *distant supervision* [Craven and Kumlien, 1999, Mintz et al., 2009], which heuristically aligns facts in a knowledge base to text in a corpus to impute positive mention-level labels for supervised learning. Previous work typically examines distant supervision in the context of binary relation extraction [Bunescu and Mooney, 2007, Riedel et al., 2010, Hoffmann et al., 2011], but we are concerned with the unary predicate "person was killed by police." As our gold standard knowledge base ($\mathcal{G}$), we use Fatal Encounters' (FE) publicly available dataset: around 18,000 entries of victim's name, age, gender and race as well as location, cause and date of death. (We use a version of the FE database downloaded Feb. 27, 2017.) We compare two different distant supervision training paradigms (Table 3.5): "hard" label training (§3.5.2) and "soft" EM-based

training (§3.5.3). This section also details mention-level models (§3.5.4,§3.5.5) and evaluation (§3.5.6).

### 3.5.1 Novel approach: latent disjunction model

Our discriminative model is built on mention-level probabilistic classifiers. Recall a single entity will have one or more mentions (i.e. the same name occurs in multiple sentences in our corpus). For a given mention $i$ in sentence $x_i$, our model predicts whether the person is described as having been killed by police, $z_i = 1$, with a binary logistic model,

$$P(z_i = 1 \mid x_i) = \sigma(\beta^\mathsf{T} f_\gamma(x_i)). \tag{3.2}$$

We experiment with both logistic regression (§3.5.4) and convolutional neural networks (§3.5.5) for this component, which use logistic regression weights $\beta$ and feature extractor parameters $\gamma$. Then we must somehow aggregate mention-level decisions to determine entity labels $y_e$.[13] If a human reader were to observe at least one sentence that states a person was killed by police, they would infer that person was killed by police. Therefore we aggregate an entity's mention-level labels with a deterministic disjunction:

$$P(y_e = 1 \mid z_{\mathcal{M}(e)}) = 1 \left\{ \vee_{i \in \mathcal{M}(e)} \; z_i \right\}. \tag{3.3}$$

At test time, $z_i$ is latent. Therefore the correct inference for an entity is to marginalize out the model's uncertainty over $z_i$:

$$P(y_e = 1 | x_{\mathcal{M}(e)}) = 1 - P(y_e = 0 | x_{\mathcal{M}(e)}) \tag{3.4}$$

$$= 1 - P(z_{\mathcal{M}(e)} = \vec{0} \mid x_{\mathcal{M}(e)}) \tag{3.5}$$

$$= 1 - \prod_{i \in \mathcal{M}(e)} (1 - P(z_i = 1 \mid x_i)). \tag{3.6}$$

[13] An alternative approach is to aggregate features across mentions into an entity-level feature vector [Mintz et al., 2009, Riedel et al., 2010]; but here we opt to directly model at the mention level, which can use contextual information.

Eq. 3.6 is the *noisyor* formula [Pearl, 1988, Craven and Kumlien, 1999]. Procedurally, it counts strong probabilistic predictions as evidence, but can also incorporate a large number of weaker signals as positive evidence as well.[14]

In order to train these classifiers, we need mention-level labels ($z_i$) which we impute via two different distant supervision labeling methods: "hard" and "soft."

### 3.5.2 "Hard" distant label training

In "hard" distant labeling, labels for mentions in the training data are heuristically imputed and directly used for training. We use two labeling rules. First, **name-only:**

$$z_i = 1 \text{ if } \exists e \in \mathcal{G}^{(train)} : \text{name}(i) = \text{name}(e). \tag{3.7}$$

This is the direct unary predicate analogue of [Mintz et al., 2009]'s *distant supervision assumption*, which assumes every mention of a gold-positive entity exhibits a description of a police killing.

This assumption is not correct. We manually analyze a sample of positive mentions and find 36 out of 100 name-only sentences did not express a police fatality event—for example, sentences contain commentary, or describe killings not by police. This is similar to the precision for distant supervision of binary relations found by [Riedel et al., 2010], who reported 10–38% of sentences did not express the relation in question.

---

[14] In early experiments, we experimented with other, more ad-hoc aggregation rules with a "hard"-trained model. The maximum and arithmetic mean functions performed worse than *noisyor*, giving credence to the disjunction model. The sum rule ($\sum_i P(z_i = 1 \mid x_i)$) had similar ranking performance as *noisyor*—perhaps because it too can use weak signals, unlike mean or max—though it does not yield proper probabilities between 0 and 1.

Our higher precision rule, **name-and-location**, leverages the fact that the location of the fatality is also in the Fatal Encounters database and requires both to be present:

$$\begin{gathered} z_i = 1 \text{ if } \exists e \in \mathcal{G}^{(train)} : \\ \text{name}(i) = \text{name}(e) \text{ and location}(e) \in x_i. \end{gathered} \tag{3.8}$$

We use this rule for training since precision is slightly better, although there is still a considerable level of noise.

### 3.5.3 "Soft" (EM) joint training

At training time, the *distant supervision assumption* used in "hard" label training is flawed: many positively-labeled mentions are in sentences that do not assert the person was killed by a police officer. Alternatively, at training time we can treat $z_i$ as a latent variable and assume, as our model states, that *at least one* of the mentions asserts the fatality event, but leave uncertainty over which mention (or multiple mentions) conveys this information. This corresponds to multiple instance learning (MIL; [Dietterich et al., 1997]) which has been applied to distantly supervised relation extraction by enforcing the *at least one* constraint at training time [Bunescu and Mooney, 2007, Riedel et al., 2010, Hoffmann et al., 2011, Surdeanu et al., 2012, Ritter et al., 2013]. Our approach differs by using exact marginal posterior inference for the E-step.

With $z_i$ as latent, the model can be trained with the EM algorithm [Dempster et al., 1977]. We initialize the model by training on the "hard" distant labels (§3.5.2), and then learn improved parameters by alternating E- and M-steps.

The **E-step** requires calculating the marginal posterior probability for each $z_i$,

$$q(z_i) := P(z_i \mid x_{\mathcal{M}(e_i)}, y_{e_i}). \tag{3.9}$$

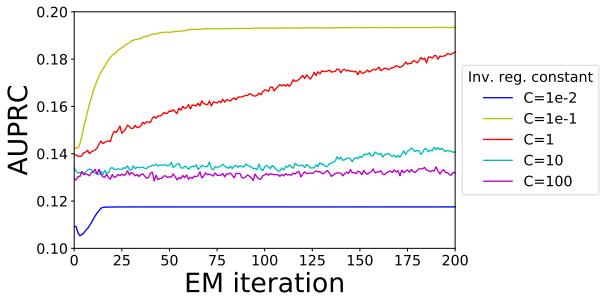



Figure 3.2: For soft-LR (EM), area under precision recall curve (AUPRC) results on the test set during training, for different inverse regularization values ($C$, the parameters' prior variance).

This corresponds to calculating the posterior probability of a disjunct, given knowledge of the output of the disjunction, and prior probabilities of all disjuncts (given by the mention-level classifier).

Since $P(z \mid x, y) = P(z, y \mid x)/P(y \mid x)$,

$$q(z_i = 1) = \frac{P(z_i = 1, y_{e_i} = 1 | x_{\mathcal{M}(e_i)})}{P(y_{e_i} = 1 | x_{\mathcal{M}(e_i)})}. \tag{3.10}$$

The numerator simplifies to the mention prediction $P(z_i = 1 \mid x_i)$ and the denominator is the entity-level *noisyor* probability (Eq. 3.6). This has the effect of taking the classifier's predicted probability and increasing it slightly (since Eq. 3.10's denominator is no greater than 1); thus the disjunction constraint implies a soft positive labeling. In the case of a negative entity with $y_e = 0$, the disjunction constraint implies all $z_{\mathcal{M}(e)}$ stay clamped to 0 as in the "hard" label training method.

The $q(z_i)$ posterior weights are then used for the **M-step**'s expected log-likelihood objective:

$$\max_\theta \sum_i \sum_{z \in \{0,1\}} q(z_i = z) \log P_\theta(z_i = z \mid x_i). \tag{3.11}$$

This objective (plus regularization) is maximized with gradient ascent as before.

This approach can be applied to any mention-level probabilistic model; we explore two in the next sections.

| | Features |
|---|---|
| $D1$ | length 3 dependency paths that include TARGET: word, POS, dep. label |
| $D2$ | length 3 dependency paths that include TARGET: word and dep. label |
| $D3$ | length 3 dependency paths that include TARGET: word and POS |
| $D4$ | all length 2 dependency paths with word, POS, dep. labels |
| $N1$ | n-grams length 1, 2, 3 |
| $N2$ | n-grams length 1, 2, 3 plus POS tags |
| $N3$ | n-grams length 1, 2, 3 plus directionality and position from TARGET |
| $N4$ | concatenated POS tags of 5-word window centered on TARGET |
| $N5$ | word and POS tags for 5-word window centered on TARGET |

Table 3.6: Feature templates for logistic regression grouped into syntactic dependencies ($D$) and N-gram ($N$) features.

### 3.5.4 Feature-based logistic regression

We construct hand-crafted features for regularized logistic regression (LR) (Table 3.6), designed to be broadly similar to the n-gram and syntactic dependency features used in previous work on feature-based semantic parsing (e.g. [Das et al., 2014, Thomson et al., 2014]). We use randomized feature hashing [Weinberger et al., 2009] to efficiently represent features in 450,000 dimensions, which achieved similar performance as an explicit feature representation. The logistic regression weights ($\beta$ in Eq. 3.2) are learned with *scikit-learn* [Pedregosa et al., 2011].[15] For EM (soft-LR) training, the test set's area under the precision recall curve converges after 96 iterations (Fig. 3.2).

[15]With *FeatureHasher*, L2 regularization, 'lbfgs' solver, and inverse strength $C = 0.1$, tuned on a development dataset in "hard" training; for EM training the same regularization strength performs best.

### 3.5.5 Convolutional neural network

We also train a convolutional neural network (CNN) classifier, which uses word embeddings and their nonlinear compositions to potentially generalize better than sparse lexical and n-gram features. CNNs have been shown useful for sentence-level classification tasks [Kim, 2014, Zhang and Wallace, 2015], relation classification [Zeng et al., 2014] and, similar to this setting, event detection [Nguyen and Grishman, 2015]. We use [Kim, 2014]'s open-source CNN implementation,[16] where a logistic function makes the final mention prediction based on max-pooled values from convolutional layers of three different filter sizes, whose parameters are learned ($\gamma$ in Eq. 3.2). We use pretrained word embeddings for initialization,[17] and update them during training. We also add two special vectors for the TARGET and PERSON symbols, initialized randomly.[18]

For training, we perform stochastic gradient descent for the negative expected log-likelihood (Eq. 3.11) by sampling with replacement fifty mention-label pairs for each minibatch, choosing each $(i, k) \in \mathcal{M} \times \{0, 1\}$ with probability proportional to $q(z_i = k)$. This strategy attains the same expected gradient as the overall objective. We use "epoch" to refer to training on 265,700 examples (approx. twice the number of mentions). Unlike EM for logistic regression, we do not run gradient descent to convergence, instead applying an E-step every two epochs to update $q$; this approach is related to incremental and online variants of EM [Neal and Hinton, 1998, Liang and Klein, 2009], and is justified since both SGD and E-steps improve the evidence lower bound (ELBO). It is also similar to [Salakhutdinov et al., 2003]'s expectation gradient method; their analysis implies the gradient calculated immediately after an

[16] https://github.com/yoonkim/CNN_sentence

[17] From the same *word2vec* embeddings used in §3.2.

[18] Training proceeds with ADADELTA [Zeiler, 2012]. We tested several different settings of dropout and L2 regularization hyperparameters on a development set, but found mixed results, so used their default values.

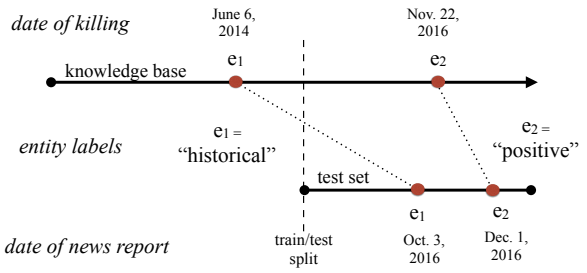



Figure 3.3: At test time, there are matches between the knowledge base and the news reports both for persons killed during the test period ("positive") and persons killed before it ("historical"). Historical cases are excluded from evaluation.

E-step is in fact the gradient for the marginal log-likelihood. We are not aware of recent work that uses EM to train latent-variable neural network models, though this combination has been explored (e.g. [Jordan and Jacobs, 1994])

### 3.5.6 Evaluation

On documents from the test period (Sept–Dec 2016), our models predict entity-level labels $P(y_e = 1 \mid x_{\mathcal{M}(e)})$ (Eq. 3.6), and we wish to evaluate whether retrieved entities are listed in Fatal Encounters as being killed during Sept–Dec 2016. We rank entities by predicted probabilities to construct a precision-recall curve (Fig. 3.5, Table 3.7). Area under the precision-recall curve (AUPRC) is calculated with a trapezoidal rule; F1 scores are shown for convenient comparison to non-ranking approaches (§3.3).

**Excluding historical fatalities:** Our model gives strong positive predictions for many people who were killed by police before the test period (i.e. before Sept 2016), when news articles contain discussion of historical police killings. We exclude these entities from evaluation, since we want to simulate an update to a fatality database (Fig 3.3). Our test dataset contains 1,148 such historical entities.

**Data upper bound:** Of the 452 gold entities in the FE database at test time, our news corpus only contained 258 (Table 3.2), hence the data upper bound of 0.57 recall, which also gives an upper bound of 0.57 on AUPRC. This is mostly a limitation

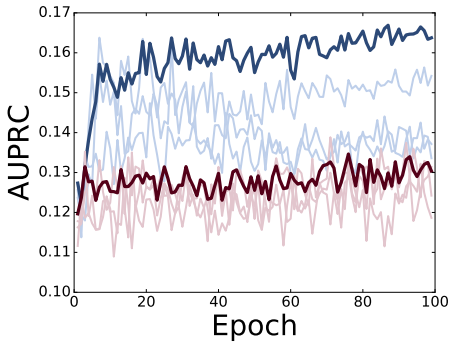



Figure 3.4: Test set AUPRC for three runs of soft-CNN (EM) (**blue**, higher in graph), and hard-CNN (**red**, lower in graph). Darker lines show performance of averaged predictions.

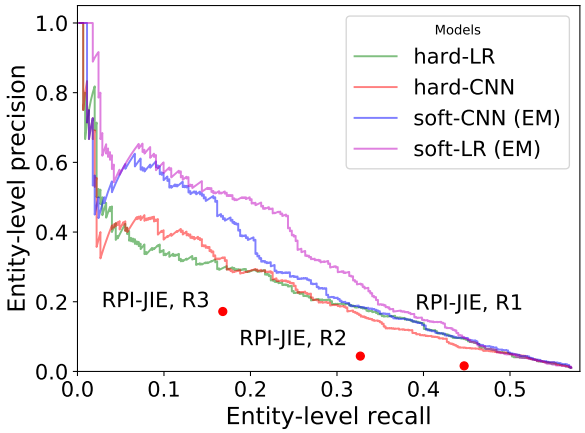



Figure 3.5: Precision-recall curves for the given models.

of our news corpus; though we collect hundreds of thousands of news articles, it turns out Google News only accesses a subset of relevant web news, as opposed to more comprehensive data sources manually reviewed by Fatal Encounters' human experts. We still believe our dataset is large enough to be realistic for developing better methods, and expect the same approaches could be applied to a more comprehensive news corpus.

| Model | AUPRC | F1 |
|---|---|---|
| hard-LR, dep. feats. | 0.117 | 0.229 |
| hard-LR, n-gram feats. | 0.134 | 0.257 |
| hard-LR, all feats. | 0.142 | 0.266 |
| hard-CNN | 0.130 | 0.252 |
| soft-CNN (EM) | 0.164 | 0.267 |
| **soft-LR (EM)** | **0.193** | **0.316** |
| Data upper bound (§3.5.6) | 0.57 | 0.73 |

Table 3.7: Area under precision-recall curve (AUPRC) and F1 (its maximum value from the PR curve) for entity prediction on the test set.

