TOWARDS ROBUST NATURAL LANGUAGE PROCESSING TO PROMOTE HEALTH EQUITY

by

Keith Harrigian

A dissertation submitted to The Johns Hopkins University

in conformity with the requirements for the degree of Doctor of Philosophy.

Baltimore, Maryland

August 2024

Abstract

Natural language processing (NLP) has rapidly become an integral component of contemporary healthcare infrastructure and is likely to become more deeply entrenched in the domain given the rise of large language models. As such, we find ourselves at a crossroads where it is perhaps more important than ever to reflect on the broader purpose of such systems. While much focus has been placed on the potential for NLP to improve general health outcomes (e.g., diagnostic accuracy, efficiency of care), there also exists an understated opportunity and responsibility to use this technology to promote health equity. How exactly do we accomplish this given the inherently complex and multifaceted nature of health disparities?

In this dissertation, I identify and make progress along two orthogonal dimensions for promoting health equity using NLP. As the first dimension, I consider the development of tools that leverage NLP to augment or supplant (potentially biased) decision making in health care. In this setting, there exists a need for "defensive"

ABSTRACT

methods that ensure NLP systems behave robustly across populations (e.g., patient populations, time periods). We ask and answer two questions – 1) how can we identify and measure sampling induced artifacts that arise in health-oriented training data, and 2) how can we counteract the effects of these artifacts or of more sweeping distribution shifts to promote model robustness? As the second dimension, I consider the development of tools that leverage NLP to detect or measure implicit bias in health care. Here, there is a need for "proactive" methods that translate hypotheses regarding implicit bias and discrimination into machine learnable tasks. We ask and answer: to what extent do these translations align with the broader pool of NLP research concerning implicit bias?

Primary Reader and Advisor: Mark Dredze

Secondary Readers: Anqi Liu & Emma Pierson

Dedication

For my grandparents – Laura, Dick, Charles, and Frances.

Acknowledgments

In all honesty, this dissertation looks nothing like I imagined it would when I started my PhD in 2019. Some of the differences are undeniably linked to the series of monumental world events that have transpired since then – a global pandemic, an attempted insurrection, multiple military conflicts, and a new civil rights movement. However, most can be attributed to the tremendous impact that my social, familial, and professional network has had on me. With that in mind, I would be remiss not to take this opportunity to thank them for their support and guidance.

First and foremost, I would like to thank my PhD advisor, Dr. Mark Dredze, who constantly encouraged me to take intellectual risks and strive for perfection, all the while tolerating my sarcastic comments and half-baked memes. One can never know for certain what life may have been like with a different advisor, but I can say quite confidently that I likely wouldn't be writing this acknowledgement section today. Mark's compassion and empathy allowed me to keep moving forward in my

personal life without ever feeling a sense of guilt for putting work on the back burner. And his ability to talk through adversity and offer creative solutions to my problems kept me from losing motivation to continue doing the technical work that brings me joy. I'm not sure I would recommend pursuing a PhD to most people, but to those dead set on doing so, I would recommend having Mark lead you through it.

Next, I'd like to recognize my committee members – Dr. Anqi Liu and Dr. Emma Pierson – for their valuable contributions to the formation of this document. Dr. Liu and Dr. Pierson's enthusiasm for this thesis kept me motivated throughout the writing process whenever I found myself struggling to write a coherent sentence. And perhaps more tangibly, their questions during my defense and thoughtful feedback thereafter were extremely helpful in identifying gaps in my logic and offering alternative perspectives to my own thought processes.

Indeed, having the opportunity to learn from and work alongside experts across a wide range of disciplines is something I feel incredibly privileged to have experienced. It started the moment I stepped on the Johns Hopkins campus, with the Once Upon a Time psychiatric dashboard team – Dr. Leslie Miller, Dr. Margaret Chisholm, Dr. Peter Zandi, Tenzin Lhaksampa, and Alex Walker – pushing me to think about and confront the practical challenges of deploying a computational system in a setting as sensitive as mental health care. It then continued with my collaborators at the

Wilmer Eye Institute – Dr. Cindy X. Cai, Dr. Tina Tang, and Anthony Gonzales – enabling me to spearhead efforts to leverage NLP in the fight against health disparities in Ophthalmology. Finally, it concluded with the Hidden in Plain Sight team – Dr. Mary Catherine Beach, Dr. Somnath Saha, Dr. Brant Chee, Dr. Aya Zirikly, Yahan Li, Anne Links, and Alya Ahmad – who provided me with the framework and context necessary to build an NLP system that has the potential to transform how healthcare providers interact with their patients. Along the way, Dr. John Ayers and Dr. Mathias Unberath helped me optimize my academic output and manifest a future that would bring me happiness.

Nevertheless, the majority of my graduate school experience was not spent with faculty or collaborators, but rather academic peers that made my day-to-day responsibilities as a PhD student enjoyable. The Westworld Squad – Carlos, Liz, Rachel, Alexandra, Nate, Abbey, Isabel, and Chris – stuck by me through several months of social-distancing and multiple out-of-state moves. Aaron and David made me feel less lonely amongst the chaos that is New York City. And Adam provided me with perspective in the times I needed it most. I feel grateful that I can call all of these people my friends.

Of course, before I could meet such an amazing group of friends and learn from the many experts mentioned above, I required help from a myriad of mentors. My

professors in the Department of Mathematics at Northeastern – Dr. Solomon Jekel and the late Dr. Christopher King – renewed my love of numerical problem solving. My undergraduate advisor, Dr. Dagmar Sternad, instilled in me a propensity to conduct thorough and meaningful research. And my colleagues at Applied Analytics in Boston – Dr. Nathan Sanders, Dr. Jonathan Foster, Arjun Sangvhi, and Abi Dawson – showed me how to apply these principles to real-world problems.

Last, but certainly not least, I must thank my incredible family for providing love and support when I found myself overwhelmed by the responsibilities of the PhD. My parents, Kriss and Gary, always knew exactly what to say when my mind was filled with self-doubt. My sibling, Nikki, kept me humble and inspired me to advocate for myself. My grandmother, Laura, allowed me to feel connected to my family in California through her letters and phone calls. And my wife, Abbey, always reminded me to live in the moment.

At times, a PhD can feel like an uphill battle with an army of one. However, when I look back at the path up the mountain, I realize just how much of a team effort went into reaching this point. To those I mentioned here and the several others I met along the way, I want to say thank you.

Contents

List of Tables

List of Figures

Part I: Introduction

Chapter 1

Introduction

1.1 Motivation

While we are far from having a cure for every disease or public health epidemic, it is rare that a dearth of scientific or medical knowledge is specifically cited as the bottleneck to maximizing human wellness. Instead, we more commonly call out issues with the allocation and quality of critical resources. For instance, we think of individuals dealing with prolonged wait times to see a trained healthcare professional (Lee et al., 2020a), or we think of individuals misdiagnosed by healthcare providers that were hesitant to believe their testimony (Plaza, 2020; Wiegand et al., 2023). Such failures are an undeniable blight on society when considered in isolation. However, they become even more reprehensible when one recognizes that they are concentrated disproportionately in certain groups of the broader population.

Whom receives the "bad" and whom receives the "good" in health care is almost ubiquitously tied to systemic and structural forms of discrimination (Braveman, 2006; McCartney et al., 2019). As a consequence, we find contemporary society filled with widespread disparities primarily affecting historically marginalized groups (Jackson et al., 2016; Ng et al., 2019). These disparities span a broad range of health outcomes – e.g., life expectancy (Dwyer-Lindgren et al., 2022), maternal mortality (MacDorman et al., 2021), substance use disorders (Buka, 2002), and self-injury (Blosnich and Bossarte, 2012). They are a significant burden on the global economy (Waidmann,

2009; Brott et al., 2011), and, perhaps more importantly, represent a tremendous moral failure on society's behalf (Powers and Faden, 2006; Jones, 2010). How can all individuals assume their unalienable right to life, liberty, and the pursuit of happiness if some individuals are systematically plagued by deleterious circumstances out of their control? As humans, we not only have a *functional need*, but also a *responsibility*, to address health disparities.

Unfortunately, most traditional strategies to date have been unable to make significant headway towards ameliorating disparities in health care (Zimmerman and Anderson, 2019; Jatoi et al., 2022), while also being extremely costly (Aluko et al., 2023). At the same time, artificial intelligence (AI) methods such as natural language processing (NLP) have been gradually revolutionizing health care, albeit with a predominant focus on improving health outcomes generally (as opposed to improving outcomes for specific groups) (Pakhomov et al., 2006; Pivovarov et al., 2015; Wei and Denny, 2015). Indeed, one does not need to look long or far to see the positive effect that NLP systems have already had – e.g., expediting biomedical literature retrieval during the COVID-19 pandemic (Chen et al., 2021), facilitating cohort selection for clinical trials (Yuan et al., 2019), and identifying adverse drug reactions in social media posts (Nikfarjam et al., 2019). Now, with language models approaching human levels of performance in natural language understanding and generation tasks (e.g.,

Achiam et al. (2023) and Touvron et al. (2023)), NLP's integration into the healthcare domain is likely to become even more rapid. Accordingly, we find ourselves at a crossroads where it is perhaps more important than ever to reflect on the broader purpose of these health-focused NLP systems.

1.2 Contributions

The lack of work thus far to leverage NLP in the fight to promote health equity belies the potential for doing so. In particular, NLP presents an opportunity to address disparities in health care by either increasing the supply and quality of resources (e.g., knowledge, services) to populations impacted by socioeconomic barriers, or by combating biased human decision making directly. Examples of this include the following:

- A dialogue agent that helps triage symptoms for members of an impoverished community while they wait to see a healthcare provider;
- A system that monitors social media posts for signs of mental distress within a population that doesn't have consistent access to mental health clinics;
- A model that measures a healthcare provider's use of language indicating doubt and highlights covert biases of which they were not previously aware.