## 3.6 Results and discussion

**Significance testing:** We would like to test robustness of performance results to the finite datasets with bootstrap testing [Berg-Kirkpatrick et al., 2012], which can accomodate performence metrics like AUPRC. It is not clear what the appropriate unit of resampling should be—for example, parsing and machine translation research in NLP often resamples sentences, which is inappropriate for our setting. We elect to resample documents in the test set, simulating variability in the generation and retrieval of news articles. Standard errors for one model's AUPRC and F1 are in the range 0.004–0.008 and 0.008–0.010 respectively; we also note pairwise significance test results. See appendix for details.

**Overall performance:** Our results indicate our model is better than existing computational methods methods to extract names of people killed by police, by comparing to F1 scores of off-the-shelf extractors (Table 3.7 vs. Table 3.3; differences are statistically significant).

We also compare entities extracted from our test dataset to the Guardian's "The Counted" database of U.S. police killings during the span of the test period (Sept.–

Dec., 2016),[19] and found 39 persons they did not include in the database, but who were in fact killed by police. This implies our approach could augment journalistic collection efforts. Additionally, our model could help practitioners by presenting them with sentence-level information in the form of Table 3.4; we hope this could decrease the amount of time and emotional toll required to maintain real-time updates of police fatality databases.

**CNN:** Model predictions were relatively unstable during the training process. Despite the fact that EM's evidence lower bound objective ($H(Q) + E_Q[\log P(Z, Y|X)]$) converged fairly well on the training set, test set AUPRC substantially fluctuated as much as 2% between epochs, and also between three different random initializations for training (Fig. 3.4). We conducted these multiple runs initially to check for variability, then used them to construct a basic ensemble: we averaged the three models' mention-level predictions before applying *noisyor* aggregation. This outperformed the individual models—especially for EM training—and showed less fluctuation in AUPRC, which made it easier to detect convergence. Reported performance numbers in Table 3.7 are with the average of all three runs from the final epoch of training.

**LR vs. CNN:** After feature ablation we found that hard-CNN and hard-LR with n-gram features (N1-N5) had comparable AUPRC values (Table 3.7). But adding dependency features (D1-D4) caused the logistic regression models to outperform the neural networks (albeit with bare significance: $p = 0.046$). We hypothesize these dependency features capture longer-distance semantic relationships between the entity, fatality trigger word, and police officer, which short n-grams cannot. Moving to sequence or graph LSTMs may better capture such dependencies.

**Soft (EM) training:** Using the EM algorithm gives substantially better performance: for the CNN, AUC improves from 0.130 to 0.164, and for LR, from 0.142 to

[19] https://www.theguardian.com/us-news/series/counted-us-police-killings, downloaded Jan. 1, 2017.

0.193. (Both improvements are statistically significant.) Logistic regression with EM training is the most accurate model. Examining the precision-recall curves (Fig. 3.5), many of the gains are in the higher confidence predictions (left side of figure). In fact, the soft EM model makes fewer strongly positive predictions: for example, hard-LR predicts $y_e = 1$ with more than 99% confidence for 170 out of 24,550 test set entities, but soft-LR does so for only 24. This makes sense given that the hard-LR model at training time assumes that many more positive entity mentions are evidence of a killing than they are in reality (§3.5.2).

**Manual analysis:** Manual analysis of false positives indicates misspellings or mismatches of names, police fatalities outside of the U.S., people who were shot by police but not killed, and names of police officers who were killed are common false positive errors (see detailed table in the appendix). This suggests many prediction errors are from ambiguous or challenging cases.[20]

## 3.7 Future work

While we have made progress on this application, more work is necessary for accuracy to be high enough to be useful for practitioners. Our model allows for the use of mention-level semantic parsing models; systems with explicit trigger/agent/patient representations, more like traditional event extraction systems, may be useful, as would more sophisticated neural network models, or attention models as an alternative to disjunction aggregation [Lin et al., 2016].

One goal is to use our model as part of a semi-automatic system, where people manually review a ranked list of entity suggestions. In this case, it is more important to focus on improving recall—specifically, improving precision at high-recall points on the precision-recall curve. Our best models, by contrast, tend to improve precision

[20]We attempted to correct non-U.S. false positive errors by using CLAVIN, an open-source country identifier, but this significantly hurt recall.

at lower-recall points on the curve. Higher recall may be possible through cost-sensitive training (e.g. [Gimpel and Smith, 2010]) and using features from beyond single sentences within the document.

Furthermore, our dataset could be used to contribute to communication studies, by exploring research questions about the dynamics of media attention (for example, the effect of race and geography on coverage of police killings), and discussions of historical killings in news—for example, many articles in 2016 discussed Michael Brown's 2014 death in Ferguson, Missouri. Improving NLP analysis of historical events would also be useful for the event extraction task itself, by delineating between recent events that require a database update, versus historical events that appear as "noise" from the perspective of the database update task. Finally, it may also be possible to adapt our model to extract other types of social behavior events.

# CHAPTER 4

# USING TEXT TO REDUCE CONFOUNDING FROM CAUSAL ESTIMATES

This chapter was originally published as Keith et al. [2020a].

## 4.1 Introduction

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

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

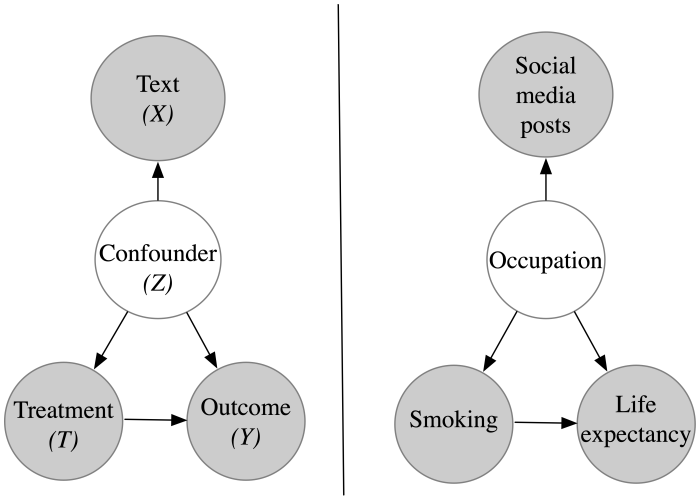



Figure 4.1: *Left:* A causal diagram for text that encodes causal confounders, the setting that is focus of this review paper. The major assumption is that latent confounders can be *measured* from text and those confounder measurements can be used in causal adjustments. *Right:* An example application in which practitioner does not have access to the confounding variable, *occupation*, in structured form but can measure confounders from unstructured text (e.g. an individual's social media posts).

the confounding bias statistically with methods such as matching, propensity score weighting, or regression adjustment (§4.5).

In causal research about human behavior and society, there are potentially many latent confounding variables that can be measured from unstructured text data. Text data could either (a) serve as a surrogate for potential confounders; or (b) the language of text itself could be a confounder. Our running example is an instance of text as a surrogate: a researcher may not have a record of an individual's occupation but could attempt to measure this variable from the individual's entire history of social media posts (see Fig. 4.1). An example of text as a direct confounder: the linguistic content of social media posts could influence censorship (treatment) and future posting rates (outcome) [Roberts et al., 2020].

A challenging aspect of this research design is the high-dimensional nature of text. Other work has explored general methods for adjusting for high-dimensional

confounders [D'Amour et al., 2021, Rassen et al., 2011, Louizos et al., 2017, Li et al., 2016, Athey et al., 2017]. However, text data differ from other high-dimensional data-types because intermediate confounding adjustments can be read and evaluated by humans (§4.6) and designing meaningful representations of text is still an open research question.[1] Even when applying simple adjustment methods, a practitioner must first transform text into a lower-dimensional representation via, for example, filtered word counts, lexicon indicators, topic models, or embeddings (§4.4). An additional challenge is that empirical evaluation in causal inference is still an open research area [Dorie et al., 2019, Gentzel et al., 2019] and text adds to the difficulty of this evaluation (§4.7).

We narrow the scope of this chapter to review methods and applications with text data as a causal *confounder*. In the broader area of text and causal inference, work has examined text as a mediator [Veitch et al., 2020], text as treatment [Fong and Grimmer, 2016, Egami et al., 2018, Wood-Doughty et al., 2018, Tan et al., 2014], text as outcome [Egami et al., 2018], causal discovery from text [Mani and Cooper, 2000], and predictive (Granger) causality with text [Balashankar et al., 2019, del Prado Martin and Brendel, 2016, Tabari et al., 2018].

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

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

- **For applied practitioners,** we collect and categorize applications with text as a causal confounder (Table 4.1 and §4.2), and we provide a flow-chart of analysts' decisions for this problem setting (Fig. 4.2).

- **For causal inference researchers working with text data,** we highlight recent work in representation learning in NLP (§4.4) and caution that this is still an open research area with questions of the sensitivity of effects to choices in representation. We also outline existing interpretable evaluation methods for adjustments of text as a causal confounder (§4.6).

- **For NLP researchers working with causal inference,** we summarize some of the most-used causal estimators that condition on confounders: matching, propensity score weighting, regression adjustment, doubly-robust methods, and causally-driven representation learning (§4.5). We also discuss evaluation of methods with constructed observational studies and semi-synthetic data (§4.7).

## 4.2 Applications

In Table 4.1, we gather and summarize applications that use text to adjust for potential confounding. This encompasses both (a) text as a surrogate for confounders, or (b) the language itself as confounders.[2]

As an example, consider Kiciman et al. [2018] where the goal is to estimate the size of the causal effect of alcohol use (treatment) on academic success (outcome) for college students. Since randomly assigning college students to binge drink is not feasible or ethical, the study instead uses observational data from Twitter, which also

[2] We acknowledge that Table 4.1 is by no means exhaustive. To construct Table 4.1, we started with three seed papers: Roberts et al. [2020], Veitch et al. [2020], and Wood-Doughty et al. [2018]. We then examined papers cited by these papers, papers that cited these papers, and papers published by the papers' authors. We repeated this approach with the additional papers we found that adjusted for confounding with text. We also examined papers matching the query "causal" or "causality" in the ACL Anthology.

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

Table 4.1: Example applications that infer the causal effects of treatment on outcome by measuring confounders (unobserved) from text data (observed). In doing so, these applications choose a representation of text (text rep.) and a method to adjust for confounding.

has the advantage of a large sample size of over sixty-three thousand students. They use heuristics to identify the Twitter accounts of college-age students and extract alcohol mentions and indicators of college success (e.g., study habits, risky behaviors, and emotions) from their Twitter posts. They condition on an individual's previous posts (temporally previous to measurements of treatment and outcome) as confounding variables since they do not have demographic data. They represent text as word counts and use stratified propensity score matching to adjust for the confounding

bias. The study finds the effects of alcohol use include decreased mentions of study habits and positive emotions and increased mentions of potentially risky behaviors.

**Text as a surrogate for confounders.** Traditionally, causal research that uses human subjects as the unit of analysis would infer demographics via surveys. However, with the proliferation of the web and social media, social research now includes large-scale observational data that would be challenging to obtain using surveys [Salganik, 2017]. This type of data typically lacks demographic information but may contain large amounts of text written by participants from which demographics can be extracted. In this space, some researchers are specific about the confounders they want to extract such as an individual's ideology [Sridhar and Getoor, 2019] or mood [Sridhar et al., 2018]. Other researchers condition on all the text they have available and assume that low-dimensional summaries capture all possible confounders. For example, researchers might assume that text encodes all possible confounders between alcohol use and college success [Kiciman et al., 2018] or psychiatric medication and anxiety [Saha et al., 2019]. We dissect and comment on this assumption in Section 4.8.

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

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

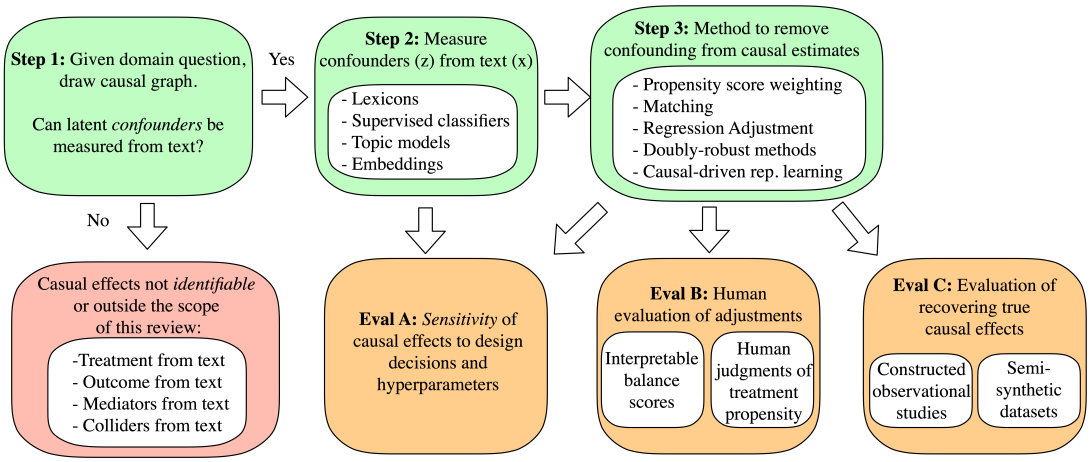



Figure 4.2: This chart is a guide to design decisions for applied research with causal confounders from text. *Step 1:* Encode domain assumptions by drawing a causal diagram (§4.3). If the application does not use text to measure latent *confounders*, the causal effects are not identifiable or the application is outside the scope of this review. *Step 2:* Use NLP to measure confounders from text (§4.4). *Step 3:* Choose a method that adjusts for confounding in causal estimates (§4.5). Evaluation should include *(A)* sensitivity analysis (§4.4), *(B)* human evaluation of adjustments when appropriate (§4.6), and *(C)* evaluation of recovering the true causal effects (§4.7).

that women write about certain topics) and the number of citations of that article ("hotter" topics will receive more citations). Other domains that analyze language as a confounder include news [Johansson et al., 2016], social media [De Choudhury et al., 2016, Olteanu et al., 2017], and loan descriptions [Pham and Shen, 2017]. See Section 4.4 for more discussion on the challenges and open problems of inferring these latent aspects of language.

## 4.3 Estimating causal effects

Two predominant causal inference frameworks are *structural causal models (SCM)* [Pearl, 2009b] and *potential outcomes* [Rubin, 1974, 2005], which are complementary and theoretically connected [Pearl, 2009b, Richardson and Robins, 2013, Morgan and Winship, 2015]. While their respective goals substantially overlap, methods from structural causal models tend to emphasize conceptualizing, expressing, and reasoning

about the effects of possible causal relationships among variables, while methods from potential outcomes tend to emphasize estimating the size or strength of causal effects.

### 4.3.1 Potential outcomes framework

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

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

$$\hat{\tau}_{\text{naive}} = \frac{1}{n_1} \sum_{i:t_i=1} y_i - \frac{1}{n_0} \sum_{j:t_j=0} y_j \tag{4.1}$$

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

---

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

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

### 4.3.2 Structural causal models framework

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

$$P(Y \mid do(T = t)) \tag{4.2}$$

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

**Identification.** In most cases, Equation 4.2 is *not* equal to the ordinary conditional distribution $P(Y \mid T = t)$ since the latter is simply filtering to the subpopulation and the former is changing the underlying data distribution via intervention. Thus, for observational studies that lack intervention, one needs an *identification strategy* in order to represent $P(Y \mid do(T = t))$ in terms of distributions of observed variables. One such identification strategy (assumed by the applications throughout this review) is the *backdoor criterion* which applies to a set of variables, $\mathcal{S}$, if they (i) block every backdoor path between treatment and outcome, and (ii) no node in $\mathcal{S}$ is a descendant of treatment. Without positive identification, the causal effects cannot be estimated and measuring variables from text is a secondary concern.