Nevertheless, achieving an objective as grand and challenging as promoting health equity first requires us to answer several key questions. For example, do current systems benefit individuals regardless of their identity or socioeconomic background, or do they only function appropriately for individuals in communities from which their training data is drawn? Do current systems generalize across time periods and clinical specialties, or do they fail in a spectacular fashion when presented with new data distributions? Do current systems address systemic disparities in health care, or do they actually exacerbate them? And if current systems are not benefiting everyone equally, what can we do about it?

The aforementioned questions are admittedly not particularly new. In fact, concerns regarding the equitable impact of AI systems in health care and beyond have existed for nearly as long as the systems themselves (Firschein et al., 1973; Thomasian et al., 2021). At the same time, these questions are certainly not closed-ended, nor does a single catch-all solution suit them. The landscape of NLP in health care is wide and evolves constantly as new methods, resources, and challenges arise. In fact, the only thing that remains fixed is the often understated responsibility and opportunity to use NLP as a tool for promoting health equity.

In practice, the latter is complicated by the inherently complex nature of health disparities. Achieving health equity requires the deconstruction and reparation of

institutional structures that have existed for tens to hundreds of years (Braveman, 2006; Braveman et al., 2011). To be very clear, in no way is this thesis intended to suggest that such injustice can be rectified by an algorithm, or in absence of collaboration with the very individuals who have been subject to wrongdoing. It is however intended to recognize and respond to the need for multifaceted efforts to arrive at an comprehensive solution. This thesis focuses specifically on two angles in which NLP may be used to mitigate inequality and promote equity in health care.

1.2.1 Defensive Tactics For Promoting Equity

As the first angle, I consider the development of tools that leverage NLP to augment or supplant (potentially biased) decision making in health care. In this setting, there exists a need for "defensive" methods – those which ensure systems leveraging NLP behave robustly across different populations (e.g., patient populations, time periods). At worst, we do not want systems to exacerbate or propagate existing disparities. And at best, we want to eradicate them altogether. Towards this end, I explore two related lines of work.

Measuring Data Bias. The first line of defensive work focuses on identifying and measuring sampling-induced artifacts in training data that may inhibit NLP model robustness. Specifically, I propose methods that practitioners can use to examine the

quality of datasets that contain non-clinically derived diagnostic annotations. My techniques reveal novel temporal artifacts and other systemic data quality issues that arise due to a combination of self-disclosure biases and other constraint-motivated data sampling decisions common within such datasets.

Counteracting Data Bias. The second line of defensive work focuses on counteracting the effects of sampling-induced artifacts and more sweeping forms of distribution shift. In particular, I introduce a new technique that leverages measurements of semantic shift between two language distributions to perform robust feature selection. I also examine the efficacy of clinical language model pretraining using out-of-domain clinical language data, obtaining results that offer alternative views to prevailing perspectives in the computational health literature regarding language model robustness and clinical data heterogeneity.

1.2.2 Proactive Tactics For Promoting Equity

As the second angle, I consider the development of tools that leverage NLP to detect and measure implicit bias in health care. These "proactive" approaches require careful translation of hypotheses regarding implicit bias and discrimination into machine learnable tasks. To what extent do these translations align with the broader pool of NLP research concerning implicit bias in language? And to what

extent do learnings regarding implicit bias generalize across different provider and patient populations? I answer these questions in the context of developing a system to characterize stigmatizing language in medical records. Not only do I highlight differences in the usage of biased language between clinical and non-clinical domains, but also differences in the usage of biased language across clinical domains.

1.2.3 Comparison of Tactics

It is true that proactive approaches for promoting health equity may require defensive techniques to maximize their efficacy. Likewise, defensive approaches for promoting health equity could arguably be seen as proactive relative to the status quo. For the purpose of this thesis, we delineate defensive and proactive approaches on the basis of their intention to counteract systemic and structural discrimination. Proactive methods are designed and deployed specifically with the goal of promoting health equity, whereas defensive methods are focused on ensuring disparities are not exacerbated or introduced. The difference between defensive and proactive methods is akin to the difference between an individual not being racist and an individual being anti-racist (King and Chandler, 2016; Cerdeña et al., 2020).

1.3 Structure

The structure of this thesis roughly parallels the development cycle of an NLP system that strives to promote equity in health care. Although not done intentionally at the time, it also largely parallels my own journey of knowledge discovery in the quest to use artificial intelligence and machine learning for social good.

Chapter 2 provides a high-level overview of health equity for readers who are less familiar with the concept. I also use this as an opportunity to make explicit the ways in which natural language processing can reduce health disparities.

Chapter 3 delves into the technical aspects of natural language processing that have the potential to exacerbate existing forms of inequity in health care (e.g., selection bias, distribution shift), as well as the classes of methods that attempt to mitigate such risks (e.g., domain adaptation).

Chapter 4 introduces my first original piece of research, a comprehensive annotation and analysis effort of over 100 social media text datasets constructed originally for the purpose of modeling mental health status (e.g., depression, PTSD). I use this review primarily as an opportunity to contextualize the data-related concerns presented in Chapter 3 and more specifically detail issues that commonly plague dataset curation in the healthcare domain. This chapter marks the end of the thesis' introduction and beginning of technical scientific contributions.

Chapters 5 and 6 present the first line of "defensive" work – measuring data bias. In these chapters, I introduce and validate methods for measuring and understanding data-centric barriers to achieving NLP model robustness (i.e., generalization) in the healthcare domain. I target issues that arise particularly as a consequence of regulatory constraints (i.e., privacy laws), difficulty acquiring annotations at scale (i.e., requirement of expert annotators), and human-factors commonly associated with health-related modeling tasks (i.e., heterogeneous clinical presentations). Both chapters consider the task of detecting depression in social media users as the specific use case.

Chapters 7 and 8 focus on the second line of "defensive" work – counteracting data bias. In Chapter 7, I introduce a new feature selection method that leverages knowledge of semantic shift to improve generalization over time. I continue using the task of detecting depression in social media users as the guiding example. In Chapter 8, I conduct a novel evaluation of clinical language model domain transfer and use these results to inform recommendations for future clinical language model development in the era of large language models. Here, I move away from social media data and instead consider the task of phenotyping in electronic health records.

Chapter 9 shifts gears further and details a proactive approach to promoting equity in health care, namely identifying and characterizing stigmatizing language

in electronic health records. I leverage knowledge and techniques from the former chapters to define appropriate boundaries for the use of the NLP models developed as part of the study. I then deploy these models on a widely adopted public clinical dataset, MIMIC-IV, and provide evidence of language that may allow clinical language models to exacerbate existing disparities.

Chapter 10 concludes the thesis by summarizing the lessons learned throughout this original research, and by enumerating opportunities for future exploration.

1.3.1 Publications

The following publications serve as the primary focus in each of the aforementioned content chapters:

- §4: "On the State of Social Media Data for Mental Health Research" (Harrigian et al., 2021)
- §5: "Do Models of Mental Health Based on Social Media Generalize?" (Harrigian et al., 2020)
- §6: "Then and Now: Quantifying the Longitudinal Validity of Self-Disclosed Depression Diagnoses" (Harrigian and Dredze, 2022b)
- §7: "The Problem of Semantic Shift in Longitudinal Monitoring of Social Media:

A Case Study on Mental Health During the COVID-19 Pandemic" (Harrigian and Dredze, 2022a)

- §8: "An Eye on Clinical BERT: Investigating Language Model Generalization for Diabetic Eye Disease Phenotyping" (Harrigian et al., 2023a)
- §9: "Characterization of Stigmatizing Language in Medical Records" (Harrigian et al., 2023b)

Where applicable, I also refer throughout the thesis to related work that I co-authored with collaborators also interested in promoting health equity using NLP. These include the following:

- "Gender and Racial Fairness in Depression Research using Social Media" (Aguirre et al., 2021)
- "Towards Understanding the Role of Gender in Deploying Social Media-Based Mental Health Surveillance Models" (Sherman et al., 2021)
- "Health Disparities in Lapses in Diabetic Retinopathy Care" (Cai et al., 2023)
- "Managing HIV during the COVID-19 pandemic: a study of help-seeking behaviors on a social media forum" (Ayers et al., 2024)

- "Recent Advances, Applications, and Open Challenges in Machine Learning for Health: Reflections from Research Roundtables at ML4H 2023 Symposium" (Jeong et al., 2024)
- "Improving the identification of diabetic retinopathy and related conditions using natural language processing methods" (Cai, Cindy X et al., 2024)
- "Are Clinical T5 Models Better for Clinical Text?" (Yi et al., 2024)

Chapter 2

What is Health Equity?

2.1 Case Study

Consider the following scenario: Ava, Barbara, and Christine were each experiencing their first pregnancy. Until the second trimester, each individual faced only minor complications (e.g., morning sickness). At that point, major bouts of fatigue, soreness, and nausea started affecting them almost daily. Ava and Barbara ultimately experienced a preterm birth that caused their children to stay in the neonatal intensive care unit (NICU) for two months. Christine carried to full term, delivered without any complications, and welcomed home a typically developing child. Do the differences in maternal outcomes that Ava, Barbara, and Christine experienced reflect a health disparity?

The answer to such a question is not necessarily straightforward. In isolation, two, three, or even thousands of people experiencing a sub-optimal health outcome does not necessarily indicate that a health disparity exists. Many of life's peaks and valleys can be attributed to simple stochasticity. For Ava and Barbara, randomly occurring genetic mutations or procedural complications during child birth could have caused the adverse maternal outcomes (Skoogh et al., 2021; Mead et al., 2023). What elevates a disparate health outcome to the level of a health disparity is the *context* in which it emerges.

With that in mind, let us learn more about the situations of Ava, Barbara, and Christine.

- Ava is a 29 year old, Queer, African-American woman with an income that puts her on the lower end of the middle-class. As a bank teller, Ava is required to stand for the majority of her shift. She considered going to her OB-GYN's office to seek help for the symptoms, but ultimately couldn't leave work before 5pm when the office closed. She could not afford to take an additional day of unpaid time off. Ava later reported feeling more stressed than usual at work due to an upcoming round of layoffs.

- Barbara is a 32 year old non-Hispanic, White woman in a heterosexual relationship. She has an Associates degree from community college, but her income puts her just above the poverty line. Barbara has more flexibility and stability in her work than Ava, but lacks health insurance and can only cover the cost of attending a local community clinic run by primary-care physicians. The physician Barbara saw downplayed the significance of her symptoms and opted not to do additional testing. Prenatal tests during her hospitalization for the birth revealed that she was suffering from high blood sugar (i.e., gestational diabetes).

- Christine is a 31 year old Native American woman in a heterosexual relationship. Her income puts her firmly in the upper class. Christine didn't deal with any of the barriers Ava and Barbara faced – her employer not only fully covered the cost

of premium health insurance that allowed her to visit nearly any OB-GYN, but also supported 1 year of paid parental leave. Christine's OB-GYN recognized the symptoms as unusual and, after running multiple tests, identified that she was also suffering from high blood sugar. Christine was put on a treatment plan that helped manage the disease.

Ava's stress at work (Ruiz and Avant, 2005; Lautarescu et al., 2020) and Barbara's high blood sugar (Yogev and Langer, 2007) made them each more likely to experience a preterm birth. While Christine also dealt with gestational diabetes, the condition was caught by her OB-GYN and she was able to manage it with appropriate treatment. Each individual expressed and deserved an equivalent need for care, but only one was able to satisfy that need.

One may presume that this mismatch between supply and demand of health resources constitutes a health disparity. However, even this juxtaposition by itself is not enough to indicate that a difference in health outcomes represents a health disparity. So, what does determine whether Ava, Barbara, and Christine's situation represents a health disparity? Let us formally define health equity, and then return to the question.

2.2 Defining Health Equity

Health equity as a scientific concept has been formalized for several decades now, albeit with some meaningful variation in definition (Pereira, 1993; Organization, 2000; Bambas, Casas, et al., 2001; Smith, 2015). The widely accepted definition in contemporary culture is presented by Braveman et al. (2018), who defines health equity in two parts. The first part states:

> *Health equity means that everyone has a fair and just opportunity to be as healthy as possible. Achieving this requires removing obstacles to health – such as poverty and discrimination and their consequences, which include powerlessness and lack of access to good jobs with fair pay; quality education, housing, and health care; and safe environments.*

This part of the definition is heavily inspired by Whitehead (1992), who defined inequities as differences that are "not only unnecessary and avoidable, but in addition, are considered unfair and unjust." Whitehead (1992) goes on to say that, "Equity is therefore concerned with creating equal opportunities for health, and with bringing health differentials down to the lowest level possible." This in turn involves "equal access to available care for equal need, equal utilization for equal need, [and] equal quality of care for all."

The "fair and just opportunity" phrase in the Braveman et al. (2018) is intended

to capture the "unnecessary and avoidable" differences referenced in the Whitehead (1992) definition. By these definitions, differences in health outcomes that happen by random chance or by an individual's own choice – e.g., participation in injury-prone pastimes – do not qualify as a possible health disparity. Only differences that are at least theoretically rectifiable given current scientific and medical knowledge qualify. That said, it is not necessary to understand the cause of the differences *a priori* – the action to address the health disparity may simply be the allocation of resources to investigate and understand the cause (Braveman, 2006).

The second part of the definition reflects observations from other scholars regarding the need for health equity/disparities to be measurable and actionable, as well as the desire to emphasize the definition's basis around disadvantaged groups (Murray et al., 1999; People, 2000; Health et al., 2006):

> *For the purposes of measurement, health equity means reducing and ultimately eliminating disparities in health and in the determinants of health that adversely affect excluded or marginalized groups.*

Here, "excluded or marginalized groups" refers specifically to socially, economically, or culturally disadvantaged subsets of the broader population that have been subject to structural or systematic forms of discrimination (Braveman, 2006). While differences in health outcomes between population subgroups may be of interest from a public

health or epidemiological perspective, they only affect overall "health equity" if the subgroups involved are those which are historically disadvantaged. For example, men having a lower life expectancy than women is worth investigating and, if possible, ameliorating. However, it does not constitute a health disparity by definition because men are the historically advantaged group. Likewise, two socioeconomically-similar cities experiencing different rates of cancer may be a public health concern, but is not a health disparity.[1]

The slate of measurable outcomes that may qualify as a health disparity is essentially infinite. There does not exist any inclusion criteria with respect to the severity of the health outcome (e.g., prevalence of papercut injuries vs. prevalence of breast cancer); even seemingly innocuous health issues can spiral into much larger society-wide disparities (Heckman and Britton, 2015; Sen, 2021). Moreover, qualifying outcomes do not necessarily need to be related directly to physical or mental health. For example, differences in literacy rates or access to community centers within a certain radius may qualify as health disparities given their indirect effect on employment and social support, respectively. Health *care* is just one piece, albeit a significant piece, of the larger health landscape.

[1] Assuming the two cities have essentially equivalent populations, economics, and social practices.

2.3 Case Study Revisited

With the definition of health equity in hand, we can now return to our example scenario involving Ava, Barbara, and Christine. In isolation, the experiences of three individuals is admittedly not enough to conclusively identify the presence of a health disparity. However, if we instead think of Ava, Barbara, and Christine as representatives of population groups sharing similar socioeconomic characteristics, then we can conclude that a health disparity exists.

First, one could reasonably argue that Ava's status as an African-American woman and minority sexuality have influenced her socioeconomic position (Berg and Lien, 2002; Akee et al., 2019; Flage, 2020), which in turn made it difficult to take additional unpaid time off to attend an unplanned OB-GYN appointment. Indeed, financial insecurity remains a major burden for marginalized groups and is frequently cited as a reason for avoiding necessary health care (Mutchler et al., 2017; Weida et al., 2020).

Next, while Barbara's race and ethnicity do not suggest her outcome is representative of a health disparity, her socioeconomic status and access to insurance do. It is possible that the generalist physician that Barbara saw was not familiar enough with her particular clinical presentation to recognize that she was at risk of experiencing a pre-term birth. Moreover, it is also possible that Barbara's treatment was influenced by her provider's unconscious bias against patients that must pay for

care out-of-pocket.

Finally, Christine's positive pregnancy experience highlights that not all members of a historically disadvantaged group, or even the disadvantaged group on the whole, will necessarily be subjected to a particular health disparity. It is possible for a health disparity to affect one marginalized group (or a subset of a marginalized group), but not others. More generally, it is recommended that disparities are measured using historically advantaged groups as the reference point (Anand et al., 2001; Keppel et al., 2005; Braveman, 2006). While it may not always be the case that these groups experience the best of a particular health outcome, it is true the overwhelming majority of the time (Williams, 2012). Counterexamples fall into the category of outcomes that are a public health concern, but not necessarily a health disparity. More recently, scholars have argued for refining reference marginalized groups to reflect intersectionality (Guan et al., 2021; Homan et al., 2021).

2.4 The Role of NLP in Promoting Health Equity

Health equity may only be achieved by eliminating the underlying causes of health disparities, which more often than not are related to deeply-ingrained systemic

and structural issues (e.g., racism, elitism, sexism, homophobia, xenophobia, etc.) (Williams, 2005; Baciu et al., 2017a). In practice, such goals are nearly impossible to achieve or, in the least, will take a long time to achieve. Current efforts to reduce disparities across a wide array of health dimensions have unfortunately been both ineffective and quite expensive (Zimmerman and Anderson, 2019; Jatoi et al., 2022; Aluko et al., 2023). The small number of efforts that have been successful have generally been narrow in scope (e.g., a single community or health system) (Haas et al., 2015; Torres-Ruiz et al., 2018; Gonzalez et al., 2021) or difficult to scale (e.g., one-on-one coaching) (Halladay et al., 2013).

Natural language processing (NLP) has tremendous potential to transform the state of disparity-reduction efforts if used appropriately and deliberately. On one hand lies the opportunity to use NLP to address issues related to **access and quality of health care**. In our running example, this may take the form of a OB-GYN-specific chatbot that allows Ava to triage her symptoms outside of work hours, or alternatively a literature retrieval and summarization system that better informs maternal care provided by Barbara's general practitioner. On the other hand lies the opportunity to use NLP as a mechanism for **uncovering discrimination and unconscious bias** within health care. As we will explore later in this thesis, perhaps the dismissive tone used by Barbara's provider could be identified to highlight a prejudice to the

provider of which they were not acutely aware. Knowledge of this trait may then inspire the provider to approach patients having similar backgrounds as Barbara with additional empathy. In both use cases, NLP offers the advantage over traditional equity-interventions of being data-driven and straightforward to scale.

At the same time, there remains a need to ensure any NLP system used for the purpose of improving health equity behaves fairly and robustly across populations. In some cases, technical complexity can provide support towards achieving such goals. However, it is important to not conflate complexity with utility. It would not make sense to spend millions of dollars developing an app that educates prospective mothers about the complications of pregnancy if there is no plan to support the individuals in need of care. Likewise, it would not make sense to train a language model to answer questions about pregnancy if it weren't likely to work for the populations who need it most. In the next chapter, we will review the underlying reasons that make it difficult to achieve robustness and highlight some common strategies for combating such challenges.