**Drawing the causal graph.** Causal graphs help clarify which variables should and should not be conditioned on. The causal graphs in Figure 4.3 illustrate how the direction of the arrows differentiates confounder, collider, and mediator variables. Identifying the differences in these variables is crucial since, by *d-separation*, conditioning on a confounder will block the treatment-confounder-outcome path, removing bias. By contrast, conditioning on a collider can create dependence between

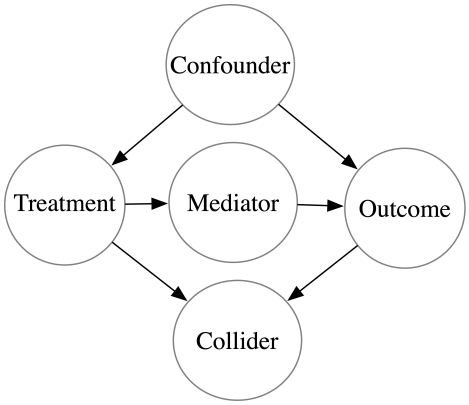



Figure 4.3: A causal diagram showing common causal relationships.

treatment-collider-outcome[5] Pearl [2009a] potentially introducing more bias [Montgomery et al., 2018, Elwert and Winship, 2014]. Mediator variables require a different set of adjustments than confounders to find the "natural direct effect" between treatment and outcome [VanderWeele, 2015, Pearl, 2014]. A practitioner typically draws a causal graph by explicitly encoding theoretical and domain assumptions as well as the results of prior data analyses.[6]

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

## 4.4 Measuring confounders via text

After drawing the causal graph, the next step is to use available text data to recover latent confounders. Some approaches *pre-specify* the confounders of interest

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

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

and measure them from text, $P(z \mid x)$. Others learn confounders *inductively* and use a low-dimensional representation of text as the confounding variable $z$ in subsequent causal adjustments.

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

***Open problems:*** Since NLP methods are still far from perfectly accurate, how can one mitigate error that arises from *approximating* confounding variables? Closely related to this question is *effect restoration* which addresses error from using proxy variables (e.g., a father's occupation) in place of true confounders (e.g, socioeconomic status) [Kuroki and Pearl, 2014, Oktay et al., 2019]. Wood-Doughty et al. [2018] build upon effect restoration for causal inference with text classifiers, but there are still open problems in accounting for error arising from other text representations and issues of calibration [Nguyen and OConnor, 2015] and prevalence estimation [Card and Smith, 2018, Keith and O'Connor, 2018] in conjunction with NLP. Ideas from

the large literature on measurement error models may also be helpful [Fuller, 1987, Carroll et al., 2006, Buonaccorsi, 2010].

**Inductively derived confounders.** Other researchers *inductively* learn confounders in order to condition on *all* aspects of text, known and unknown. For example, some applications condition on the entirety of news [Johansson et al., 2016] or scientific articles [Veitch et al., 2020, Roberts et al., 2020]. This approach typically summarizes textual information with text representations common in NLP. Ideally, this would encode all aspects of language (meaning, topic, style, affect, etc.), though this is an extremely difficult, open NLP problem. Typical approaches include the following. (1) *Bag-of-words* representations discard word order and use word counts as representations. (2) *Topic models* are generative probabilistic models that learn latent topics in document collections and represent documents as distributions over topics [Blei et al., 2003, Boyd-Graber et al., 2014, Roberts et al., 2014]. (3) *Embeddings* are continuous, vector-based representations of text. To create vector representations of longer texts, off-the-shelf word embeddings such as *word2vec* [Mikolov et al., 2013] or *GloVe* [Pennington et al., 2014] or combined via variants of weighted averaging [Arora et al., 2017] or neural models [Iyyer et al., 2015, Bojanowski et al., 2017, Yang et al., 2016]. (4) Recently, fine-tuned, large-scale neural language models such as BERT [Devlin et al., 2019] have achieved state-of-the-art performance on semantic benchmarks, and are now used as text representations. Each of these text representations is a real-valued vector that is used in place of the confounder, $z$, in a causal adjustment method (§4.5)

***Open problems:*** Estimates of causal effects are contingent on the "garden of forking paths of data analysis, meaning any "paths an analyst did not take could have resulted in different conclusions [Gelman and Loken, 2013]. For settings with causal confounders from text, the first fork is the choice of representation (e.g., topic

models or embeddings) and the second fork is the pre-processing and hyperparameter decisions for the chosen representations.

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

## 4.5 Adjusting for confounding bias

Given a set of variables $Z$ that satisfy the backdoor criterion (§4.3.2), one can use the *backdoor adjustment* to estimate the causal quantity of interest,

$$\begin{aligned} &P(Y = y \mid do(T = t)) \;= \\ &\int P(Y = y \mid T = t, Z = z)\; P(Z = z)\; dz \end{aligned} \tag{4.3}$$

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

***Open problems:*** Causal studies typically make an assumption of *overlap*, also known as *common support* or *positivity*, meaning that any individual has a non-zero probability of assignment to each treatment condition for all possible values of the covariates: $\forall z, \quad 0 < P(T = 1 \mid Z = z) < 1$. D'Amour et al. [2021] show that as

the dimensionality of covariates grows, strict overlap converges to zero. What are the implications of these results for high-dimensional text data?

### 4.5.1 Propensity scores

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

$$\pi \equiv P(T = 1 \mid Z). \tag{4.4}$$

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

$$w_i = t_i/\hat{\pi}_i + (1 - t_i)/(1 - \hat{\pi}_i), \tag{4.5}$$

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

$$\hat{\tau}_{\text{IPTW}} = \frac{1}{n_1} \sum_{i:t_i=1} w_i y_i - \frac{1}{n_0} \sum_{j:t_j=0} w_j y_j \tag{4.6}$$

### 4.5.2 Matching and stratification

*Matching* aims to create treatment and control groups with similar confounder assignments; for example, grouping units by observed variables (e.g., age, gender, occu-

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

pation), then estimating effect size within each stratum [Stuart, 2010]. *Exact matching* on confounders is ideal but nearly impossible to obtain with high-dimensional confounders, including those from text. A framework for matching with text data is described by Mozer et al. [2020] and requires choosing: (1) a text representation (§4.4); (2) a distance metric (cosine, Eucliean, absolute difference in propensity score etc.); and (3) a matching algorithm. As Stuart [2010] describes, the matching algorithm involves additional decisions about (a) greedy vs. optimal matching; (b) number of control items per treatment item; (c) using calipers (thresholds of maximum distance); and (d) matching with or without replacement. *Coarsened exact matching (CEM)* matches on discretized raw values of the observed confounders [Iacus et al., 2012].

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

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

$$\hat{y}_i(k) = \begin{cases} y_i & \text{if } t_i = k \\ \frac{1}{|\mathcal{M}_i|} \sum_{j \in \mathcal{M}_i} y_j & \text{if } t_i \neq k \end{cases} \tag{4.7}$$

which is plugged into the matching estimator,[8]

$$\hat{\tau}_{\text{match}} = \frac{1}{n} \sum_{i}^{n} \Big( \hat{y}_i(1) - \hat{y}_i(0) \Big). \tag{4.8}$$

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

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

### 4.5.3 Regression adjustment

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

$$q(t, z) \equiv \mathbb{E}(Y \mid T = t, Z = z) \tag{4.9}$$

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

$$\hat{\tau}_{\text{reg}} = \frac{1}{n} \sum_{i}^{n} (\hat{q}(1, z_i) - \hat{q}(0, z_i)) \tag{4.10}$$

### 4.5.4 Doubly-robust methods

Unlike methods that model only treatment (IPTW) or only outcome (regression adjustment), doubly robust methods model both treatment and outcome, and have the desirable property that if either the treatment or outcome models are unbiased then the effect estimate will be unbiased as well. These methods often perform very well in practice [Dorie et al., 2019]. *Adjusted inverse probability of treatment weighting (A-IPTW)* combines estimated propensity scores (Eqn. 4.4) and conditional outcomes (Eqn. 4.9), while the more general *targeted maximum likelihood estimator (TMLE)*

updates the conditional outcome estimate with a regression on the propensity weights (Eqn. 4.5) and $\hat{q}$ [Van der Laan and Rose, 2011].

### 4.5.5 Causal-driven representation learning

Several research efforts design representations of text specifically for causal inference goals. These approaches still initialize their models with representations of text described in Section 4.4, but then the representations are updated with machine learning architectures that incorporate the observed treatment assignment and other causal information. Johansson et al. [2016] design a network with a multi-task objective that aims for low prediction error for the conditional outcome estimates, $q$, and minimizes the discrepancy distance between $q(1, z_i)$ and $q(0, z_i)$ in order achieve balance in the confounders. Roberts et al. [2020] combine structural topic models (STM; [Roberts et al., 2014]), propensity scores, and matching. They use the observed treatment assignment as the content covariate in the STM, append an estimated propensity score to the topic-proportion vector for each document, and then perform coarsened exact matching on that vector. Veitch et al. [2020] fine-tune a pretrained BERT network with a multi-task loss objective that estimates (a) the original masked language-modeling objective of BERT, (b) propensity scores, and (c) conditional outcomes for both treatment and control. They use the predicted conditional outcomes and propensity scores in regression adjustment and the TMLE formulas.

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

## 4.6 Human evaluation of intermediate steps

Text data has the advantage of being *interpretable*—matched pairs and some low-dimensional representations of text can be read by humans to evaluate their quality.

When possible, we suggest practitioners use (1) interpretable balance metrics and/or (2) human judgements of treatment propensity to evaluate intermediate steps of the causal estimation pipeline.

### 4.6.1 Interpretable balance metrics

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

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

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

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

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

### 4.6.2 Judgements of treatment propensity

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

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

## 4.7 Evaluation of causal methods

Because the true causal effects in real-world causal inference are typically unknown, causal *evaluation* is a difficult and open research question. As algorithmic complexity grows, the expected performance of causal methods can be difficult to estimate theoretically [Jensen, 2019]. Other causal evaluations involve *synthetic data*. However, as Gentzel et al. [2019] discuss, synthetic data has no "unknown unknowns" and many researcher degrees of freedom, which limits their effectiveness. Thus, we encourage researchers to evaluate with *constructed observational studies* or *semi-synthetic datasets*, although measuring latent confounders from text increases the difficulty of creating realistic datasets that can be used for empirical evaluation of causal methods.

### 4.7.1 Constructed observational studies

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

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

### 4.7.2 Semi-synthetic datasets

Semi-synthetic datasets use real covariates and synthetically generate treatment and outcome, as in the 2016 Atlantic Causal Inference Competition [Dorie et al., 2019]. Several applications in this review use real metadata or latent aspects of text to simulate treatment and outcome: Johansson et al. [2016] simulate treatment

and outcome from two centroids in topic model space from newswire text; Veitch et al. [2020] use indicators of an article's "buzzy" keywords; Roberts et al. [2020] use "quantitative methodology" categories of articles that were hand-coded by other researchers.

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

## 4.8 Discussion and conclusion

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

# CHAPTER 5

# CAUSAL RESEARCH DESIGN FOR EFFECTS OF DIFFERENTIAL TREATMENT OF SOCIAL GROUPS VIA LANGUAGE MEDIATORS

## 5.1 Introduction

Interactions between individuals are key components of social structure [Hinde, 1976]. While we rarely have access to individuals' internal thoughts during these interactions, we often can observe the language they use. Using observed language to better understand interpersonal interactions is important in high-stakes decision making—for instance, judges' decisions within the United States legal system [Danescu-Niculescu-Mizil et al., 2012] or police interaction with citizens during traffic stops [Voigt et al., 2017].

Important decision makers sometimes treat some social groups (e.g. women, racial minorities, or ideological communities) differently than others [Gleason, 2020]. Yet, quantitative analyses of this problem often do not account for all possible mechanisms that could induce this differential treatment. For instance, one might ask, *Is a U.S. Supreme Court justice interrupting female advocates more because they are female, because of the advocates' language content, or because of the advocate's language delivery?* (Fig. 5.1B). Accounting for these language mechanisms could help separate the remaining "gender bias" of justices.

We reformulate the previous question as a general *counterfactual* [Pearl, 2009b, Morgan and Winship, 2015] about two speakers: *How would Speaker 2 respond if the signal they received of Speaker 1's social group flipped from A to B but Speaker 1 still used language typical of social group A?* Here, our question is about the direct

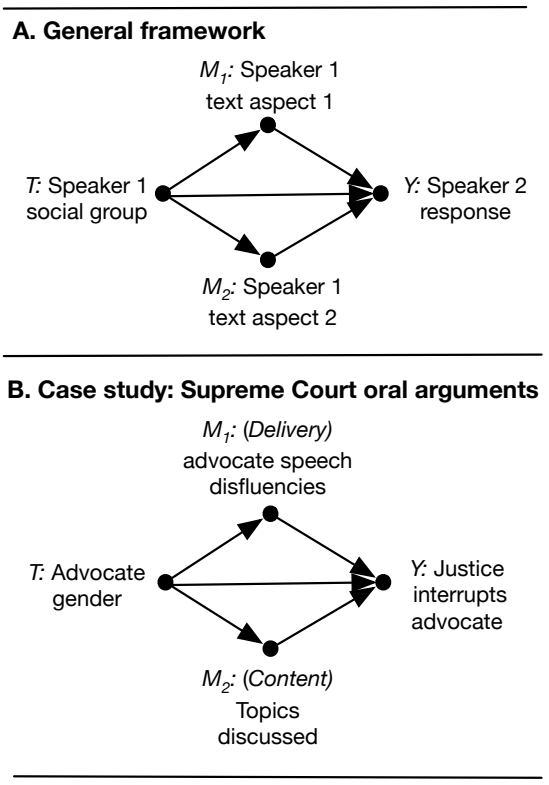



Figure 5.1: Causal diagrams in which nodes are random variables and arrows denote causal dependence for **A.** proposed general framework for *differential treatment of social groups via language aspects* and **B.** instantiation of the framework for a case study of Supreme Court oral arguments. In both diagrams, $T$ is the treatment variable, $Y$ is the outcome variable, and $M$ are mediator variables. This is a simplified schema; see Fig. 5.2 for an expanded diagram.

causal effect of *treatment*—Speaker 1's signalled social group—on *outcome*—Speaker 2's response—that is not through the causal pathway of the *mediator*—an aspect of language (Fig. 5.1A).

The fundamental problem with this and any counterfactual question is that we cannot go back in time and observe an individual counterfactual while holding all other conditions the same [Holland, 1986]. Furthermore, in many high-stakes, real-world settings (e.g. the U.S. Supreme Court), we cannot run experiments to randomly assign treatment and approximate these counterfactuals. Instead, in these settings, causal estimation must rely on *observational* (non-experimental) data.

In this work, we focus on this observational setting and build from causal mediation methods [Pearl, 2001, Imai et al., 2010, VanderWeele, 2016] to specify a research design of causal estimates of *differential treatment of social groups via language aspects*.[1] We address critiques of the design in §5.4.2 and §5.5 including: flaws in using social groups as a causal treatment, dependence between mediators in conversations, and dependence between perception of social groups and linguistic perception.

Overall, we make the following contributions:

- We propose a new causal research design for estimating the natural indirect and direct effects of social group signal on a conversational outcome with separate aspects of language as causal mediators (§5.3).
- We illustrate the promises and challenges of this framework via a theoretical case study of the effect of an advocate's gender on interruptions from justices in Supreme Court oral arguments. (§5.2).
- We discuss challenges researchers might face conceptualizing and operationalizaing the causal variables in this research design (§5.4).
- We directly address critiques of using social groups (e.g. race or gender) as treatment and construct gender and language as *constitutive* variables, building from Sen and Wasow [2016], Hu and Kohler-Hausmann [2020].
- We articulate potential open challenges in this research design including temporal dependence between mediators in conversations, causal dependence between multiple language mediators, and dependence between social group perception and language perception (§5.5).

[1]Other work has used mediation analysis to understand NLP components [Vig et al., 2020, Finlayson et al., 2021]; however, this work is more closely aligned with recent work examining the role of text in causal estimates [Veitch et al., 2020, Roberts et al., 2020, Keith et al., 2020a, Zhang et al., 2020, Pryzant et al., 2021].

| |
|---|
| ***(A) Case: Kennedy v. Plan Administrator for DuPont Sav. and Investment Plan (2008-07-636)*** |
| **Mark Irving Levy:** [...] The QDRO provision is an objective checklist that is easy for – for plan administrators to follow. |
| **Antonin Scalia:** What if they had agreed to the waiver apart from [...] We'd be in the same suit that you're - - that you say we have to avoid, wouldn't we? |
| **Mark Irving Levy:** I don't think so. I mean I think that would be an alienation. |
| **Antonin Scalia:** Well, if it's an alienation, but his point is that a waiver is not an alienation. |
| ***(B) Case: Lozano v. Montoya Alvarez (2013-12-820)*** |
| **Ann O'Connell Adams:** Well - - |
| **Antonin Scalia:** I mean, it seems to me it just makes that article impossible to apply consistently country to country. |
| **Ann O'Connell Adams:** - - No, I don't think so. And - - and, the other signatories have - - have almost all, I mean I think the Hong Kong court does say that it doesn't have discretion, but it said in that case nevertheless it would, even if it had discretion, it wouldn't order the children returned. But the other courts of signatory countries that have interpreted Article 12 have all found a discretion, whether it be in Article 12 or in Article 8. And if I - - |
| **Antonin Scalia:** Have they exercised it? Have they exercised it, that discretion which they say is there? |

Table 5.1: Selected utterances from the oral arguments of two Supreme Court cases, A [Oyez, a] and B [Oyez, b], with advocates Mark Irving Levy (male) and Ann O'Connell Adams (female) respectively. Justice Antonin Scalia responds to both advocates. Hedging language is highlighted in blue. Speech disfluencies are highlighted in red. Gray-colored utterances directly proceed the target utterances (non-gray colored) in the oral arguments.