Chapter 3

Technical Challenges in Promoting Health Equity

Researchers interested in using NLP and other forms of AI to improve health equity are frequently the same researchers concerned about the risk of such technologies worsening health equity (Zhang et al., 2017; Adamson and Smith, 2018; Veinot et al., 2018; Chen et al., 2020; Ibrahim and Pronovost, 2021; Röösli et al., 2021; Celi et al., 2022; Thamman et al., 2023). In general, these fears stem from two undesirable scenarios – 1) propagating and exacerbating existing health disparities, and 2) introducing completely new health disparities altogether.

3.1 Risk of Propagating Disparities

The first scenario refers to the situation in which an algorithm learns to mimic behavior that is already systematically biased or discriminatory in nature. As an automated technology capable of being applied at scale, an algorithm's negative impact may quickly surpass that of the behavior it originally learned to replicate. For example, consider a hypothetical situation in which a historically marginalized community experiences difficulty accessing traditional forms of mental health care and in turn receiving diagnoses that enable them to collect social welfare benefits. In response, an individual may consider training a chat bot on historical provider-patient interactions to offer diagnoses in place of a human mental health professional. If humans systematically discriminate against the marginalized community (e.g., underdiagnose

a condition, question a patient's credibility), then an algorithm trained to replicate existing human behavior will only strengthen the position of the already advantaged community. This situation can be difficult to address from both a technical and moral perspective.

From the technical perspective, guarding against the exacerbation of existing disparities requires training models such that they explicitly ignore selective aspects of their training data. This issue has been explored extensively outside health applications – e.g., removing gender bias in language models (Bordia and Bowman, 2019; Sun et al., 2019b), counteracting racial bias in models used to screen resumes (Deshpande et al., 2020), learning to ignore group-specific predictors in general classification settings (Sagawa et al., 2019). Unfortunately, the efficacy of "debiasing" models in these applications tends to vary quite significantly (Gonen and Goldberg, 2019; Zhang et al., 2022). Moreover, not all health applications have a clear cut "ideal" target distribution that can be used to guide the debiasing process. For example, several outcomes in healthcare have prevalences that have been measured to vary across demographics (Akhtar-Danesh and Landeen, 2007; Colleran et al., 2007), but it not always clear whether these differences reflect a physiological difference that should be preserved (e.g., comorbid conditions) or a social difference (e.g., underreporting, unconscious prejudice) that should be rectified.

Indeed, ambiguity in systemic biases and technical limitations on the type of systemic biases that can be addressed (Schrouff et al., 2022) places an onus on the model developer to cast a value judgement regarding what aspects of training data are problematic from an equity perspective. What happens in the mental health chat bot example above if two historically marginalized communities are subject to systemic issues, but a practitioner is only able to successfully ameliorate issues for one of the marginalized communities? Is it morally acceptable to deploy the resulting model? Should the practitioner attempt to reduce the efficacy of the debiasing method for one community if it increases the efficacy of the debiasing method for the other community? Such questions are often best left for philosophers and ethicists instead of machine learning and NLP practitioners.

In lieu of attempting to directly ameliorate existing social biases themselves, technologists may instead focus on identifying systemic disparities that are undesirable to perpetuate. For instance, one may use NLP to extract social determinants of health (SDoH) from an electronic medical record and feed the output into an analysis regarding clinical outcomes (Patra et al., 2021). Alternatively, one may measure disparities in the degree of empathy expressed in responses to patient questions in an online health forum as a function of the patient's demographics (Nobles et al., 2020). As I will show later in this thesis with the development of a system for characterizing

stigmatizing language in medical records (Chapter 9), NLP does not always need to be the solution for directly eliminating bias; it can instead be used to highlight disparities that weren't already known or measured.

3.2 Risk of Introducing Disparities

The second scenario of concern to NLP practitioners interested in improving health equity is one in which a model or system introduces a disparity that didn't already exist or re-introduces a disparity that was better managed by current practice (e.g., countering bias in surgical decision making using biological and physiological data (Hisam et al., 2016)). These concerns are generally well-motivated given the multitude of infamous cases in which ML and NLP models have behaved suboptimally or in a discriminatory manner for marginalized communities – e.g., COMPAS systematically assigning Black people a higher risk of recidivism than White people (Flores et al., 2016; Dressel and Farid, 2018), AI hiring software rejecting qualified minority candidates over similarly-qualified majority group candidates (Kodiyan, 2019; Chen, 2023), facial recognition and classification models performing worse for certain genders and skin colors in a variety of settings (e.g., criminal surveillance, emotion recognition) (Buolamwini and Gebru, 2018), and language models showing distinct gender and race biases when referencing some occupations (Kotek et al., 2023). Similar failures

have also occurred in the health domain – e.g., predictive models wrongfully cutting off insurance payouts for medicare patients (Mello and Rose, 2024), and COVID-19 chest x-ray classifiers performing worse for particular age groups (Santa Cruz et al., 2021; Arias-Garzón et al., 2023).

The ubiquity of these issues may suggest that ML and NLP models being unfair or even discriminatory is an inevitability. In truth, the fundamental mechanisms underlying such behavior are actually relatively well understood by the research community. Models may learn non-generalizable relationships at training time that negatively affect performance in some populations at test time due to distribution shift. Addressing these issues requires us to first detect and measure their presence, and then strategically counteract their effect. Methods for doing so that are both practical and actually effective on real, non-synthetic data are the focus of Chapters 5 through 8 of this thesis.

3.2.1 What Causes Models to Fail?

NLP models have the potential to introduce or re-introduce health disparities when 1) they perform differently for different groups, and 2) the difference in performance has an effect on downstream health outcomes. Differences in performance across groups may be purely quantitative (e.g., classification accuracy, probability calibration)

or something more abstract (e.g., tone of a dialogue agent). Differences in model performance should not be equated with differences in model behavior; there are cases in which models arguably should behave differently for different groups to achieve an equitable downstream outcome (e.g., personalized education, different treatment recommendations based on lifestyle factors) (Arslan, 2023; Dumont and Ready, 2023). The main point of the aforementioned statement is that inequity arises when there exists a systematic difference in the *quality* of a model's outputs across groups, where quality may itself be independently defined by each group.

So, what causes these differences? For the sake of our discussion, let us define a model M to be a function having parameters θ such that for each input $x \in \mathcal{X}$, M uses θ to transform x into an associated output $y \in \mathcal{Y}$. Succinctly, we have $M(x;\theta) : \mathcal{X} \to \mathcal{Y}$. The parameters θ may be defined manually by a practitioner (e.g., heuristics based on domain knowledge) or learned from data. In either case, we assume that θ has been chosen based on a finite amount of historical evidence drawn from a distribution $p(x, y)$ for the purpose of facilitating a transformation from x to y. Informally, we can say that M's "performance" is defined by its ability to transform x to y, where y is the correct and/or desired output for a given input x (e.g., accurate prediction, context-appropriate generation). We are generally not interested in a model's performance on the data used to choose θ, but rather its performance on

an unseen set of data – i.e., how well the model generalizes. We say that a model generalizes if it achieves an equivalent level of performance on the unseen test data as it does on the data used for training (within a reasonable margin of variation).

Distribution Shift. The ability for a model to generalize to unseen data depends on the degree of alignment between the distribution used for learning θ and the test distribution from which the unseen data is sampled (Elsahar and Gallé, 2019; Garg et al., 2021). Informally, any degree of misalignment between training and test distributions is referred to as *distribution shift*. Formally, we say that distribution shift has occurred from one joint distribution $p(x_a, y_a)$ to a different joint distribution $p(x_b, y_b)$ if $p(x_a, y_a) \neq p(x_b, y_b)$. An inequality between two joint distributions can arise for three possible reasons, each of which we denote as a specific form of distribution shift.

Covariate Shift. Formally, $p(x_a) \neq p(x_b)$, but $p(y_a \mid x_a) = p(y_b \mid x_b)$. Informally, we say there is a systematic change in the distribution of inputs, but the relationship between the inputs and outputs remains the same.

Prior Shift. Formally, $p(y_a) \neq p(y_b)$, but $p(x_a \mid y_a) = p(x_b \mid y_b)$. Informally, we say that there is a systematic change in the distribution of the outputs, but the relationship between inputs and outputs remains the same. Prior shift is also referred to as *Label Shift*.

Concept Shift. Formally, $p(y_a \mid x_a) \neq p(y_b \mid x_b)$, but $p(x_a) = p(x_b)$ (in $x \rightarrow y$ problems). Alternatively, $p(x_a \mid y_a) \neq p(x_b \mid y_b)$, but $p(y_a) = p(y_b)$ (in $y \rightarrow x$ problems). Informally, we say that the relationship between the inputs and outputs has changed.

Selection Bias. Each type of distribution shift described above can be viewed as a consequence of *selection bias* (Moreno-Torres et al., 2012; Wan et al., 2022). At a high level, selection bias denotes any systematic error in the outcome of a study that results from an investigator's decisions for curating the study's underlying dataset (Tripepi et al., 2010).[1] The consequence of these errors is the curation of a dataset (i.e., training data) which is not representative of the population it is intended to capture (i.e., test data).

Zadrozny (2004) was the first to formalize selection bias in the context of machine learning. They posit a situation in which instances (x, y, s) are drawn independently from a probability distribution $\mathcal{D}$ with domain $\mathcal{X} \times \mathcal{Y} \times \mathcal{S}$, where $x \subseteq \mathcal{X}$ is the feature space, $y \subseteq \mathcal{Y}$ is the label space, and $s \subseteq \mathcal{S}$ is a binary space for indicating whether an instance is selected. They then define selection bias as occurring when a sample from the space ($s = 1$) is not drawn at random from $\mathcal{D}$, but rather one of three scenarios

[1] This does not necessarily mean that the investigator made incorrect decisions at the time of the study. It is possible that a dataset was representative of the target population when it was collected, but later became non-representative due to temporal dynamics.

occur.

1. **Selection depends only the feature set:** $P(s \mid x, y) = P(s \mid x)$
2. **Selection depends only on the label:** $P(s \mid x, y) = P(s \mid y)$
3. **No independence assumptions are satisfied:** Selection cannot be explained purely as a function of the observed feature set or labels. Rather, it may be the case $P(s \mid x_s, x, y) = P(s \mid x_s)$, where x_s is an unobserved set of features which controls the selection.

In machine learning literature, the first and second scenarios are tied directly to notions of covariate shift (Sugiyama et al., 2006) and prior shift (Garg et al., 2020), respectively. The third scenario has the potential to introduce all three of the forms of distribution shift (Wan et al., 2022). Statisticians and those interested in causal inference may recognize the first and third settings as an issue of missing data – Missing At Random (MAR) and Missing Not At Random (MNAR), respectively (Little and Rubin, 2002; Bhattacharya et al., 2020).

To make the connection between distribution shift and selection bias explicitly clear, we can extend the context of the aforementioned examples. Directed Acyclic Graphs (DAGs) for each form of distribution shift and selection bias are visualized in Figure 3.1 to guide the expansion.

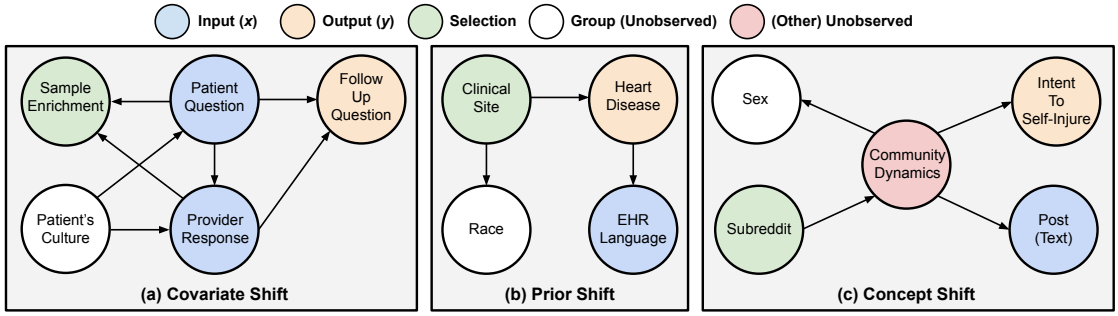


Figure 3.1: Examples of (a) covariate shift, (b) prior shift, and (c) concept shift caused by selection bias. Each scenario is visualized using a Directed Acyclic Graph (DAG). The joint distribution of observed variables $p(x, y)$ is subject to change depending on the selection outcome.

Covariate Shift. Consider developing a model to predict whether a clinician's response to a secure patient message will be sufficient, or alternatively will require an additional follow-up response. Such a model could allow clinicians to optimize their responses and reduce time-consuming back-and-forth with patients. We consider the scenario in which not all patient question and provider responses (x) meet criteria (s) that the practitioner deems relevant for inclusion in their effort to train a model to predict whether a follow-up question (y) will be received. For example, the practitioner may exclude data points in which provider responses are below a certain length, or exclude data points in which the patient question doesn't include a known medical concept (e.g., using it as a flawed proxy for administrative questions). We suspect the patient's culture influences how they frame questions to the provider (e.g., word choice) and

how the provider responds to them (e.g., succinct vs. detailed) (Duvall et al., 2021; Heisey-Grove et al., 2021; Armstrong et al., 2023). As shown in panel (a) of Figure 3.1, $s \perp\!\!\!\perp y \mid x$ because when x is observed, all paths through x from s are blocked by rules of d-separation. Thus, $p(s \mid x, y) = p(s \mid x)$. At the same time, it is not true that $s \perp\!\!\!\perp x \mid y$ because there is a direct path from $x \to s$. Therefore, $p(s \mid x, y) \neq p(s \mid y)$. If we deploy our model on data that isn't subject to the same exclusion criteria (s), the model will be subject to covariate shift. Informally, we can see that given s, the model would be trained in such a way that it wasn't exposed to specific types of questions and responses (x), which in turn could cause it to underperform for particular cultural groups (e.g., predicting that a response is sufficient when it actually isn't, and in turn causing disparate levels of frustration).

Prior Shift. Consider developing a model to estimate the risk that an individual will develop heart disease in a predetermined time frame based on their electronic medical records. We consider the scenario in which electronic health record data (x) and information about whether an individual will contract heart disease (y) is drawn from a non-random set of clinical sites (s). A clinical site's location and policies (e.g., insurance accepted) are likely to influence the patient population they see. As shown in panel (b) of Figure 3.1, $s \perp\!\!\!\perp x \mid y$ by rules

of d-separation, which implies $p(s \mid x, y) = p(s \mid y)$. However, it is not true that $s \perp\!\!\!\perp y \mid x$ because there is a direct path from $s \rightarrow y$, which in turn implies $p(s \mid x, y) \neq p(s \mid x)$. This causal scenario sets up the possibility that we deal with prior shift. Informally, we see that sampling non-randomly from certain clinical sites will expose us disproportionately to one outcome label more than another. If the clinics in our sample have a majority White patient population, we may ultimately find that the trained model systematically under-predicts the risk of heart disease when we apply it at clinics with a majority Black patient population (Graham, 2015).

Concept Shift. Consider developing a model to detect intent to self-injure in social media posts on Reddit. We consider the scenario in which social media posts (x) and labels regarding the presence of intent to self-injure (y) are drawn from a single online community (s) (i.e., subreddit). We may pull primarily from a single online community because we suspect that statements of suicidal intent are extremely rare outside of mental health support communities. We assume that the community's dynamics (e.g., culture, topicality) influence the distribution of sexes represented in the training data (e.g., majority Male). We also assume that the community's latent dynamics (d) influence the manner in which intent to self-injure is expressed (e.g., sarcasm, word-sense

disambiguation). We see in panel (c) of Figure 3.1 that in the absence of measuring the community's dynamics, there exists paths $s \to d \to y$ and $s \to d \to x$, meaning that both x and y are not independent of s. Had we been able to observe community dynamics in the feature set, we would have $s \perp\!\!\!\perp x, y \mid d$ which implies $p(s \mid d, x, y) = p(s \mid d)$ (i.e., the third selection bias scenario). Informally, we have $p(y \mid x, s) \neq p(y \mid x)$, meaning that the relationship $p(y \mid x)$ depends on the context provided by the community s. This indicates that the model trained on the sampled dataset is subject to issues with concept shift. It may underperform on a community that has a majority Female or Non-binary user base because the semantic relationship $p(y \mid x)$ is different from what was previously learned.

Theory vs. Practice. While it is often convenient to assume that distribution shift will not occur between training and test distributions (i.e., i.i.d. assumption), such an assumption is often far too strong when human language is involved. Language data from different groups of people or the even same group of people at different time points is rarely generated from perfectly equivalent distributions. In some cases, group-specific relationships emerge as a consequence of innate, immutable characteristics (e.g., genetic conditions, physiological traits) (Raza et al., 2021; Sirugo et al., 2021). In other cases, they emerge as a consequence of external influence

(e.g., culture, geography, economics, politics, interpersonal social dynamics) (Chen et al., 2006; Perales and Campbell, 2020). Generally, however, they emerge due to a confluence of multiple interdependent factors that aren't as straightforward to explain or model as the ones in the examples discussed previously.

Furthermore, while these types of phenomena may be relatively ubiquitous within a group, they are not necessarily guaranteed to be omnipresent. Intersectionality can confound the generality of the human experience (Hankivsky, 2022), in turn promoting additional subdivision with a single "group's" language distribution (Tan and Celis, 2019). The number of possible intersectional groups is infinite, given that individuals may identify with or experience varying degrees of membership in a particular group (Settles and Buchanan, 2014).

When these realities of humanity are juxtaposed with the practical challenges that arise when collecting data to train a machine learning or NLP model (e.g., finite sample sizes, cost, labor), it starts to become clear why distribution shift is such a concern to those interested in promoting health equity with NLP. To further drive home this point, Chapter 4 of this thesis serves as a case study on the ubiquity of sample selection bias that arises due to data acquisition constraints in health-related applications.

3.2.2 Detecting Distribution Shift

The first challenge presented by distribution shift and selection bias is detecting whether it exists in a given deployment scenario (i.e., train $\rightarrow$ test). When distributions of variables are parametric (e.g., Gaussian, Binomial), one can typically use off-the-shelf statistical tests taught during an "Introduction to Probability" course to estimate whether the training and test distributions are equivalent. Unfortunately, such tests are generally insufficient for our problem domain. Real language data is not only non-parametric, but also high-dimensional. These characteristics necessitate more clever methods for detecting whether distribution shift has occurred (Lavergne and Patilea, 2008; Balakrishnan and Wasserman, 2018).

Yang et al. (2021a) argue that several problem setups in machine learning (e.g., anomaly detection (Chandola et al., 2009; Pang et al., 2021), outlier detection (Boukerche et al., 2020), and open-set-recognition (Geng et al., 2020)) can be unified under the umbrella of out-of-distribution (OOD) detection. Here, the goal is to determine whether a new instance aligns with an existing distribution (e.g., a distribution used for training). Instances that are not determined to come from the existing distribution may be passed to models trained on data that better aligns with the instance (Suárez-Cetrulo et al., 2023), or directed to humans for manual review (Geifman and El-Yaniv, 2017; Kompa et al., 2021). Methods for making this

determination may use, amongst other attributes (Gama et al., 2014; Nair et al., 2019), an instance's distance from a model's decision boundary (i.e., uncertainty) in combination with an alerting threshold (Nicora et al., 2022), or alternatively a separate classifier that infers whether an input comes from a particular distribution (Hojjati et al., 2022).