## 5.2 Theoretical case study of gender bias in U.S. Supreme Court interruptions

To motivate our causal research design and illustrate challenges that arise with it, we focus on a specific theoretical case study—the effect of advocate gender on justice interruptions via advocates' language in Supreme Court oral arguments (Fig. 5.1B). Previous work found female lawyers are interrupted earlier in oral arguments, allowed to speak for less time, and subjected to longer speeches by justices [Patton and Smith, 2017], and justices are more likely to vote for a female advocate's party when the female advocate uses emotional language [Gleason, 2020].

**Counterfactual questions.** We present a novel causal approach to understanding gender bias in Supreme Court oral arguments that corresponds to the following counterfactual questions:

1. *(NDE)*: How would a justice's interruptions of an advocate change if the signal of the advocate's gender the justice received flipped from male to female, but the advocate still used language typical of a male advocate?

2. *(NIE)*: How would a justice's interruptions of an advocate change if a male advocate used language typical of a female advocate but the signal of the advocate's gender the justice received remained male?

which we show correspond to the *natural direct effect* (NDE) and *natural indirect effect* (NIE) respectively in §5.3. In §5.4, we walk through the theoretical conceptualization and empirical operationalization of advocate gender (treatment), interruption (outcome), and advocate language (mediators).

**Intuitive example.** We describe intuitive challenges of our causal research design with the example in Table 5.1. In Example A [Oyez, a], Levy—a male advocate—is not interrupted by Justice Antonin Scalia but in Example B [Oyez, b], Adams—a female advocate—is interrupted. *Why was the female advocate interrupted? Was it because of her gender or because of* what *she said or* how *she said it*? We hypothesize one causal pathway between gender and interruption is through the mediating variable hedging—expressions of deference or politeness.[2] Suppose we operationalize hedging as certain key phrases, e.g. "I don't think so" and "I mean I think." An initial causal design might assign a binary hedging indicator to utterances and then compare average interruption outcomes for male and female advocates conditional on the hedging indicator.

[2]Previous work has shown hedging is used more often by women [Lakoff, 1973, Poos and Simpson, 2002], and we hypothesize judges might respond more positively to more authoritative language (less hedging) from advocates.

However, advocate utterances matched on this hedging indicator could have a number of latent mediators and confounders. In Table 5.1, Adams has speech disfluencies ("and - - and" and "have - - have" shown in red) which might cause Scalia to get frustrated and interrupt. The cases are from different areas of the law[3] and Scalia may interrupt more for case issue areas he cares more about. The advocate utterance in Ex. B is longer (more tokens) and longer utterances may be more likely to be interrupted. In Ex. B, Scalia interrupts Adams just prior to the target utterance which possibly indicates a more "heated" portion of the oral arguments during which interruptions occur more on average. With these confounding and additional mediator challenges, a simple causal matching approach (e.g. Stuart [2010], Roberts et al. [2020]) is unlikely to work and we advocate for the causal estimation strategy presented in §5.3.3. We move from this case study to a formalization of our causal research design in §5.3.

## 5.3 Causal mediation formalization, identification, and estimation

Many causal questions involve *mediators*—variables on a causal path between treatment and outcome. For example, what is the effect of gender[4] (treatment) on salary (outcome) with and without considering merit (a mediator)? If one intervenes on treatment, then one would activate both the "direct path" from gender to salary *and* the "indirect path" from gender through merit to salary. Thus, a major focus of causal mediation is specifying conditions under which one can separate estimates of the *direct effect* from the *indirect effect*—the former being the effect of treatment on outcome *not* through mediators and the later the effect through mediators.

---

[3]The Supreme Court Database codes Ex. A as "economic activity" and Ex. B as "civil rights" [Spaeth et al., 2021].

[4]See §5.4 for discussion of operationalizing difficult causal variables such as gender.

We use this causal mediation approach to formally define our framework. For each unit of analysis (see §5.4.1), $i$, let $T_i$ represent the treatment variable—the social group, e.g. gender of an advocate—and $Y_i$ represent the outcome variable—the second speaker's response, e.g. a judge's interruption or non-interruption of an advocate. For each defined mediator $j$, let $M_i^j$ represent the mediating variable—an aspect of language, e.g. an advocate's speech disfluencies or the topics of an utterance. Let $X_i$ represent any other confounders between any combination of the other variables.

Because causal mediation consists of inquiries about counterfactual *paths* and not interventions of *variables*,[5] we use the potential outcomes framework [Rubin, 1974] to define the effects of interest. Let $M_i(t)$ represent the (counterfactual) potential value the mediator would take if $T_i = t$. Then $Y_i(t, M_i(t'))$ is a doubly-nested counterfactual that represents the potential outcome that results from both $T_i = t$ and potential value of the mediator variable with $T_i = t'$. With this formal notation, we define the individual *natural direct effect (NDE)* and *natural indirect effect (NIE)*:[6]

$$\text{NDE}_i = Y_i(1, M_i(0)) - Y_i(0, M_i(0)) \tag{5.1}$$

$$\text{NIE}_i = Y_i(0, M_i(1)) - Y_i(0, M_i(0)) \tag{5.2}$$

These correspond to the two counterfactual questions from §5.2 if $T_i = 0$ and $T_i = 1$ represent the gender signal of the advocate being male and female respectively.

---

[5] In the words of Pearl [2001], a mediation research question "cannot be represented in the standard syntax of $do(x)$ operators—it does not involve fixing any of the variables in the model but, rather, modifying the causal paths in the model."

[6] We note Pearl et al. [2016] defines the NDE and NIE in terms of the non-treatment condition, $T = 0$. Others (e.g. Imai et al. [2010] and Van der Laan and Rose [2011]) give alternate definitions of these quantities in terms of $T = 1$. We follow Pearl et al.'s definitions in the remainder of this work.

### 5.3.1 Interpretation of the NDE as "bias"

Many applications of causal mediation aim to quantify "implicit bias" or "discrimination" via the natural direct effect. However, if all relevant mediators are not accounted for, one cannot interpret the estimand of the natural direct effect as the actual direct causal effect [Van der Laan and Rose, 2011, p.135]. Nevertheless, separating the total effect into the proportion that is the NDE and the NIE with the mediators we can measure moves our analysis *closer* to estimating the true direct effect between treatment and outcome. Thus, in this work we emphasize the value of having interpretable mediators (i.e. language aspects) for which the NIE is a meaningful quantity to analyze in itself.

### 5.3.2 Identification

Like any causal inference problem, we first examine the *identification assumptions* necessary to claim an estimate as causal. The key assumption particular to causal mediation is that of *sequential ignorability* [Imai et al., 2010]:

1. Potential outcomes and mediators are independent of treatment given confounders

$$\{Y_i(t', m), M_i(t)\} \perp\!\!\!\perp T_i \mid X_i = x \tag{5.3}$$

2. Potential outcomes are independent of mediators given treatment and confounders

$$Y_i(t', m) \perp\!\!\!\perp M_i(t) \mid \{T_i = t, X_i = x\} \tag{5.4}$$

for $t, t' \in \{0, 1\}$ and all values of $x$ and $m$.

*Mediator Independence Assumption:*[7] For our particular framework, we make an additional assumption that for each language aspect we study, the mediators are

[7]This is similar to the assumptions Pryzant et al. [2021] make for linguistic properties of text as treatment.

independent conditional on the treatment and confounders

$$\forall j, j' : \; M_i^j(t) \perp\!\!\!\perp M_i^{j'}(t) \mid \{T_i = t, X_i = x\} \tag{5.5}$$

With this assumption, we can estimate the NIE and NDE of each mediator successively, ignoring the existence of other mediators. [Imai et al., 2010, Tingley et al., 2014]. We discuss the validity of this assumption in §5.5.

### 5.3.3 Estimation

Given the satisfaction of sequential ignorability, mediator independence, and other standard causal identification assumptions,[8] we propose using the following estimators of population-level natural direct and indirect effects for each mediator $j$ [Imai et al., 2010, Pearl et al., 2016]:

$$\begin{aligned} \text{SA-NDE}^j = \frac{1}{N} \sum_{i=1}^{N} \sum_{x \in \mathcal{X}} \sum_{m \in \mathcal{M}^j} \Big( & \hat{f}^j(Y | M_i^j = m, T_i = 1, X_i = x) \\ & - \hat{f}^j(Y | M_i^j = m, T_i = 0, X_i = x) \Big) \hat{g}^j(m | T_i = 0, X_i = x) \end{aligned} \tag{5.6}$$

$$\begin{aligned} \text{SA-NIE}^j = \frac{1}{N} \sum_{i=1}^{N} \sum_{x \in \mathcal{X}} \sum_{m \in \mathcal{M}^j} & \hat{f}^j(Y | M_i^j = m, T_i = 0, X_i = x) \\ & \Big( \hat{g}^j(m | T_i = 1, X_i = x) - \hat{g}^j(m | T_i = 0, X_i = x) \Big) \end{aligned} \tag{5.7}$$

Each is a **S**ample **A**verage estimate from $N$ data points, relying on models trained to predict mediator and outcome given confounders and treatment: $\hat{g}^j$ infers mediator $j$'s probability distribution, while $\hat{f}^j$ infers the expected outcome conditional on mediator $j$. The estimators marginalize over confounder and mediator from their respective domains ($x \in \mathcal{X}$, $m \in \mathcal{M}^j$), which for our discrete variables is feasible with explicit sums (see Imai et al. for the continuous case).

[8] Overlap, SUTVA etc.; see Morgan and Winship [2015].

**Model fitting.** When fitting models $\hat{f}$ and $\hat{g}$, we highly recommend using a cross-sample or cross-validation approach in which one part of the sample is used for training/estimation ($S_{\text{train}}$) and the other is used for testing/inference ($S_{\text{test}}$) in order to avoid overfitting [Chernozhukov et al., 2017, Egami et al., 2018]. With text, one must also fit a model for the mediators conditional on text, $h(m|\text{text})$ using $S_{\text{train}}$. In some cases, such as measuring advocate speech disfluencies, $h$ may be a simple deterministic function. However, when using NLP and other probabilistic models (e.g topic models or embeddings), $h$ could be a difficult function to fit and have a certain amount of measurement error. A major open question is whether to jointly fit $h$ and $g$ at training time as advocated by previous work [Veitch et al., 2020, Roberts et al., 2020] or if $h$ and $g$ should be treated as separate modules. At inference time, we do not use the inference text from $S_{\text{test}}$ since Eqns. 5.6 and 5.7 only rely on the mediators through estimates from $\hat{g}$.

## 5.4 Conceptualization and operationalization of causal variables

For any causal research design—and particularly those in the social sciences—there are often challenges *conceptualizing* the theoretical causal variables of interest. Even after these theoretical concepts are made concrete, there are often multiple ways to *operationalize* these concepts. We discuss conceptual and operational issues for our both our general research design and case study. In particular, we recommend researchers formalize variables such as gender and language as *constitutive* variables made of multiple components, building from Sen and Wasow [2016], Hu and Kohler-Hausmann [2020] (e.g. see Fig. 5.2).

### 5.4.1 Unit of analysis

As with most causal research designs, one starts by conceptualizing the *unit of analysis*—the smallest unit about which one wants to make counterfactual inquiries. In our framework, the *unit of analysis* is language ($L$) between speakers of two categories: the first ($P_1$) being a social group of interest (e.g. advocates) for which treatment values (e.g. female and male) will be assigned; and the second ($P_2$) being the set of decision-makers responding to the first speaker (e.g. judges).

**Operationalizations.** There are several possible operationalizations of $L$: pairs of single utterances—whenever a $P_1$ speaks and a $P_2$ responds; a thread of several utterances between a $P_1$ and a $P_2$ within a conversation; or the entire conversation between a $P_1$ and a $P_2$. In §5.5, we note that selecting the unit of language could have implications for modeling temporal dependence between mediators.

### 5.4.2 Treatment

At the most basic level, *treatment*, $T$, in our research design is *the social group* of $P_1$ (Fig. 5.1). However, inspired by the *causal consistency* arguments from Hernán [2016],[9] we examine several competing versions of treatment for our theoretical case study and explain the reasons we eventually choose version #5 (in bold):

1. Do judges interrupt at different rates based on an advocate's *gender*?
2. Based on an advocate's *biological sex assigned at birth*?
3. An advocate's *perceived gender*?
4. An advocate's *gender signal*?

---

[9] *Consistency* is the condition that for observed outcome $Y$ and treatment $T$, the potential outcome equals the observed outcome, $Y(t) = Y$ for each individual with $T = t$. Hernán [2016] presents eight versions of treatment for the causal question "Does water kill?" to illustrate the deceptiveness of this apparently simple consistency condition. Hernán points out that "declaring a version of treatment sufficiently well-defined is a matter of agreement among experts based on the available substantive knowledge" and is inherently (and frustratingly) subjective.

5. **An advocate's *gender signal* as defined by (hypothetical) manipulations of the advocate's clothes, hair, name, and voice pitch?**
6. An advocate's *gender signal* by (hypothetical) manipulations of their entire physical appearance, facial features, name, and voice pitch?
7. An advocate's *gender signal* by setting their physical appearance, facial features, name, and voice pitch to specific values (e.g. all facial features set to that of the same 40-year-old, white female and clothes set to a black blazer and pants).

In critique of treatment version #1, most social groups (e.g. gender or race) reflect highly contextual social constructs [Sen and Wasow, 2016, Kohler-Hausmann, 2018, Hanna et al., 2020]. For gender in particular, researchers have shown social, institutional, and cultural forces shape gender and gender perceptions [Deaux, 1985, West and Zimmerman, 1987], and thus viewing gender as a binary "treatment" in which individuals can be randomly assigned is methodologically flawed. In critique of version #2, *biological sex assigned at birth* is a characteristic that is not manipulable by researchers and the "at birth" timing of treatment assignment means all other variables about the individual are post-treatment. Thus, researchers have warned against estimating the causal effects of these "immutable characteristics" [Berk et al., 2005, Holland, 2008].

Greiner and Rubin [2011] propose overcoming the issues in versions #1 and #2 by shifting the unit of analysis to the *perceived gender* of the decision-maker (#3) and defining treatment assignment as the moment the decision-maker first perceives the social group of the other individual. Hu and Kohler-Hausmann [2020] critique this *perceived gender* variable and emphasize that we, as researchers, cannot actually change the internal, psychological state of the decision-makers, but rather we can change the *signal* about race or gender those decision-makers receive (#4). However, as Sen and Wasow [2016] discuss, defining treatment as the *gender signal* (#4) is dismissive of the many components that make up a social construct like gender.

Instead, Sen and Wasow recommend articulating the specific variables one would potentially manipulate. For *gender* in our case study, this could mean hypothetical manipulations of an advocate's dress, name, and voice pitch (#5).

Shifting from versions #5 to #6 and #7, we define treatment in terms of more specific manipulations. However, we also enter the realm of Hernán's argument that precisely defining the treatment never ends, and some aspects of #6 and #7 are impossible to manipulate in the real-world setting of the U.S. Supreme Court. What does it mean to manipulate an advocate's "entire physical appearance?"[10] When we define treatment very specifically—e.g. using the same 40-year old white woman as the treatment for "female advocate" (#7)—are we estimating a causal effect of gender *in general*? Thus, we back-off from versions #6 and #7, and advocate using #5 as our definition of treatment.