Nonetheless, a single instance being detected as OOD may itself be an anomaly. To draw conclusions regarding the presence of a broader distribution shift, it is necessary to consider multiple instances simultaneously. We provide a high-level overview of such methods below based on the primary type of shift they target.

Covariate Shift. Covariate shift detection is unique from other forms of distribution shift detection in the sense that it does not explicitly require access to ground truth outputs y in either the training or test distributions. In seminal work, Ben-David et al. (2006) introduced the $\mathcal{H}$-divergence as a theoretical measure of difference between two input distributions, as well as the $\mathcal{A}$-distance as an empirically-calculated proxy for the $\mathcal{H}$-divergence. They showed that the $\mathcal{A}$-distance could be estimated by training a linear classifier (i.e., logistic regression) to discriminate between domains, and then using domain classification accuracy on a held-out sample as a measure of the distributional similarity. Gözüaçık et al. (2019) use this foundation for identifying distributional shifts in x over time, namely by leveraging

the discriminatory strength between moving windows of data as an online alerting mechanism. Rabanser et al. (2019) demonstrated that accurately estimating the divergence using a domain classifier requires potentially prohibitively large sample sizes. However, they also noted that domain classifiers provide an opportunity to reason about the nature of the shift (e.g., by examining the most and least certain examples).

Various alternatives to the domain-discriminative classifier approach have been proposed. Gokhale et al. (2022) generate a curve of predicted probabilities under a trained model by varying the degree of interpolation between a training and test example (i.e., convex combination, token-wise swapping), and then using the area under this curve as a measure of covariate shift for the particular example; Tian et al. (2021) uses the ℓ_2 norm of the last embedding layer of a pretrained neural model to compute a covariate shift score for unseen examples; and Fei and Liu (2015) transforms individual test documents into a similarity space by measuring similarity to training documents under varying n-gram representations, and then using a one-class SVM classifier to determine whether the unseen sample is OOD. In each of these three cases, instance-level measures are then aggregated across a sample to draw a conclusion about the presence of distribution shift (e.g., rate of detection, mean similarity score). Other practitioners opt to avoid the instance-level measure and instead pass an average feature representation from two data samples into a

two-sample hypothesis test (e.g., Maximum Mean Discrepancy (MMD)) (Gretton et al., 2012; Castle et al., 2021). For a more comprehensive and relatively modern review of covariate shift detection mechanisms, I recommend consulting Nair et al. (2019).

Prior Shift. Prior shift detection has relevance in the medical and public health communities (e.g., conditions cause symptoms, a condition's prevalence may vary over time) (Lipton et al., 2018), as well as relevance in imbalanced or rare classification problems (Cao et al., 2019; Ye et al., 2024). It is often viewed as a difficult phenomena to detect because, unlike in covariate shift detection, we cannot assume having access to ground truth from the output space. The prevailing approaches for detecting prior shift are Black Box Shift Estimation (BBSE) (Lipton et al., 2018) and Maximum Likelihood Label Shift (MLLS) Estimation (Garg et al., 2020). Practitioners may opt to focus on covariate shift detection or concept shift detection given that the presence of label shift implies that at least one of covariate shift or concept shift has occurred.

> *Claim:* $p(y_a) \neq p(y_b)$ and $p(x_a \mid y_a) = p(x_b \mid y_b) \Rightarrow p(x_a) \neq p(x_b)$ or $p(y_a \mid x_a) \neq p(y_b \mid x_b)$.

Proof:

$$p(y_a) \neq p(y_b) \qquad \text{(1. By assumption)}$$

$$\frac{p(y_a \mid x_a)p(x_a)}{p(x_a \mid y_a)} \neq \frac{p(y_b \mid x_b)p(x_b)}{p(x_b \mid y_b)} \qquad \text{(2. Bayes theorem)}$$

$$p(y_a \mid x_a)p(x_a) \neq p(y_b \mid x_b)p(x_b) \qquad \text{(3. By assumption)}$$

$$p(x_a) = p(x_b) \Rightarrow p(y_a \mid x_a) \neq p(y_b \mid x_b) \qquad \text{(4. Case if no covariate shift)}$$

$$p(y_a \mid x_a) = p(y_b \mid x_b) \Rightarrow p(x_a) \neq p(x_b) \qquad \text{(5. Case if no concept shift)}$$

$$p(y_a) \neq p(y_b) \Rightarrow p(y_a \mid x_a) \neq p(y_b \mid x_b)$$

$$\text{or } p(x_a) \neq p(x_b) \quad \square \qquad \text{(6. Putting together)}$$

Note that (3) follows by assumption that $p(x_a \mid y_a) = p(x_b \mid y_b)$. To show that at least one of covariate shift or concept shift must be present, we assume that either one is not present and show that the other must exist.

Concept Shift. Similar to prior shift, concept shift detection is made difficult by the absence of observations from the output space. In general, practitioners approach the problem of detecting concept shifts using an existing model of the conditional relationship between x and y. For example, Elsahar and Gallé (2019) uses intermediate representations from a task model (e.g., last embedding layer) as the representation for computing the proxy $\mathcal{A}$-distance; Huang et al. (2021) computes the gradient of the KL-divergence between the softmax output from a pretrained model and a

uniform distribution over the class outputs and then uses the norm of this gradient as a measure of distributional similarity; and Sethi and Kantardzic (2017) measures the density of unseen examples that are placed in the margin of a task-specific SVM classifier. While a model of the conditional x and y relationship may be necessary to confirm the presence of concept shift, we will show in Chapter 7 that, given existing domain knowledge, a model of the conditional relationship between x and y may not be explicitly necessary to at least raise concerns that a concept shift has occurred.

Comments on Detection. The ability to detect that a distribution shift has occurred may be extremely useful to those monitoring a deployed model, or alternatively for gauging whether an existing model will work on a new distribution (e.g., a new patient population). However, detecting that a distribution shift has occurred is generally not enough in most applications. Once a shift is detected, there exists a need to thoroughly *understand* the nature of the distribution shift (e.g., features that changed, features that remained stable, causes of the shift). The task of understanding a distribution shift often cannot be done using statistical tests alone. Instead, practitioners must leverage external knowledge – either about their model, data, or the world more generally – to reason about the nature of the shift. This is particularly true for language data, where modern measures of shift may be based on embeddings or neuron activations that can be difficult to interpret (Jain

and Wallace, 2019; Vijayakumar, 2023). As we will highlight in Chapters 5 through 7, data exploration and experimentation rooted in the application itself can reveal significant information that a formal statistical test would not.

3.2.3 Addressing Distribution Shift

Once a practitioner knows that distribution shift is present or likely to be present in their deployment scenario, they are faced with the challenge of addressing it. There are broadly two trains of thought with respect to how this should be done. On one hand lies the perspective that machine learning models should learn general, sample-invariant relationships that are robust to distribution shifts – i.e., *domain generalization*. On the other hand lies the perspective that machine learning models should be adapted to specific target distributions, and these adaptations can vary across target distributions – i.e., *domain adaptation*. We will review both independently before comparing them to one another.[2]

Domain Generalization. The objective of training a model on one or more distributions that will generalize to *unseen* target distributions is referred to as domain

[2] For language data, it is particularly difficult to define what constitutes a domain (Baldwin et al., 2013). It is perhaps better to say we are interested in *distribution adaptation* or *distribution generalization*. In this way, we would at least account for the fact that traditional "domains" (e.g., social media, news, clinical notes) are actually quite heterogeneous (e.g., varied vocabularies and topicality).

generalization. It has received significant attention as a subfield of machine learning in recent years due to the proliferation of models being deployed on distributions of data for which they were not originally trained (Zunic et al., 2020; Sushil et al., 2021).

One of the central ways in which domain generalization is achieved is by training a robust model. A machine learning model $f(\cdot)$ is said to be *robust* if for any two inputs x and x', $f(x) = f(x')$ as long as x and x' are similar enough (Ben-Tal and Nemirovski, 1998). Robustness is proven to be a necessary and sufficient condition for generalization (Xu and Mannor, 2012). In the context of NLP, robustness research typically focuses on ensuring models do not rely on spurious correlations and instead learn core feature sets (Tu et al., 2020; Wang et al., 2022b). For instance, in the two examples below, a robust model for detecting depressive language should classify both statements in the affirmative based on their semantic implications, regardless of the exact tokens used to express that meaning.

> *Example 1.* Going to be another rough night. Sadness sucks the life out of me.

> *Example 2.* Going to be another rough afternoon. Sadness sucks the soul out of me.

Existing attempts to learn robust relationships in language data fall into three categories: instance-level corrections, feature-level corrections, and model optimization strategies. We provide examples of these approaches below, and refer the reader to

the several available literature reviews on the subject if they are interested in learning more (Liu et al., 2021b; Wang et al., 2022a; Wang et al., 2022c; Zhou et al., 2022).

Instance-level corrections are applied by scoring the predictability of examples in the training data and then excluding the "easy" examples during training to enhance generalization (Le Bras et al., 2020; Yaghoobzadeh et al., 2021). This class of methods also includes counterfactual data augmentation, where counterfactual examples (i.e., similar x, different y) are constructed – either manually or automatically using heuristics – and added to the training data sample (Kaushik et al., 2019; Teney et al., 2020; Joshi and He, 2022).

Feature-level adjustments leverage gradient supervision in combination with interpretability metrics to identify influential covariates and remove those that lack causal relevance with respect to the task – manually or with help of a knowledge base (Wang and Culotta, 2020; Gardner et al., 2021; Han and Tsvetkov, 2021; Wang et al., 2022b).