**Constitutive causal diagrams.** With these considerations, drawing a causal diagram in which a *gender* is represented as a single node seems methodologically flawed. Instead, building from Sen and Wasow [2016], Hu and Kohler-Hausmann [2020], we represent treatment (the social group) as cloud of components (a *constitutive* variable), some of which are latent, some observable, and some manipulable. In Fig. 5.2, we shade the "outward" components of *gender*—hair, appearance, clothes, voice pitch, and name—that are our hypothetical manipulations and would influence the latent variable of a judge's perceived gender of the advocate. Other "background" components of gender—gender norms, education, and socialization—are the components that causally influence language.

[10]Would justices have to interact with advocates through a computer-mediated system in which one could customize avatars of the advocates? We note, using computer-mediated avatars to signal social group identity has been used effectively in other causal studies, e.g. Munger [2017].

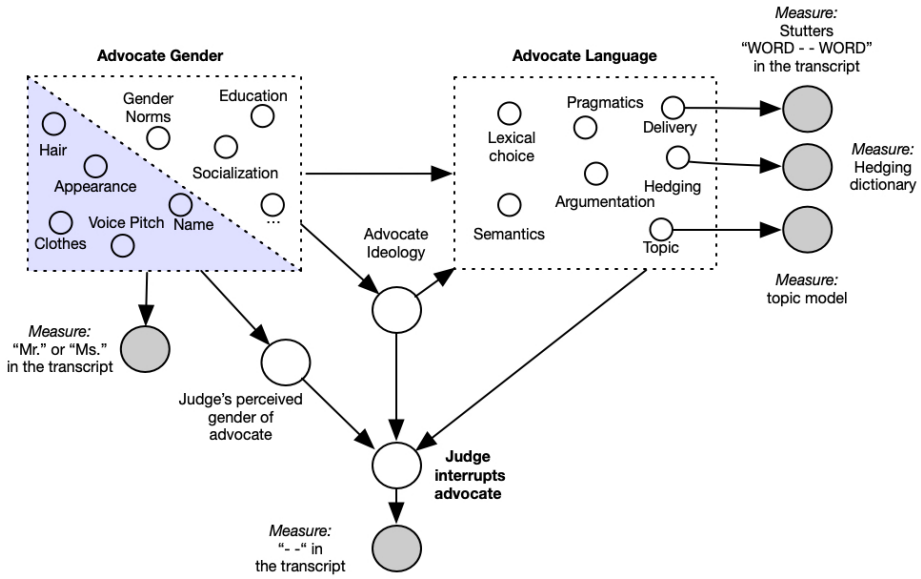



Figure 5.2: *Constitutive* causal diagram for gendered interruption in Supreme Court oral arguments. Latent theoretical concepts are unshaded circles and observed measurements are shaded circles. The causal variables *gender* and *language* are represented as dashed lines around their constituent parts. The shaded portion of *gender* consists of the gender variables that one could manipulate in a hypothetical intervention.

**Case study operationalizations.** Even after selecting version #5 as our conceptualization of treatment, there are still multiple operationalizations for our theoretical case study:

**Treatment operationalization 1:** Previous work operationalizes gender in Supreme Court oral arguments by using norm that the Chief Justice introduces an advocate as "Ms." and "Mr." before their first speaking turn [Patton and Smith, 2017, Gleason, 2020]. The advantage of this operationalization is that it is simple, clean, and consistent, and occurs direct before an advocate's first utterance.[11]

[11]The treatment assignment timing is potentially important for the rest of the causal diagram. If we can define *gender signal* and thus latent *perceived gender* as happening right before an advocate first speaks, and then is not adapted or updated by the judge over the course of the oral arguments and we can eliminate the causal arrow between variables "language" and "perceived gender."

**Treatment operationalization 2:** Alternatively, one could focus on even more specific components of gender for (hypothetical) manipulations. For instance, Chen et al. [2016] and Chen et al. [2019] measure voice pitch when studying gender on the U.S. Supreme Court. While being more cumbersome to measure, this operationalizes gender as a real-valued (instead of binary) variable and thus potentially measures more subtle gender biases.

### 5.4.3 Outcome

In our general framework, we define the *outcome*, $Y$, as *the response of the second speaker* (Fig. 5.1A), and we intentionally leave this variable vague and domain-specific. However, if making the leap from *differential treatment* to claiming *discrimination* or *bias*, conceptualizing a causal outcome requires normative commitments and a moral theory of what is harmful [Kohler-Hausmann, 2018, Blodgett et al., 2020]. In our case study, we conceptualize the outcome variable as a judge interrupting an advocate. This outcome is of substantive interest because, in general, interruptions can indicate and reinforce status in conversation [Mendelberg et al., 2014], and, specifically to the U.S. Supreme Court, there is interest in connecting justice's behavior in oral arguments to case outcomes.

**Outcome operationalization 1:** Previous work uses the transcription norm of a double-dash ("- -") at the end of a advocate utterance when a justice interrupts in the next utterance [Patton and Smith, 2017]. However, the validity of this operationalization relies on consistent transcription standards.

**Outcome operationalization 2:** An alternative operationalization could classify interruptions into positive (agreeing with the first speaker's comment), negative (disagreeing, raising an objection, or completely changing the topic), or neutral [Stromer-Galley, 2007, Mendelberg et al., 2014]. While estimating the effects of only negative interruptions could further refine the causal question—*Do justices nega-*

*tively interrupt female advocates more?*—this operationalization could also introduce measurement error since it could prove difficult difficult to design an accurate NLP classifier for this task.

### 5.4.4 Language mediators

Our framework explicit focuses on *language as a mediator* in differential treatment of social groups. Yet, language consists of multiple levels of linguistic structure [Bender, 2013, Bender and Lascarides, 2019], so as with social groups (§5.4.2) it is a variable that is non-modular and should be represented as constituent parts (Fig. 5.2).

**Mediator Operationalizations:** We focus on three potential language aspects for our Supreme Court case study: (A) *hedging*—expressions of deference or politeness—with an operationalization as lexical matches from a single-word hedging dictionary (e.g. Prokofieva and Hirschberg [2014]); (B) *speech disfluencies*—repetitions of syllables, words, or phrases—which we operationalize as the transcript noting a repeated unigram with a double dash, "*word - - word*"; and (C) semantic *topics* operationalized as a topic model [Blei et al., 2003] applied to utterances.

**Recommendations.** We discuss the choice of these particular language aspects, $M^j$, for our case study as well as general recommendations for researchers operationalizing language as a mediator.

Is $M^j$ *interpretable*? Is there a *hypothetical manipulation*[12] of $M^j$? In contrast to prior work that treats language as a black-box in causal mediation estimates [Veitch et al., 2020], we advocate for using interpretable aspects of language so that the NIE is meaningful.

[12]To be precise, the *controlled direct effect* is the estimand in which the mediator is manipulated, $do(M)$ [Pearl, 2001]. In contrast, the *natural* direct and indirect effects are counterfactuals on paths. However, we still find thinking through potential manipulations is helpful in refining the conceptualization of a language aspect.

Is there *substantive theory* for causal pathways $T \rightarrow M^j$ and from $M^j \rightarrow Y$? Without such theory, studying certain aspects of language is not meaningful. See §5.2 for our theoretical reasoning through the causal connections between gender, hedging, and interruption.

To what extent does one expect *measurement error* of $M^j$ when using automatic NLP tools? Our operationaliztions of hedging lexicons and speech disfluences are deterministic; however, topic model inferences are probabilistic and sensitive to changes in hyperparameters and pre-processing decisions [Schofield et al., 2017, Denny and Spirling, 2018], and these kinds of measurement errors are still open questions (although there is work that examines measurement error when text is treatment [Wood-Doughty et al., 2018]).

Is $M^j$ *causally independent* from other measured language aspects, $M^{j'}$? If not, our proposed estimator from §5.3.3 is invalid. Thus, one must scrutinise which aspects of language are separable and thus able to be included in the causal analysis—e.g. we could include content (topics) versus delivery (speech disfluencies) since one could hypothetically modify one without affecting the other. We discuss this assumption further in §5.5.

### 5.4.5 Non-language mediators

Returning to §5.3.1, there is often a tendency to interpret the NDE as something like "pure" *gender bias*—What is the effect of gender on interruption when all other possible causal pathways are stripped away? Conceptualizing and operationalizing language aspects as mediators (§5.4.4) moves the NDE towards the desired "gender bias." However, there may be other mediator pathways that explain these effects. For example, in our case-study, two additional mediators of interest are advocate ideology (e.g. liberal or conservative) and the level of "eliteness" of the advocate's law firm. A major validity issue is the *causal independence* of these mediators from the language

mediators. For instance, ideology could influence certain aspects of language (topic), and "eliteness" of the advocate's law firm could be a proxy for level of training which could influence the advocate's delivery.

## 5.5 Challenges and threats to validity

**Temporal dependence of utterances.** So far, we assumed the "units of analysis" of text are independent (§5.4.1). However, previous utterances in a conversation often influence the target utterances. For our case study, if Judge A interrupted Advocate B often in $t' < t$, interruption at $t$ is more likely (the two speakers are possibly in a "heated" part of the conversation) and Advocate B's speech disfluencies at $t$ are also more likely (the advocate could be mentally fatigued). Potential avenues forward include changing the unit of analysis to the entire conversational thread between the two target speakers or building extensions to the multiple mediator literature, i.e. Imai and Yamamoto [2013], VanderWeele and Vansteelandt [2014], Daniel et al. [2015], VanderWeele [2016].

**Dependence between multiple language mediators.** Our framework assumes one can computationally separate aspects of language.[13] However, some sociolinguists argue aspects of language such as "style" cannot be separated from "content" because style originates in the content of people's lives and different ways of speaking signal socially meaningful differences in content [Eckert, 2008, Blodgett, 2021]. If our mediator independence assumption (Eqn. 5.5) is violated, then we would have to turn to alternate estimation strategies from the multiple mediator literature to deal with this dependence.

**Dependence between social group perception and linguistic perception.** Separating the direct and indirect causal paths in our framework relies on there being

[13]This assumption is made in other NLP applications such as style transfer or machine translation [Prabhumoye et al., 2018, Li et al., 2018, Hovy et al., 2020].

a *decision-maker's latent perception of social group* variable on the direct path and this is independent from a *decision-maker's latent perception of language* variable on the indirect path. However, in sociolinguistics, "indexical inversion" considers "how language ideologies associated with social categories produce the perception of linguistic signs" [Inoue, 2006, Rosa and Flores, 2017]. Suppose Judge A perceives Advocate B as female, then Judge A might perceive Advocate B's language as more feminine even if it is linguistically identical to language used by male advocates. Furthermore, latent gender perception and latent language perception might interact in affecting the outcome through mechanisms such as rewarding "conforming to gender norms"—an advocate who is perceived as a man might get penalized for using feminine language whereas an advocate perceived as a woman might get rewarded, e.g. Gleason [2020].

## 5.6 Conclusion

In this work, we specify a causal research design for *differential treatment of social groups with language as a mediator*. We believe this research design is important for studying the direct and indirect causal effects in high-stakes decision making such as gender bias in the United States Supreme Court. Separating the indirect effect of treatment on outcome through interpretable language aspects allows us to estimate counterfactual inquiries about differential treatment when speakers use and do not use the same language. Despite open technical challenges, we remain optimistic that researchers can build upon this framework and continue to improve our understanding of differential treatment in settings of high-stakes decision making.

# CHAPTER 6

# CONCLUSION

This thesis has been motivated by real-world social science applications that use text data. In order to support social science needs, this thesis has addressed gaps between methods in natural language processing (NLP) used to automate and scale-up these quantitative studies of text and themes of measurement and causal inference. We have made progress in closing this gap via models for document class prevalence estimation that are more robust to shifts in class priors between training and inference (Chapter 2); methods for entity-event measurement with a new latent disjunction model that aggregates mention-level inferences to determine entity-level labels (Chapter 3); a review, guidelines, and open problems for using text to reduce confounding from causal estimates (Chapter 4); and a new causal research design for language as causal mediators in estimating the effects of social group signals on differential treatment (Chapter 5). While these are incremental contributions towards better "corpus-centered" NLP, there are numerous future directions in closing this gap.

## 6.1 Future work and discussion

We see fruitful future work in improved characterization of the relationship between noisy text measurements and causal estimation (§6.1.1); improved empirical evaluation of corpus-level measurement, causal evaluation, and some text-based causal inference assumptions (§6.1.2); and extensions of text measurement applications and approaches (§6.1.3).

### 6.1.1 Relationship between measurement and causal inference with text

In this thesis, we have treated our two themes—measurement and causal inference with text—as somewhat separate endeavors. However, the two are inextricably linked. Causal questions help direct which measurements are important to construct (even for purely descriptive studies), and measurement of text is the necessary component that converts raw text data into variables that can be incorporated into a causal model.

**Modular vs. joint learning of measurement and causal estimation with text.** In many cases, the accuracy of a noisy measurement component will affect the error and validity of causal estimates. With text, it is unclear in what situations one should *jointly* learn text measures and causal estimates or if measurement should be treated as a *separate module* that can be plugged into a causal estimator.

In favor of the modular approach, work on "effect restoration" adjusts causal estimates by relying on obtaining the conditional probabilities of the proxy (in our case text) given confounders that govern the error mechanism [Pearl, 2010, Kuroki and Pearl, 2014]. Wood-Doughty et al. [2018] extend this approach to measurement errors in text classifiers. Yet, it is unclear how to use this approach when text encodes multiple causal variables simultaneously (e.g. confounding and treatment). In this entangled case, the particular structure of measurement error (e.g. whether it is "independent" or "nondifferential") may be helpful in guiding which methods can be used to correct it [Hernán and Robins, 2020, Chapter 9]

In favor of the joint approach, recent work with text as a proxy for treatment or confounders finds (in semi-synthetic experiments) that joint learning of text representations and causal variables results in decreased error of causal estimates. Veitch et al. [2020] develop a method for "causally sufficient embeddings" that learn aspects of text predictive of treatment and outcome, and find jointly-learned representations have much lower causal estimate errors than representations learned without training

on treatment and outcome. Roberts et al. [2020] use a structured topic model to incorporate topic and treatment assignment for text as confounders and they find that this joint learning has improved mean squared error, bias, and coverage compared to matching on only topics or only on treatment assignment. With text as treatment, Pryzant et al. [2021] find causal estimates "can lose fidelity when then proxy is less than 80% accurate." While these preliminary experimental results are promising, future work needs better empirical evaluation (see §6.1.2) and methods to manually validate these jointly-learned measures of text.

**Sensitivity of causal estimates to text measurement decisions.** One potential direction forward is to evaluate the *sensitivity* of measurement decisions in causal estimates across a suite of real-world empirical applications. As we mention in Chapter 4, estimates of causal effects are contingent on the "garden of forking paths" of data analysis [Gelman and Loken, 2013], meaning any "paths" an analyst did not take could result in different conclusions, differences which are important to characterize for valid social science. While recent work uses synthetic or semi-synthetic data for sensitivity analysis in text-based causal inference [Wood-Doughty et al., 2021], evaluations with with real-world data may be more meaningful. For instance, Keith et al. [2020b] empirically examine an established economic index which measures "economic policy uncertainty" from keyword occurrences in news articles [Baker et al., 2016]. Keith et al. swap the measurement module from keyword-matching to a supervised classifier (which has higher F1 and accuracy on the training and test sets), and show the two measurement modules have very low correlation (0.38 Pearson's $\rho$), a concerning conclusion for the validity of the index. Future work could extend this method to other text-as-data applications to better characterize the sensitivity of measurement decisions in causal estimates.

### 6.1.2 Empirical evaluation

**Corpus-level evaluations.** As we describe in the introduction, one of the challenges of shifting from "downstream-centered" to "corpus-centered" NLP is that the later is concerned with inferences not at the individual phrase, sentence, or document level but at the *corpus*-level. Yet, corpus-level evaluation is still underaddressed in NLP and requires large amounts of labeled data. In Chapter 2, we evaluate our prevalence estimation approaches by constructing a natural prevalence estimation task—inferring the prevalence of positive sentiment from Yelp reviews for individual businesses. However, this evaluation required an extremely large collection of pseudo-labels (Yelp stars) and a large number of test groups (500 businesses) each of which consisted of 200 to 2000 reviews. Other recent work on corpus-level evaluation by this thesis author also required large amounts of labeled data; Halterman et al. [2021] annotate all 21,391 sentences for police activity from one-month of *Times of India* news reports in order to evaluate temporal trends of event counts. Both these corpus-level evaluations are extremely data-hungry, exemplifying a major barrier to corpus-level evaluation at scale. Possible future work could gather previous data annotation efforts by social scientists and build suites of corpus-level evaluation benchmarks. Other promising avenues could be distant supervision (such as the approach used in Chapter 3) using existing structured social science databases.