Optimization strategies focus on regularizing away features that vary in importance across different subsets of the full distribution (i.e., random samples, fixed covariate clusters), operating under an assumption that invariant features are "core" to solving the task (Sagawa et al., 2019; Sohoni et al., 2020; Liu et al., 2021a; Duchi et al., 2023).

Given the focus on learning invariant relationships, many of these methods are inspired by work on causal learning (Bühlmann, 2020; Zhang et al., 2021; Sheth et al., 2022).

Domain Adaptation. The objective of transferring knowledge from one observed set of distributions to another is referred to as domain adaptation (or domain transfer in some ML literature). Unlike domain generalization approaches where the goal is to achieve equitable levels of performance within arbitrary unseen distributions, domain adaptation approaches assume knowledge of the target data distribution and attempt to maximize performance specifically within this target data distribution. A quick review of the literature will highlight the progress that has been made toward this goal for both supervised and unsupervised settings (Jiang, 2008; Sun et al., 2015; Kouw and Loog, 2019; Ramponi and Plank, 2020; Wilson and Cook, 2020; Farahani et al., 2021).

Approaches to facilitate domain adaptation are quite diverse. Importance weighting methods re-weight training examples based on their similarity to examples from the target distribution (Jiang and Zhai, 2007; Zhang et al., 2013). Feature augmentation and transformation methods generate domain-specific and domain-agnostic versions of existing features and use this augmented representation to disentangle domain-specific relationships from those that generalize across domains (Blitzer et al., 2006; 2007; Daumé III, 2007; Daumé III et al., 2010). And in adversarial learning setups, neural models attempt to learn a feature representation that is both

predictive of the target outcome and not predictive of domain membership (Tzeng et al., 2017; Long et al., 2018; Zhao et al., 2018a; HassanPour Zonoozi and Seydi, 2023).

Two additional classes of techniques are most relevant for this thesis. The first class of methods, stable feature selection, involves the identification of features that are unlikely to experience a change in either $p(x)$ or $p(y \mid x)$ from the source to target data distribution (Kouw et al., 2016; Sun et al., 2019a; Fu et al., 2021; Yan et al., 2021). In Chapter 7, I introduce a method that leverages quantitative measures of semantic shift (i.e., changes in the meaning of words, or changes in the prevalence of difference word senses) as a stable feature selection mechanism. The second class of methods involves continued pretraining of a language model on a target data distribution for the purpose of improving the latent representation of text in the target domain (Gururangan et al., 2020; Rietzler et al., 2020; Diao et al., 2021; Madan et al., 2021). In Chapter 8, I use a combination of domain adaptive and task adaptive pretraining (Gururangan et al., 2020) to investigate contemporary claims that pretraining a language model on out-of-domain clinical data provides benefits in clinical tasks over pretraining on non-clinical data.

Choosing an Approach. Because both domain generalization and domain adaptation have the same end goal – training models that generalize to data that practitioners care most about – it can be difficult to decide which approach one

should take. In practice, the decision may not have anything to do with the efficacy of the approaches themselves, but rather the deployment scenarios for which they are most appropriate.

Domain generalization implicitly assumes that for a given predictive task, there exists a single model to "rule them all." This property is appealing to practitioners primarily for two reasons. First, it means there is less maintenance required upon deployment. In comparison, domain adaptation may require practitioners to re-train models (or adapt them in other ways) whenever distribution shift occurs – a process which can be impractical for models deployed continuously over time or in a large number of settings (e.g., hospitals, social media communities) (Schelter et al., 2018; Mensah et al., 2020). Second, the causal underpinnings (i.e., invariance) of domain generalization are likely to engender higher levels of trust from stakeholders in comparison to models that have potentially distribution-specific predictors (Gilpin et al., 2018).

Wherein lies the issue with domain generalization is its actually efficacy on real-world data. For instance, robust optimization methods have been successful at identifying and removing spurious relationships in synthetic or highly-curated datasets (Arjovsky et al., 2019; Sagawa et al., 2019), but fail when applied on datasets not having very specific causal structures (Choe et al., 2020; Guo et al., 2021). In

other cases, it has been shown that robustness on multiple distributions comes at the expense of optimal performance (Li and Li, 2024). It may even exacerbate performance disparities between groups (e.g., racial and gender minorities) (Khani and Liang, 2021). Domain adaptation allows us to account for the fact that some distributions (and groups of people) have reasonable forms of variation, or at least reasonable forms of variation amongst the covariates we can practically measure (e.g., semantics, but not a speaker's intent) (Hovy and Yang, 2021).

Practitioners should ultimately opt for the approach that works best for their task and deployment setting. However, I would argue that the ideal scenario involves training a "general" model and then adapting it to particular distributions thereafter. For instance, in Yi et al. (2024), we show that language models pretrained on narrow samples of clinical data are less adaptable to new tasks and distributions than language models pretrained on large amounts of non-clinical data. These results align with much of the contemporary work on domain-specific language models – e.g., BloombergGPT is trained on generic language and then adapted to the financial domain (Wu et al., 2023); MedPalm builds directly from Palm instead of being trained completely from scratch (Singhal et al., 2023b). Pure domain generalization approaches may be the answer when we have unambiguous and straightforward causal-structures, but, at least for now, domain adaptation approaches better reflect and address the

complexities of real world data.

3.3 Equity Takes More Than a Model

Negative effects on health equity may arise at several points in an NLP system's overall pipeline and deployment, not just within the underlying models and algorithms. For example, one culture may easily understand a system's user interface, while another culture may struggle with it (Ishak et al., 2012; Alexander et al., 2021). Alternatively, a well-resourced institution with adequate compute infrastructure may be able to fully utilize a system, while an under-resourced institution may be forced to sacrifice functionality due to compute limitations. In general, negative effects on health equity do not arise out of malicious intent (e.g., purposely generating misinformation for a population). Rather, they arise because practitioners focused on developing models and solving computational tasks don't always consider the larger context for their work.

While challenges related to the deployment of NLP systems and their effect on health equity are out of this thesis' scope, I would be remiss not to take this opportunity to highlight their existence. Below is a non-comprehensive list of questions that practitioners may find helpful when considering the potential impact of their work on health equity. Ideally, introspection regarding such questions should occur well before a system's development process begins – not in the hour leading up to a

paper submission deadline while frantically putting together an impact statement.

- Will both historically advantaged and disadvantaged communities reasonably be able to use the system I develop to its fullest extent (e.g., sufficient compute, technical expertise, maintenance support)?
- Were the models and algorithms included in my system originally developed for a purpose that aligns with the system's planned use (e.g., construct validity of a classification outcome)?
- Can my system be used in an unintended manner for malevolent reasons? What safeguards can I put in place to minimize the risk of this usage?
- Are there scenarios or distributions for which my system should *not* be used? If so, do these limits affect historically disadvantaged communities disproportionately?
- How does my system perform relative to existing solutions, automated or otherwise, for the same task?
- Is it reasonable to expect that data passed as an input to my system upon deployment will be similar to the data used for developing it? If not, will it be possible to update the system to accommodate such data?

Chapter 4

The State of Social Media Data for Mental Health Research: A Case Study

4.1 Overview

To understand why distribution shift is such a prominent concern in the health domain, it is useful to understand the challenges faced by practitioners when curating health datasets, as well as the approaches commonly used to circumvent those challenges. To make these challenges tangible and provide context for the technical contributions set forth in Chapters 5 through 7, we will now present a case study on the availability of social media datasets used for the purpose of modeling mental health status (e.g., depression, schizophrenia). This work was presented originally in Harrigian et al. (2021) and contains minor modifications to more acutely highlight generalities experienced in health-related applications.

4.2 Background

The last decade has seen exponential growth in computational research devoted to modeling mental health phenomena using clinical (Poulin et al., 2014; Garriga et al., 2022) and non-clinical language data (Bucci et al., 2019). Studies analyzing data from the web, such as social media platforms and peer-to-peer messaging services, have been particularly appealing to the research community due to their scale and deep entrenchment within contemporary culture (Fuchs, 2015; Graham et al., 2015; Perrin,

2015). Such studies have yielded novel preliminary insights into population-level mental health (De Choudhury et al., 2013b; Amir et al., 2019) and shown promising avenues for the incorporation of data-driven analyses in the treatment of psychiatric disorders (Eichstaedt et al., 2018).

Nonetheless, complexities specific to the health domain, and more particularly the mental-health domain, have generally made it difficult to obtain data that supports broad and definitive conclusions regarding mental health phenomena. For instance, behavioral disorders are known to display variable clinical presentations amongst different populations (De Choudhury et al., 2017), but methods for curating "ground truth" mental health status annotations often do not reflect this heterogeneity (Arseniev-Koehler et al., 2018). Semi-automated methods for annotating an individual's mental health status have allowed ML and NLP practitioners to circumvent logistical challenges imposed by involving domain experts to label data (Coppersmith et al., 2015c; Kumar et al., 2015), but they typically rely on oversimplifications that lack the same clinical validation and robustness as something like a mental health battery (Zhang et al., 2014; Ernala et al., 2019). When researchers have been able to curate heterogeneous datasets with clinically valid annotations (Eichstaedt et al., 2018), ethical considerations related to the sensitive nature of the data have made it difficult to share them (Benton et al., 2017a).

4.3 Motivation and Contribution

Prior reviews of computational research for mental health have noted several of the aforementioned challenges, but have predominantly discussed technical methods (e.g., model architectures, feature engineering) developed to surmount existing constraints (Guntuku et al., 2017; Wongkoblap et al., 2017). Work from Chancellor and De Choudhury (2020), completed concurrently with our own, examines the shortcomings of *data* for mental health research. However, it critically does not focus on *language* found in social media data, nor does it measure the *accessibility* of datasets in the domain. To this end, we construct a new open-source directory of mental health datasets, annotated using a standardized schema that enables researchers to identify relevant and accessible datasets.[1] We draw upon this resource to offer nuanced insights regarding current shortcomings of data in the "social media for mental health" application domain.