**Methods to examine some text-based causal assumptions.** Although most causal assumptions are untestable, there are a few that have the possibility to be assessed. For instance, Hill and Su [2013] assess the causal assumption of *overlap*—the conditional probability of treatment given confounding covariates is bounded between 0 and 1—since there is nothing to prevent some causal estimation methods from extrapolating over areas of the confounder space in which overlap does not exist; Hill and Su propose a solution that removes observations that have large standard deviations of model-inferred Bayesian posteriors. Veitch and Zaveri [2020] create

plots that help researchers qualitatively reason about how unobserved confounders could compare to observed confounders. Since overlap violations and unobserved confounding are major concerns for high-dimensional text data, interesting future avenues could explore expanding these methods to text.

**Causal evaluation with text.** As we mention in Chapter 4, evaluation of causal methods is a difficult and open problem and extending this to text is even more complicated. There could be future efforts in competition-based benchmarks of causal inference with text similar to the approach of Dorie et al. [2019]. This direction is promising given newly created venues such as the *First Workshop on Causal Inference & NLP*[1] that could support these types of competitions.

### 6.1.3 Measurement extensions

**Heterogeneous perception of text.** Although we hint at this in the introduction when discussing linguistic ambiguity (measurement challenge #3), this thesis does not address the possibility of *heterogeneous* perception of texts. Modeling this ambiguity is crucial both in incorporating uncertainty in text measurements and the importance of heterogeneous perception measurement in some causal estimation settings (e.g. latent perception of social groups and language which we mention in Chapter 5). Future efforts could build upon recent work in NLP that focuses on propagating annotator uncertainty to downstream inferences [Dumitrache et al., 2018, Paun et al., 2018, Pavlick and Kwiatkowski, 2019] or Bayesian models for rational speech acts which formalize communication as recursive reasoning between a speaker and listener [Andreas and Klein, 2016, Monroe, 2018].

**Collecting counterdata from text.** A potentially impactful application area of the methods presented in this thesis is collecting *counterdata*—ground-up collection of data that is missing or not collected by central governments or institutions

[1]To appear at EMNLP in November, 2021 https://causaltext.github.io/2021/

[Currie et al., 2016, D'Ignazio and Klein, 2020]. Chapter 3—updating a database of police fatalities—can be see as an instance of counterdata collection, and counterdata collection efforts can be beneficial to those aiming for data-driven policy changes. Expanding these text-based counterdata collection applications is a rich avenue of future work. For instance, Halterman et al. [2021] use event extraction techniques to detect police actions during riots in Gujurat, India in 2002—data not released by the Indian government. Future work could expand these initial efforts and build a broader set of NLP tools to augment existing manual counterdata collection projects. For example, D'Ignazio and Klein [2020] provide an example of a single citizen in Mexico generating a map of femicides from manually reading news reports.[2]

## 6.2 Final thoughts

It is an invigorating era for computational social science and text-as-data research. The explosion of available text data that has accompanied the digital age has provided many opportunities to quantitatively analyze the relationships between language use and human thought, actions, and societal structure. Key to these quantitative conclusions are improved measurement and causal inference, the foci of this thesis. However, we believe the research community must be vigilant of replication issues, not necessarily because of lack of transparency or open data, but because of the "garden of forking paths" that different text methods may result in wildly different conclusions. Additionally, we echo the cautions of D'Ignazio and Klein [2020] and Crawford [2021] that what we—as researchers and society—decide to measure often becomes the basis for policy-making and resource allocation and what is not measured risks becoming invisible. Holding these tensions as we decide what text-based measures and causal estimates to focus on is crucial moving forward.

[2]https://feminicidiosmx.crowdmap.com

# APPENDIX

# POLICE FATALITY APPENDIX

## A.1 Document retrieval from Google News

Our news dataset is created using documents gathered via Google News. Specifically, we issued search queries to Google News[1] United States (English) regional edition throughout 2016. Our scraper issued queries with terms from two lists: (1) a list of 22 words closely related to police officers and (2) a list of 21 words closely related to killing. These lists were semi-automatically constructed by looking up the nearest neighbors of "police" and "kill" (by cosine distance) from Google's public release of *word2vec* vectors pretrained on a very large (proprietary) Google News corpus,[2] and then manually excluding a small number of misspelled words or redundant capitalizations (e.g. "Police" and "police").

Our list of police words includes: police, officer, officers, cop, cops, detective, sheriff, policeman, policemen, constable, patrolman, sergeant, detectives, patrolmen, policewoman, constables, trooper, troopers, sergeants, lieutenant, deputies, deputy.

Our list of kill words includes: kill, kills, killing, killings, killed, shot, shots, shoot, shoots, shooting, murder, murders, murdered, beat, beats, beating, beaten, fatal, homicide, homicides.

We construct one word queries using single terms drawn from one of the two lists, as well as two-word queries which consist of one word drawn from each list (e.g. "police

[1] https://news.google.com/

[2] https://code.google.com/archive/p/word2vec/

| rank | name | positive | analysis |
|---|---|---|---|
| 1 | **Keith Scott** | **true** | |
| 2 | **Terence Crutcher** | **true** | |
| 3 | Alfred Olango | true | |
| 4 | Deborah Danner | true | |
| 5 | Carnell Snell | true | |
| 6 | Kajuan Raye | true | |
| 7 | Terrence Sterling | true | |
| 8 | Francisco Serna | true | |
| 9 | Sam DuBose | false | name mismatch |
| 10 | Michael Vance | true | |
| 11 | Tyre King | true | |
| 12 | Joshua Beal | true | |
| 13 | Trayvon Martin | false | killed, not by police |
| 14 | **Mark Duggan** | **false** | **non-US** |
| 15 | Kirk Figueroa | true | |
| 16 | Anis Amri | false | non-US |
| 17 | **Logan Clarke** | **false** | **shot not killed** |
| 18 | Craig McDougall | false | non-US |
| 19 | Frank Clark | true | |
| 20 | Benjamin Marconi | false | name of officer |

Table A.1: Top 20 entity predictions given by soft-LR (excluding historical entities) evaluated as "true" or "false" based on matching the gold knowledge base. False positives were manually analyzed. See Table 7 in the main paper for more detailed information regarding bold-faced entities.

shoot" or "cops gunfire"), yielding 505 different queries (22×21 + 22 + 21), each of which was queried approximately once per hour throughout 2016.[3] This yielded a list of recent results matching the query; the scraper downloaded documents whose URL it had not seen before, eventually collecting 1,162,300 web pages (approx. 3000 per day).

| Model | AUPRC | SE-1 | SE-2 | SE-3 | F1 | SE-1 | SE-2 | SE-3 |
|---|---|---|---|---|---|---|---|---|
| (m1) hard-LR, dep. feats. | 0.117 | (0.018) | (0.005) | (0.004) | 0.229 | (0.021) | (0.009) | (0.008) |
| (m2) hard-LR, n-gram feats. | 0.134 | (0.020) | (0.006) | (0.005) | 0.257 | (0.022) | (0.011) | (0.009) |
| (m3) hard-LR, all feats. | 0.142 | (0.021) | (0.006) | (0.005) | 0.266 | (0.023) | (0.010) | (0.009) |
| (m4) hard-CNN | 0.130 | (0.019) | (0.006) | (0.005) | 0.252 | (0.022) | (0.009) | (0.009) |
| (m5) soft-CNN (EM) | 0.164 | (0.023) | (0.007) | (0.007) | 0.267 | (0.023) | (0.009) | (0.009) |
| **(m6) soft-LR (EM)** | **0.193** | (0.025) | (0.008) | (0.008) | **0.316** | (0.025) | (0.011) | (0.010) |
| Data upper bound (§3.5.6) | 0.57 | – | – | – | 0.73 | – | – | – |

Table A.2: Area under precision-recall curve (AUPRC) and F1 (its maximum value from the PR curve) for entity prediction on the test set with bootstrap standard errors (SE) sampling from (1) entities (2) documents (3) documents without replacement.

## A.2 Document preprocessing

Once documents are downloaded from URLs collected via Google news queries, we apply text extraction with the Lynx browser[4] to extract text from HTML. (Newer open-source packages, like Boilerpipe and Newspaper, exist for text extraction, but we observed they often failed on our web data.)

## A.3 Mention-level preprocessing

From the corpus of scraped news documents, to create the mention-level dataset (i.e. the set of sentences containing candidate entities) we :

1. Apply the Lynx text-based web browser to extract all a webpage's text.
2. Segment sentences in two steps:
   (a) Segment documents to fragments of text (typically, paragraphs) by splitting on Lynx's representation of HTML paragraph, list markers, and other dividers: double newlines and the characters -,*, |, + and #.

[3]We also collected data during part of 2015; the volume of search results varied over time due to changes internal to Google News. After the first few weeks in 2016, the volume was fairly constant.

[4]Version 2.8

(b) Apply spaCy's sentence segmenter (and NLP pipeline) to these paragraph-like text fragments.

3. De-duplicate sentences as described in detail below.
4. Remove sentences that have fewer than 5 tokens or more than 200.
5. Remove entities (and associated mentions) that
   (a) Contain punctuation (except for periods, hyphens and apostrophes).
   (b) Contain numbers.
   (c) Are one token in length.
6. Strip any "'s" occurring at the end of named entity spans.
7. Strip titles (i.e. Ms., Mr. Sgt., Lt.) occurring in entity spans. (HAPNIS sometimes identifies these types of titles; this step basically augments its rules.)
8. Filter to mentions that contain at least one police keyword and at least one fatality keyword.

Additionally, we remove literal duplicate sentences from our mention-level dataset, eliminating all but one duplicated sentence. We select the earliest sentence by download time of its scraped webpage.

## A.4 *Noisyor* numerical stability

Under "hard" training, many entities at test time have probabilities very close to 1; in some cases, higher than $1 - e^{-1000}$. This happens for entities with a very large number of mentions, where the naive implementation of *noisyor* as $p = 1 - \prod_i (1 - p_i)$ has numerical underflow, causing many ties with entities having $p = 1$. In fact, random tie-breaking for ordering these entity predictions can give moderate variance to the AUPRC. (Part of the issue is that floating point numbers have worse tolerance near 1 than near 0.)

Instead, we rank entity predictions by the log of the complement probability (i.e. 1000 for $p = 1 - e^{-1000}$):

$$\begin{aligned} &\log\left(1 - P(y_e = 1 \mid x_{\mathcal{M}(e)})\right) \\ &= \sum_i \log P(z_i = 0 \mid x_i) \end{aligned}$$

This is more stable, and while there are a small number of ties, the standard deviation of AUPRC across random tie breakings is less than $10^{-10}$.

## A.5 Manual analysis of results

Manual analysis is available in Table A.1.

## A.6 Bootstrap

We conduct three different methods of bootstrap resampling, varying the objects being sampled:

1. Entities
2. Documents
3. Documents, with deduplication of mentions.[5]

We resample both test-set entities and test-set documents because we are currently unaware of literature that provides reasoning for one over the other, and both are arguably relevant in our context. The bootstrap sampling model assumes a given dataset represents a finite sample from a theoretically infinite population, and asks what variability there would be if a finite sample were to be drawn again from the

[5]To implement, we take the 10,000 samples (with replacement) of documents, and reduce them to the unique set of drawn documents. This effectively removes duplicate mentions that occur in method 2 when the same document is drawn more than once in a sample.

population. This has different interpretations for entity and document resampling. Resampling entities measures robustness due to variability in the names that occur in the documents. Resampling documents measures robustness due to variability in our data source—for example, if our document scraping procedure was altered, or potentially, if the news generation process was changed. Since both entities and documents are not i.i.d., these are both dissatisfying assumptions.

We also conduct resampling of documents with deduplication of mentions since, during development, we found our noisy-or metric was sensitive to duplicate mentions; this deduplication step effectively includes running our analysis pipeline's sentence deduplication for each bootstrap sample.

In Fig. A.2, we augment the results from Fig. 3.7 with standard errors calculated from $B = 10,000$ bootstrap samples given the three methods for sampling described above. Document resampling tends to give smaller standard errors than entity resampling, which is to be expected since there is a larger number of documents than entities. We analyze our results using the standard errors and significance tests from method 3.

We examine the statistical significance of difference between models with a one-sided hypothesis test. Our statistic is

$$T_{ij} = AUPRC_{\text{model } j} - AUPRC_{\text{model } i}.$$

We use hypotheses $H_0 : T \leq 0$ and $H_1 : T > 0$. As above, we take 10,000 bootstrap samples and find $T^b$ statistic of each sample $b \in \{1..10000\}$. Then we compute p-values

$$\text{p-value}_{ij} = \frac{\text{Count}(T_{ij}^b \leq 0)}{10000}.$$

Finally, since in the observed data, one model is better than the other, we are interested the null hypothesis that the apparently-worse model outperforms the apparently-

better model. Therefore the final p-value comparing systems $i$ and $j$ is actually calculated as $\min(p_{ij}, p_{ji})$, since the different directions correspond to the fraction of bootstrap samples with $T_{ij} \leq 0$ versus $T_{ij} > 0$; these values are shown in Fig. A.3. (Note $p_{ji} = 1 - p_{ij}$ in expectation.) While this seems to follow standard practice in bootstrap hypothesis testing in NLP [Berg-Kirkpatrick et al., 2012], we note that MacKinnon [2009] argues to instead multiply that by two (i.e., calculate $2\min(p_{ij}, p_{ji})$) to conduct a two-sided test that correctly gives $p \sim \text{Unif}(0, 1)$ when a null hypothesis of equivalent performance is true.

| | m2 | m3 | m4 | m5 | m6 |
|---|---|---|---|---|---|
| m1 | 2.7e-1 | 1.8e-1 | 3.1e-1 | 6.0e-2 | 6.2e-3 |
| m2 | | 3.8e-1 | 4.5e-1 | 1.7e-1 | 3.2e-2 |
| m3 | | | 3.3e-1 | 2.5e-1 | 5.8e-2 |
| m4 | | | | 1.4e-1 | 2.2e-2 |
| m5 | | | | | 1.9e-1 |

(a) Entity resampling

| | m2 | m3 | m4 | m5 | m6 |
|---|---|---|---|---|---|
| m1 | 3.5e-2 | 1.7e-3 | 5.0e-2 | 0 | 0 |
| m2 | | 1.8e-1 | 4.1e-1 | 3.6e-3 | 0 |
| m3 | | | 1.2e-1 | 3.1e-2 | 0 |
| m4 | | | | 2.1e-3 | 0 |
| m5 | | | | | 1.2e-2 |

(b) Document resampling

| | m2 | m3 | m4 | m5 | m6 |
|---|---|---|---|---|---|
| m1 | 2.2e-2 | 8.2-4 | 9.3e-2 | 1e-4 | 0 |
| m2 | | 1.5e-1 | 2.6e-1 | 7.3e-3 | 0 |
| m3 | | | 4.6e-2 | 5.9e-2 | 0 |
| m4 | | | | 1.6e-3 | 0 |
| m5 | | | | | 2.7e-3 |

(c) Document resampling with deduplication

Table A.3: One-sided p-values for for the difference between two models using statistic $T_{ij}$ where $AUPRC_{\text{model } j} > AUPRC_{\text{model } i}$; each cell in the table shows $\min(p_{ij}, p_{ji})$.

# BIBLIOGRAPHY


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

Jacob Andreas and Dan Klein. Reasoning about pragmatics with neural listeners and speakers. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1173–1182, 2016.

Galen Andrew and Jianfeng Gao. Scalable training of L1-regularized log-linear models. In *Proceedings of the 24th International Conference on Machine Learning*, 2007.

Joshua D Angrist and Jörn-Steffen Pischke. *Mostly harmless econometrics: An empiricist's companion*. Princeton university press, 2008.

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

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

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

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

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

Scott R Baker, Nicholas Bloom, and Steven J Davis. Measuring economic policy uncertainty. *The quarterly journal of economics*, 131(4):1593–1636, 2016.

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

David Bamman, Brendan O'Connor, and Noah A. Smith. Learning latent personas of film characters. In *Proceedings of ACL*, 2013.

David Bamman, Ted Underwood, and Noah A. Smith. A Bayesian mixed effects model of literary character. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, June 2014. URL http://www.aclweb.org/anthology/P14-1035.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. Abstract meaning representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, Sofia, Bulgaria, August 2013. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W13-2322.

Duren Banks, Paul Ruddle, Erin Kennedy, and Michael G. Planty. Arrest-related deaths program redesign study, 2015–16: Preliminary findings, 2016.

Antonio Bella, Cesar Ferri, José Hernández-Orallo, and Maria Jose Ramirez-Quintana. Quantification via probability estimators. In *IEEE 10th International Conference on Data Mining (ICDM)*, 2010.

Emily M Bender. Linguistic fundamentals for natural language processing: 100 essentials from morphology and syntax. *Synthesis lectures on human language technologies*, 6(3):1–184, 2013.

Emily M Bender and Alex Lascarides. Linguistic fundamentals for natural language processing ii: 100 essentials from semantics and pragmatics. *Synthesis Lectures on Human Language Technologies*, 12(3):1–268, 2019.

Taylor Berg-Kirkpatrick, David Burkett, and Dan Klein. An empirical investigation of statistical significance in NLP. In *Proceedings of EMNLP*, 2012.

Richard Berk, Azusa Li, and Laura J Hickman. Statistical difficulties in determining the role of race in capital cases: A re-analysis of data from the state of maryland. *Journal of Quantitative Criminology*, 21(4):365–390, 2005.

George Bissias, Brian Levine, Marc Liberatore, Brian Lynn, Juston Moore, Hanna Wallach, and Janis Wolak. Characterization of contact offenders and child exploitation material trafficking on five peer-to-peer networks. *Child Abuse & Neglect*, 52: 185 – 199, 2016. ISSN 0145-2134. doi: https://doi.org/10.1016/j.chiabu.2015.10.022. URL http://www.sciencedirect.com/science/article/pii/S0145213415003804.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

John Blitzer, Mark Dredze, and Fernando Pereira. Biographies, bollywood, boomboxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th annual meeting of the association of computational linguistics*, pages 440–447, 2007.

Su Lin Blodgett. Sociolinguistically driven approaches for just natural language processing. *PhD Thesis*, 2021.

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

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

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

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

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

Elizabeth Boschee, Premkumar Natarajan, and Ralph Weischedel. Automatic extraction of events from open source text for predictive forecasting. *Handbook of Computational Approaches to Counterterrorism*, page 51, 2013.

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

Razvan Bunescu and Raymond Mooney. Learning to extract relations from the web using minimal supervision. In *Proceedings of ACL*, pages 576–583, Prague, Czech Republic, June 2007. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P07-1073.

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

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

Dallas Card and Noah A Smith. The importance of calibration for estimating proportions from annotations. In *Proceedings of Empirical Methods in Natural Language Processing*, 2018.

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

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

Andrea Ceron, Luigi Curini, and Stefano M Iacus. Using sentiment analysis to monitor electoral campaigns: Method matters—evidence from the United States and Italy. *Social Science Computer Review*, 33(1):3–20, 2015.

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

Daniel Chen, Yosh Halberstam, and Alan CL Yu. Perceived masculinity predicts us supreme court outcomes. *PloS one*, 11(10):e0164324, 2016.

Daniel L Chen, Yosh Halberstam, Manoj Kumar, and Alan Yu. Attorney voice and the us supreme court. *Law as Data, Santa Fe Institute Press, ed. M. Livermore and D. Rockmore*, 2019.

Danqi Chen and Christopher Manning. A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 740–750, Doha, Qatar, October 2014. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/D14-1082.

Sean X. Chen and Jun S. Liu. Statistical applications of the Poisson-binomial and conditional Bernoulli distributions. *Statistica Sinica*, pages 875–892, 1997.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, and Whitney Newey. Double/debiased/neyman machine learning of treatment effects. *American Economic Review*, 107(5):261–65, 2017.

Laura Chiticariu, Yunyao Li, and Frederick R. Reiss. Rule-based information extraction is dead! long live rule-based information extraction systems! In *Proceedings of EMNLP*, 2013. URL http://www.aclweb.org/anthology/D13-1079.

Mark Craven and Johan Kumlien. Constructing biological knowledge bases by extracting information from text sources. In *ISMB*, pages 77–86, 1999.

Mark Craven, Andrew McCallum, Dan PiPasquo, Tom Mitchell, and Dayne Freitag. Learning to extract symbolic knowledge from the World Wide Web. In *Proceedings of AAAI*, 1998.

Kate Crawford. *The Atlas of AI*. Yale University Press, 2021.

Morgan Currie, Britt S Paris, Irene Pasquetto, and Jennifer Pierre. The conundrum of police officer-involved homicides: Counter-data in los angeles county. *Big Data & Society*, 3(2):2053951716663566, 2016.

Alexander D'Amour, Peng Ding, Avi Feller, Lihua Lei, and Jasjeet Sekhon. Overlap in observational studies with high-dimensional covariates. *Journal of Econometrics*, 221(2):644–654, 2021.

Cristian Danescu-Niculescu-Mizil, Lillian Lee, Bo Pang, and Jon Kleinberg. Echoes of power: Language effects and power differences in social interaction. In *WWW*, 2012.

Rhian M Daniel, Bianca L De Stavola, SN Cousens, and Stijn Vansteelandt. Causal mediation analysis with multiple mediators. *Biometrics*, 71(1):1–14, 2015.

Dipanjan Das, Desai Chen, Andre F. T. Martins, Nathan Schneider, and Noah A. Smith. Frame-semantic parsing. *Computational Linguistics*, 2014.

Hal Daumé III. Frustratingly easy domain adaptation. *ACL 2007*, page 256, 2007.

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

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

Kay Deaux. Sex and gender. *Annual review of psychology*, 36(1):49–81, 1985.

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

Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (methodological)*, pages 1–38, 1977.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

Thomas G Dietterich, Richard H Lathrop, and Tomás Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial intelligence*, 89 (1):31–71, 1997.

Catherine D'Ignazio and Lauren F Klein. *Data feminism*. MIT press, 2020.

George R Doddington, Alexis Mitchell, Mark A Przybocki, Lance A Ramshaw, Stephanie Strassel, and Ralph M Weischedel. The automatic content extraction (ACE) program-tasks, data, and evaluation. In *LREC*, volume 2, page 1, 2004.

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

Anca Dumitrache, Lora Aroyo, and Chris Welty. Crowdsourcing ground truth for medical relation extraction. *ACM Transactions on Interactive Intelligent Systems*, 8(2):1–20, 2018.

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

Penelope Eckert. Variation and the indexical field 1. *Journal of sociolinguistics*, 12 (4):453–476, 2008.

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

Jacob Eisenstein. *Introduction to natural language processing*. MIT Press, 2019.

Jacob Eisenstein, Amr Ahmed, and Eric P. Xing. Sparse additive generative models of text. In *Proceedings of ICML*, pages 1041–1048, 2011.

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

Andrea Esuli and Fabrizio Sebastiani. Optimizing text quantifiers for multivariate loss functions. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 9 (4):27, 2015.

James A Evans and Pedro Aceves. Machine translation: Mining text for social theory. *Annual Review of Sociology*, 42:21–50, 2016.

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

Charles J. Fillmore, Christopher R. Johnson, and Miriam R.L. Petruck. Background to FrameNet. *International Journal of Lexicography*, 2003.

Matthew Finlayson, Aaron Mueller, Stuart Shieber, Sebastian Gehrmann, Tal Linzen, and Yonatan Belinkov. Causal analysis of syntactic agreement mechanisms in neural language models. In *ACL-IJCNLP*, 2021.

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

George Forman. Counting positives accurately despite inaccurate classification. In *European Conference on Machine Learning*, 2005.

George Forman. Quantifying counts and costs via classification. *Data Mining and Knowledge Discovery*, 17(2):164–206, 2008.

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

Katrin Fundel, Robert Küffner, and Ralf Zimmer. Relex—relation extraction using dependency parse trees. *Bioinformatics*, 23(3):365–371, 2006.

John J. Gart and Alfred A. Buck. Comparison of a screening test and a reference test in epidemiologic studies. II. A probabilistic model for the comparison of diagnostic tests. *American Journal of Epidemiology*, 83(3):593–602, 1966. doi: https://doi.org/10.1093/oxfordjournals.aje.a120610.

Timnit Gebru, Jonathan Krause, Yilun Wang, Duyun Chen, Jia Deng, Erez Lieberman Aiden, and Li Fei-Fei. Using deep learning and Google Street View to estimate the demographic makeup of neighborhoods across the United States. *Proceedings of the National Academy of Sciences*, 2017.

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

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

Matthew Gentzkow and Jesse M Shapiro. What drives media slant? evidence from us daily newspapers. *Econometrica*, 78(1):35–71, 2010.

Matthew Gentzkow, Bryan Kelly, and Matt Taddy. Text as data. *Journal of Economic Literature*, 57(3):535–74, 2019.

Sean M Gerrish. *Applications of Latent Variable Models in Modeling Influence and Decision Making*. PhD thesis, Princeton University, 2013.

Anindya Ghose, Panagiotis Ipeirotis, and Arun Sundararajan. Opinion mining using econometrics: A case study on reputation systems. In *Proceedings of ACL*, Prague, Czech Republic, 2007. URL http://www.aclweb.org/anthology/P07-1053.

Kevin Gimpel and Noah A. Smith. Softmax-margin CRFs: Training log-linear models with cost functions. In *Proceedings of NAACL-HLT*, pages 733–736. Association for Computational Linguistics, 2010.

Shane A Gleason. Beyond mere presence: Gender norms in oral arguments at the us supreme court. *Political Research Quarterly*, 73(3):596–608, 2020.

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

Yoav Goldberg. A primer on neural network models for natural language processing. *Journal of Artificial Intelligence Research*, 57:345–420, 2016.

Pablo González, Alberto Castaño, Nitesh V. Chawla, and Juan José Del Coz. A review on quantification learning. *ACM Computing Surveys*, 50(5):74:1–74:40, September 2017a. ISSN 0360-0300. doi: 10.1145/3117807. URL http://doi.acm.org/10.1145/3117807.

Pablo González, Jorge Díez, Nitesh Chawla, and Juan José del Coz. Why is quantification an interesting learning problem? *Progress in Artificial Intelligence*, 6(1):53–58, 2017b.

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

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

Justin Grimmer. We are all social scientists now: How big data, machine learning, and causal inference work together. *PS, Political Science & Politics*, 48(1):80, 2015.

Justin Grimmer and Brandon M Stewart. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political analysis*, 21(3):267–297, 2013.

Justin Grimmer, Solomon Messing, and Sean J Westwood. How words and money cultivate a personal vote: The effect of legislator credit claiming on constituent credit allocation. *American Political Science Review*, 106(4):703–719, 2012.

Justin Grimmer, Margaret E. Roberts, and Brandon Stewart. *Text as Data: A New Framework for Machine Learning and the Social Sciences*. Princeton University Press, 2021.

Kadri Hacioglu. Semantic role labeling using dependency trees. In *Proceedings of the 20th international conference on Computational Linguistics*. Association for Computational Linguistics, 2004.

Jan Hajic, Eva Hajicová, Jarmila Panevová, Petr Sgall, Ondrej Bojar, Silvie Cinková, Eva Fucíková, Marie Mikulová, Petr Pajas, Jan Popelka, et al. Announcing prague czech-english dependency treebank 2.0. In *LREC*, pages 3153–3160, 2012.

Alon Halevy, Peter Norvig, and Fernando Pereira. The unreasonable effectiveness of data. *IEEE Intelligent Systems*, 24(2):8–12, 2009.

Andrew Halterman, Katherine A. Keith, Sheikh Sarwar, and Brendan O'Connor. Corpus-level evaluation for event QA: The IndiaPoliceEvents corpus covering the 2002 Gujarat violence. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2021.

David J Hand. Classifier technology and the illusion of progress. *Statistical science*, 21(1):1–14, 2006.

Alex Hanna. MPEDS: Automating the generation of protest event data. *SocArXiv*, 2017.

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

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

Miguel A Hernán and James M Robins. *Causal Inference: What If*. Boca Raton: Chapman & Hall/CRC, 2020.

Jennifer Hill and Yu-Sung Su. Assessing lack of common support in causal inference using bayesian nonparametrics: Implications for evaluating the effect of breastfeeding on children's cognitive outcomes. *The Annals of Applied Statistics*, pages 1386–1420, 2013.

R. A. Hinde. Interactions, relationships and social structure. *Man*, 11(1):1–17, 1976.

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

Raphael Hoffmann, Congle Zhang, Xiao Ling, Luke Zettlemoyer, and Daniel S. Weld. Knowledge-based weak supervision for information extraction of overlapping relations. In *Proceedings of ACL*, 2011. URL http://www.aclweb.org/anthology/P11-1055.

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

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

Daniel J. Hopkins and Gary King. A method of automated nonparametric content analysis for social science. *American Journal of Political Science*, 54(1):229–247, January 2010. ISSN 00925853. doi: 10.1111/j.1540-5907.2009.00428.x. URL http://dash.harvard.edu/handle/1/4142694.

Dirk Hovy, Federico Bianchi, and Tommaso Fornaciari. You Sound Just Like Your Father Commercial Machine Translation Systems Include Stylistic Biases. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1686–1690, 2020.

Eduard Hovy, Teruko Mitamura, Felisa Verdejo, Jun Araki, and Andrew Philpot. Events are not simple: Identity, non-identity, and quasi-identity. In *Workshop on Events: Definition, Detection, Coreference, and Representation*, pages 21–28, Atlanta, Georgia, June 2013. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W13-1203.

Lily Hu and Issa Kohler-Hausmann. What's sex got to do with machine learning? In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 513–513, 2020.

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

Kosuke Imai and Teppei Yamamoto. Identification and sensitivity analysis for multiple causal mechanisms: Revisiting evidence from framing experiments. *Political Analysis*, 21(2):141–171, 2013.

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

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

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

Miyako Inoue. *Vicarious language*. University of California Press, 2006.

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

Abigail Z Jacobs and Hanna Wallach. Measurement and fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 375–385, 2021.

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

Connor T Jerzak, Gary King, and Anton Strezhnev. An improved method of automated nonparametric content analysis for social science. *Political Analysis*, 2019.

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

Michael I Jordan and Robert A Jacobs. Hierarchical mixtures of experts and the em algorithm. *Neural computation*, 6(2):181–214, 1994.

Daniel Jurafsky and James H Martin. *Speech and Language Processing: An introduction to speech recognition, computational linguistics and natural language processing*. Pearson Prentice Hall, 2 edition, 2019.

Katherine Keith and Brendan O'Connor. Uncertainty-aware generative models for inferring document class prevalence. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4575–4585, 2018.

Katherine Keith, Abram Handler, Michael Pinkham, Cara Magliozzi, Joshua McDuffie, and Brendan O'Connor. Identifying civilians killed by police with distantly supervised entity-event extraction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1547–1557, 2017.

Katherine Keith, Su Lin Blodgett, and Brendan O'Connor. Monte carlo syntax marginals for exploring and using dependency parses. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 917–928, 2018.

Katherine Keith, David Jensen, and Brendan O'Connor. 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*, pages 5332–5344, Online, July 2020a. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.474. URL https://www.aclweb.org/anthology/2020.acl-main.474.

Katherine Keith, Christoph Teichmann, Brendan O'Connor, and Edgar Meij. Uncertainty over uncertainty: Investigating the assumptions, annotations, and text measurements of economic policy uncertainty. In *Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science*, pages 116–131, 2020b.

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

Yoon Kim. Convolutional neural networks for sentence classification. In *Proceedings of EMNLP*, 2014.

Gary King, Jennifer Pan, and Margaret E. Roberts. How censorship in china allows government criticism but silences collective expression. *American Political Science Review*, 107:1–18, 2013.

Paul Kingsbury and Martha Palmer. From TreeBank to PropBank. In *LREC*, pages 1989–1993, 2002.

Pang Wei Koh, Shiori Sagawa, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pages 5637–5664. PMLR, 2021.

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

Alexander Konovalov, Benjamin Strauss, Alan Ritter, and Brendan O'Connor. Learning to extract events from knowledge base revisions. In *Proceedings of WWW*, 2017.

Klaus Krippendorff. *Content analysis: An introduction to its methodology*. Sage publications, 2018.

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

Robin Lakoff. Language and woman's place. *Language in society*, 2(1):45–79, 1973.

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

David Lazer, Alex Sandy Pentland, Lada Adamic, Sinan Aral, Albert Laszlo Barabasi, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, et al. Life in the network: the coming age of computational social science. *Science (New York, NY)*, 323(5915):721, 2009.

Heeyoung Lee, Marta Recasens, Angel Chang, Mihai Surdeanu, and Dan Jurafsky. Joint entity and event coreference resolution across documents. In *Proceedings of EMNLP*, 2012.

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

Juncen Li, Robin Jia, He He, and Percy Liang. Delete, retrieve, generate: a simple approach to sentiment and style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1865–1874, 2018.

Qi Li and Heng Ji. Incremental joint extraction of entity mentions and relations. In *Proceedings of ACL*, 2014.

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

Percy Liang and Dan Klein. Online EM for unsupervised models. In *Proceedings of NAACL*, Boulder, Colorado, 2009. URL http://www.aclweb.org/anthology/N/N09/N09-1069.

Dekang Lin and Patrick Pantel. DIRT – discovery of inference rules from text. In *Proceedings of KDD*, 2001.

Yankai Lin, Shiqi Shen, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. Neural relation extraction with selective attention over instances. In *Proceedings of ACL*, pages 2124–2133, Berlin, Germany, August 2016. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P16-1200.

Jane Loevinger. Objective tests as instruments of psychological theory. *Psychological reports*, 3(3):635–694, 1957.

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

Kristian Lum and Patrick Ball. Estimating undocumented homicides with two lists and list dependence. *Human Rights Data Analysis Group*, April 2015. URL https://hrdag.org/wp-content/uploads/2015/07/2015-hrdag-estimating-undoc-homicides.pdf.

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

James G MacKinnon. Bootstrap hypothesis testing. *Handbook of Computational Econometrics*, 2009.

Benjamin Mandel, Aron Culotta, John Boulahanis, Danielle Stark, Bonnie Lewis, and Jeremy Rodrigue. A demographic analysis of online sentiment during Hurricane Irene. In *Proceedings of the Second Workshop on Language in Social Media*, pages 27–36. Association for Computational Linguistics, 2012.

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

Mitchell Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. Building a large annotated corpus of English: The Penn Treebank. 1993.

Andrew McCallum and Kamal Nigam. A comparison of event models for naive Bayes text classification. In *AAAI-98 Workshop on Learning for Text Categorization*, volume 752, pages 41–48, 1998.

Tali Mendelberg, Christopher F Karpowitz, and J Baxter Oliphant. Gender inequality in deliberation: Unpacking the black box of interaction. *Perspectives on Politics*, 12(1):18–44, 2014.

Samuel Messick. Validity. *ETS Research Report Series*, 1987(2):i–208, 1987.

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

Mike Mintz, Steven Bills, Rion Snow, and Daniel Jurafsky. Distant supervision for relation extraction without labeled data. In *Proceedings of ACL*, Suntec, Singapore, 2009. URL http://www.aclweb.org/anthology/P/P09/P09-1113.

B. L. Monroe, M. P. Colaresi, and K. M. Quinn. Fightin' Words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, 16(4):372, 2008.

Will Monroe. *Learning in the rational speech acts model*. PhD thesis, Stanford University, 2018.

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

Jose G. Moreno-Torres, Troy Raeder, Rocío Alaiz-Rodríguez, Nitesh V. Chawla, and Francisco Herrera. A unifying view on dataset shift in classification. *Pattern Recognition*, 45(1):521–530, 2012.

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

Shunji Mori, Hirobumi Nishida, and Hiromitsu Yamada. *Optical character recognition*. John Wiley & Sons, Inc., 1999.

Frederick Mosteller and David L Wallace. Inference in an authorship problem: A comparative study of discrimination methods applied to the authorship of the disputed federalist papers. *Journal of the American Statistical Association*, 58(302): 275–309, 1963.

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

Kevin Munger. Tweetment effects on the tweeted: Experimentally reducing racist harassment. *Political Behavior*, 39(3):629–649, 2017.

Preslav Nakov, Alan Ritter, Sara Rosenthal, Fabrizio Sebastiani, and Veselin Stoyanov. Semeval-2016 Task 4: Sentiment analysis in twitter. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, June 2016. URL http://www.aclweb.org/anthology/S16-1001.

Radford M Neal and Geoffrey E Hinton. A view of the EM algorithm that justifies incremental, sparse, and other variants. In *Learning in graphical models*, pages 355–368. Springer, 1998.

Kimberly A Neuendorf. *The content analysis guidebook*. SAGE, 2017.

Andrew Ng and Michael Jordan. On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes. *Advances in neural information processing systems*, 14:841, 2002.

Dong Nguyen, Maria Liakata, Simon DeDeo, Jacob Eisenstein, David Mimno, Rebekah Tromble, and Jane Winters. How we do things with words: Analyzing text as social and cultural data. *Frontiers in Artificial Intelligence*, 3:62, 2020.

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

Thien Huu Nguyen and Ralph Grishman. Event detection and domain adaptation with convolutional neural networks. In *Proceedings of ACL*, 2015.

Brendan O'Connor, David Bamman, and Noah A. Smith. Computational text analysis for social science: Model assumptions and complexity. In *Second Workshop on Comptuational Social Science and the Wisdom of Crowds (NIPS 2011)*, 2011.

Brendan O'Connor, Brandon Stewart, and Noah A. Smith. Learning to extract international relations from political context. In *Proceedings of ACL*, 2013.

Stephan Oepen, Marco Kuhlmann, Yusuke Miyao, Daniel Zeman, Dan Flickinger, Jan Hajic, Angelina Ivanova, and Yi Zhang. Semeval 2014: Task 8, broad-coverage semantic dependency parsing. In *Proceedings of SemEval*, 2014. URL http://www.aclweb.org/anthology/S14-2008.

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

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

Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kiciman. Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data*, 2: 13, 2019.

Oyez. Kennedy v. Plan Administrator for DuPont Sav. and Investment Plan., a. URL https://www.oyez.org/cases/2008/07-636. Accessed 15 Jul. 2021.

Oyez. Lozano v. Montoya Alvarez, b. URL https://www.oyez.org/cases/2013/12-820. Accessed 15 Jul. 2021.

Terence Parsons. *Events in the Semantics of English*. Cambridge, MA: MIT Press, 1990.

Dana Patton and Joseph L Smith. Lawyer, interrupted: Gender bias in oral arguments at the us supreme court. *Journal of Law and Courts*, 5(2):337–361, 2017.

John W Patty and Elizabeth Maggie Penn. Analyzing big data: social choice and measurement. *PS, Political Science & Politics*, 48(1):95, 2015.

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

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

Ellie Pavlick, Heng Ji, Xiaoman Pan, and Chris Callison-Burch. The Gun Violence Database: A new task and data set for NLP. In *Proceedings of EMNLP*, 2016. URL https://aclweb.org/anthology/D16-1106.

Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1988. ISBN 0-934613-73-7.

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

Judea Pearl. Direct and indirect effects. In *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*, pages 411–420, 2001.

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

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

Judea Pearl. On measurement bias in causal inference. In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*, pages 425–432, 2010.

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

Judea Pearl. The seven tools of causal inference, with reflections on machine learning. *Communications of the ACM*, 62(3):54–60, 2019.

Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. *Causal inference in statistics: A primer*. John Wiley & Sons, 2016.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

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

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

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

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

Deanna Poos and Rita Simpson. Cross-disciplinary comparisons of hedging. *Using corpora to explore linguistic variation*, pages 3–23, 2002.

Shrimai Prabhumoye, Yulia Tsvetkov, Ruslan Salakhutdinov, and Alan W Black. Style transfer through back-translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 866–876, 2018.

Anna Prokofieva and Julia Hirschberg. Hedging and speaker commitment. In *5th Intl. Workshop on Emotion, Social Signals, Sentiment & Linked Open Data, Reykjavik, Iceland*, 2014.

Reid Pryzant, Dallas Card, Dan Jurafsky, Victor Veitch, and Dhanya Sridhar. 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*, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.323. URL https://aclanthology.org/2021.naacl-main.323.

Kevin M Quinn, Burt L Monroe, Michael Colaresi, Michael H Crespin, and Dragomir R Radev. How to analyze political attention with minimal assumptions and costs. *American Journal of Political Science*, 54(1):209–228, 2010.

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

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

Kevin Reschke, Martin Jankowiak, Mihai Surdeanu, Christopher D. Manning, and Daniel Jurafsky. Event extraction using distant supervision. In *Language Resources and Evaluation Conference (LREC)*, 2014.

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

Sebastian Riedel, Limin Yao, and Andrew McCallum. Modeling relations and their mentions without labeled text. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 148–163. Springer, 2010.

Sebastian Riedel, Limin Yao, Andrew McCallum, and Benjamin M. Marlin. Relation extraction with matrix factorization and universal schemas. In *Proceedings of NAACL*, Atlanta, Georgia, 2013. URL http://www.aclweb.org/anthology/N13-1008.

Alan Ritter, Luke Zettlemoyer, Oren Etzioni, et al. Modeling missing data in distant supervision for information extraction. *TACL*, 2013.

Margaret E Roberts, Brandon M Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson, and David G Rand. Structural topic models for open-ended survey responses. *American Journal of Political Science*, 58(4):1064–1082, 2014.

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

Jonathan Rosa and Nelson Flores. Unsettling race and language: Toward a raciolinguistic perspective. *Language in society*, 46(5):621–647, 2017.

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

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

Sara Rosenthal, Noura Farra, and Preslav Nakov. Semeval-2017 Task 4: Sentiment analysis in twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, August 2017. URL http://www.aclweb.org/anthology/S17-2088.

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

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

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

Marco Saerens, Patrice Latinne, and Christine Decaestecker. Adjusting the outputs of a classifier to new a priori probabilities: a simple procedure. *Neural computation*, 14(1):21–41, 2002.

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

Ruslan Salakhutdinov, Sam T Roweis, and Zoubin Ghahramani. Optimization with EM and expectation-conjugate-gradient. In *Proceedings of ICML*, 2003.

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

Christian Sandvig and Eszter Hargittai. How to think about digital research. *Digital confidential: The secrets of studying online behavior*, pages 1–25, 2015.

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

Alexandra Schofield, Måns Magnusson, and David Mimno. Pulling out the stops: Rethinking stopword removal for topic models. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 432–436, 2017.

Philip A. Schrodt. Precedents, progress, and prospects in political event data. *International Interactions*, 38(4):546–569, 2012.

Philip A. Schrodt and Deborah J. Gerner. Validity assessment of a machine-coded event data set for the Middle East, 1982-1992. *American Journal of Political Science*, 1994.

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

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

Yanchuan Sim, Brice Acree, Justin H. Gross, and Noah A. Smith. Measuring ideological proportions in political speeches. In *Proceedings of EMNLP*, 2013.

Rion Snow, Dan. Jurafsky, and Andrew Y. Ng. Learning syntactic patterns for automatic hypernym discovery. In *Advances in Neural Information Processing Systems*, 2005.

Harold J Spaeth, Lee Epstein, Andrew D Martin, Jeffrey A Segal, Theodore J Ruger, and Sara C Benesh. Supreme Court Database, Version 2021 Release 01. *Database at http://scdb.wustl.edu/*, 2021.

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

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

SS Stevens. On the theory of scales of measurement. *SCIENCE*, 103(2684), 1946.

Jennifer Stromer-Galley. Measuring deliberation's content: A coding scheme. *Journal of public deliberation*, 3(1), 2007.

Elizabeth A Stuart. Matching methods for causal inference: A review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 25(1):1, 2010.

Milan Sulc and Jiri Matas. Improving cnn classifiers by estimating test-time priors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019.

Mihai Surdeanu, Julie Tibshirani, Ramesh Nallapati, and Christopher D. Manning. Multi-instance multi-label learning for relation extraction. In *Proceedings of EMNLP*, 2012. URL http://www.aclweb.org/anthology/D12-1042.

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

Matt Taddy. Multinomial inverse regression for text analysis. *Journal of the American Statistical Association*, 108(503):755–770, 2013.

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

Dirk Tasche. Fisher consistency for prior probability shift. *Journal of Machine Learning Research*, 18(95):1–32, 2017. URL http://jmlr.org/papers/v18/17-048.html.

Sam Thomson, Brendan O'Connor, Jeffrey Flanigan, David Bamman, Jesse Dodge, Swabha Swayamdipta, Nathan Schneider, Chris Dyer, and Noah A. Smith. CMU: Arc-factored, discriminative semantic dependency parsing. In *Proceedings of SemEval*, 2014. URL http://www.aclweb.org/anthology/S14-2027.

Dustin Tingley, Teppei Yamamoto, Kentaro Hirose, Luke Keele, and Kosuke Imai. Mediation: R package for causal mediation analysis. 2014.

Mark J Van der Laan and Sherri Rose. *Targeted learning: causal inference for observational and experimental data*. Springer Science & Business Media, 2011.

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

Tyler VanderWeele and Stijn Vansteelandt. Mediation analysis with multiple mediators. *Epidemiologic methods*, 2(1):95–115, 2014.

Tyler J VanderWeele. Mediation analysis: a practitioner's guide. *Annual review of public health*, 37:17–32, 2016.

Victor Veitch and Anisha Zaveri. Sense and sensitivity analysis: Simple post-hoc analysis of bias due to unobserved confounding. *Neural Information Processing Systems (NeurIPS)*, 2020.

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

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart M Shieber. Investigating gender bias in language models using causal mediation analysis. In *NeurIPS*, 2020.

Rob Voigt, Nicholas P Camp, Vinodkumar Prabhakaran, William L Hamilton, Rebecca C Hetey, Camilla M Griffiths, David Jurgens, Dan Jurafsky, and Jennifer L Eberhardt. Language from police body camera footage shows racial disparities in officer respect. *Proceedings of the National Academy of Sciences*, 114(25):6521–6526, 2017.

Slobodan Vucetic and Zoran Obradovic. Classification on data with biased class distribution. In *European Conference on Machine Learning*, pages 527–538. Springer, 2001.

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

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, 2018.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. Superglue: a stickier benchmark for general-purpose language understanding systems. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 3266–3280, 2019.

William Yang Wang, Elijah Mayfield, Suresh Naidu, and Jeremiah Dittmar. Historical analysis of legal opinions with a sparse mixed-effects latent variable model. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL)*, July 2012. URL http://www.aclweb.org/anthology/P12-1078.

Larry Wasserman. *All of statistics*. Springer Science & Business Media, 2011.

Duncan J Watts. *Everything is obvious:* Once you know the answer*. Atlantic Books, 2011.

Kilian Weinberger, Anirban Dasgupta, John Langford, Alex Smola, and Josh Attenberg. Feature hashing for large scale multitask learning. In *Proceedings of ICML*, 2009.

Candace West and Don H Zimmerman. Doing gender, gender and society. *Vol. 1, No. 2.(Jun*, page 125, 1987.

Steven Wilkinson. *Votes and violence: Electoral competition and ethnic riots in India*. Cambridge University Press, 2006.

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

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

Jack Chongjie Xue and Gary M Weiss. Quantification and semi-supervised classification methods for handling changes in class distribution. In *Proceedings of KDD*, 2009.

Bishan Yang, Claire Cardie, and Peter Frazier. A hierarchical distance-dependent Bayesian model for event coreference resolution. *TACL*, 3, 2015.

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

Limin Yao, Sebastian Riedel, and Andrew McCallum. Collective cross-document relation extractionwithout labelled data. In *Proceedings of EMNLP*, 2010.

Matthew D. Zeiler. Adadelta: An adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012.

Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, and Jun Zhao. Relation classification via convolutional deep neural network. In *Proceedings of COLING*, 2014.

Justine Zhang, Sendhil Mullainathan, and Cristian Danescu-Niculescu-Mizil. Quantifying the causal effects of conversational tendencies. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):1–24, 2020.

Ye Zhang and Byron Wallace. A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. *arXiv preprint arXiv:1510.03820*, 2015.

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