4.4 Methods

Unlike some computational fields that have a surplus of well-defined and uniformly-adopted benchmark datasets, mental health researchers have thus far relied

[1] https://github.com/kharrigian/mental-health-datasets

on a decentralized medley of resources. This fact, spurred in part by the variable presentations of psychiatric conditions and in part by the sensitive nature of mental health data, thus requires that we compile a new database of literature. In this section, we detail our literature search methodology, establish inclusion/exclusion criteria, and define a list of dataset attributes to analyze.

4.4.1 Dataset Search

Datasets were identified using a breadth-focused literature search. After including data sources from the three aforementioned systematic reviews (Guntuku et al., 2017; Wongkoblap et al., 2017; Chancellor and De Choudhury, 2020), we searched for literature that lie primarily at the intersection of natural language processing (NLP) and mental health communities. We sought peer-reviewed studies published between January 2012 and December 2019 in relevant NLP conferences (e.g., NAACL, EMNLP, ACL, COLING), NLP workshops (e.g., CLPsych, LOUHI), and health-focused journals (e.g., JMIR, PNAS, BMJ). We also searched Google Scholar, ArXiv, and PubMed using the following two query structures to identify additional candidate articles:

1. (mental health|`DISORDER`) + (social|electronic) + media

2. (machine learning|prediction|inference|detection) + (mental health|`DISORDER`)

Here, the "|" indicates a logical or, and `DISORDER` was replaced by one of 13 mental health keywords.[2] Additional literature was identified using snowball sampling from the citations of these papers. To moderately restrict the scope of this work, computational research regarding neurodegenerative and cognitive disorders (e.g., Dementia, Parkinson's Disease) was ignored.

4.4.2 Selection Criteria

To enhance parity amongst datasets considered in our meta-analysis, we require datasets found within the literature search to meet three additional criteria. While excluded from subsequent analysis, datasets that do not meet this criteria are maintained with complete annotations in the aforementioned digital directory.

1. Datasets must contain non-clinical electronic media (e.g., social media, SMS, online forums, search query text).

2. Datasets must contain written language (i.e., text) within each unit of data.

3. Datasets must contain a dependent variable that captures or proxies a psychiatric condition listed in the DSM-5 (APA, 2013).

[2] Depression, Suicide, Anxiety, Mood, PTSD, Bipolar, Borderline Personality, ADHD, OCD, Panic, Addiction, Eating, Schizophrenia

Our first criteria excludes research that examines electronic health records or digitally-transcribed interviews (Gratch et al., 2014; Holderness et al., 2019). Our second criteria excludes research that, for example, primarily analyzes search query volume or mobile activity traces (Ayers et al., 2013; Renn et al., 2018). It also excludes research based on speech data (Iter et al., 2018). Our third criteria excludes research in which annotations are only loosely associated with their stated mental health condition. For instance, we filter out research that seeks to identify diagnosis dates in self-disclosure statements (MacAvaney et al., 2018), in addition to research that proposes using sentiment as a proxy for mental illness (Davcheva et al., 2019). This last criteria also inherently excludes datasets that lack annotation of mental health status altogether (e.g., data dumps of online mental health support platforms and text-message counseling services) (Loveys et al., 2018; Demasi et al., 2019).

4.4.3 Annotation Schema

We develop a high-level schema to code properties of each dataset. In addition to standard reference information (i.e., Title, Year Published, Authors), we keep track of the following characteristics:

Platforms: Electronic media source (e.g., Twitter, SMS)

Tasks: The mental health disorders included as dependent variables (e.g., depression, suicidal ideation, PTSD)

Annotation Method: Method for defining and annotating mental health variables (e.g., regular expressions, community participation/affiliation, clinical diagnosis)

Annotation Level: Resolution at which ground-truth annotations are made (e.g., individual, document, conversation)

Size: Number of data points at each annotation resolution for each task class

Language: The primary language of text in the dataset

Data Availability: Whether the dataset can be shared and, if so, the mechanism by which it may be accessed (e.g., data usage agreement, reproducible via API, distribution prohibited by collection agreement)

If a characteristic is not clear from a dataset's associated literature, we leave the characteristic blank; missing data points are denoted where applicable. While we simplify these annotations for a standardized analysis – e.g., different psychiatric batteries used to annotate depression in individuals (e.g., PHQ-9, CES-D) are simplified as "Survey (Clinical)" – we maintain specifics in the digital directory.

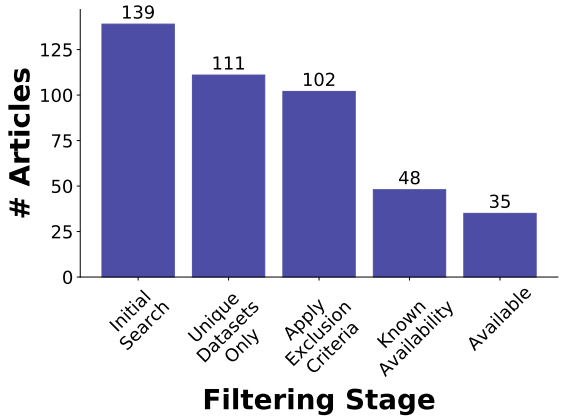


Figure 4.1: The number of articles (i.e., datasets) remaining after each stage of our search procedure. We were unable to readily discern the external availability of datasets for over half of the studies identified by our search procedure.

4.5 Results

Our literature search yielded 139 articles referencing 111 nominally-unique datasets. Application of exclusion criteria left us with 102 datasets. A visualization of dataset volume remaining after each filtering criteria is applied is provided in Figure 4.1.

Publication Trends. The majority of datasets were released after 2012, with an average of 12.75 per year, a minimum of 1 (2012), and a maximum of 23 (2017). The 2015 CLPsych Shared Task (Coppersmith et al., 2015c), Reddit Self-reported Depression Diagnosis (Yates et al., 2017), and "Language of Mental Health" (Gkotsis et al., 2016) datasets were the most reused resources, serving as the basis of 7, 3, and 3 additional publications respectively.

Platforms. We identified 20 unique electronic media platforms across the 102

datasets. Twitter (47 datasets) and Reddit (22 datasets) were the most widely studied platforms. YouTube, Facebook, and Instagram were relatively underutilized for mental health research – each found less than ten times in our analysis – despite being the three most-widely adopted social media platforms globally (Perrin and Anderson, 2019). We expect our focus on NLP to moderate the presence of YouTube and Instagram datasets, though not entirely given both platforms offer expansive text fields (i.e., comments, tags) in addition to their primary content of video and images (Chancellor et al., 2016a; Choi et al., 2016). It is more likely that use of these platforms (and Facebook) for research is hindered by increasingly stringent privacy policies and ethical concerns (Panger, 2016; Benton et al., 2017a).

Tasks. We identified 36 unique mental health related modeling tasks across the 102 datasets: Antisocial Behavior (ABHV), Attention Deficit Hyperactivity Disorder (ADHD), Alcoholism (ALC), Anxiety (ANX), Social Anxiety (ANXS), Asperger's (ASP), Autism (AUT), Bipolar Disorder (BI), Borderline Personality Disorder (BPD), Cognitive Distortion (COG), Depression (DEP), Eating Disorder (EAT), Recovery from Eating Disorder (EATR), Grief (GRF), Loneliness (LONE), General Mental Health Disorder (MHGEN), Mood Instability (MOOD), Obsessive Compulsive Disorder (OCD), Opiate Addiction (OPAD), Opiate Usage (OPUS), Post Traumatic Stress Disorder (PTSD), Panic Disorder (PAN), Postpartum Depression

(PPD), Psychosis (PSY), Trauma from Rape (RS), Schizophrenia (SCHZ), Seasonal Affective Disorder (SAD), Self-Esteem (SELF), Self Harm (SH), Sleep Disorder (SLP), Stress (STR), Stressor Subjects (STRS), Substance Use (SUB), Suicide Attempt (SA), Suicidal Ideation (SI), Trauma (TRA). Only 27 of these conditions or mental states were represented in the subset of datasets that were available for distribution.

While the majority of tasks were examined less than twice, a few tasks were considered quite frequently. Depression (42 datasets), suicidal ideation (26 datasets), and eating disorders (11 datasets) were the most common psychiatric conditions examined. Anxiety, PTSD, self-harm, bipolar disorder, and schizophrenia were also prominently featured conditions, each found within at least four unique datasets. A handful of studies sought to characterize finer-grained attributes associated with higher-level psychiatric conditions (e.g., symptoms of depression, stress events and stressor subjects) (Mowery et al., 2015; Lin et al., 2016). The dearth of anxiety-specific datasets was somewhat surprising given the condition's prevalence and the abundance of pyschometric batteries for assessing anxiety (Cougle et al., 2009; Antony and Barlow, 2020). That said, generalized anxiety disorder (GAD) only accounts for a small proportion of the overall prevalence of anxiety disorders (Bandelow and Michaelis, 2015) and many other types of anxiety disorders (e.g., social anxiety, PTSD, OCD, etc.) were typically treated as independent conditions (Coppersmith

et al., 2015a; De Choudhury et al., 2016).

Annotation. We identified 24 unique annotation mechanisms. It was common for several annotation mechanisms to be used jointly to increase precision of the defined task classes and/or evaluate the reliability of distantly supervised labeling processes. For example, some form of regular expression matching was used to construct 43 of datasets, with 23 of these including manual annotations as well. Community participation/affiliation (24 datasets), clinical surveys (22 datasets), and platform activity (3 datasets) were also common annotation mechanisms. The majority of datasets contained annotations made on the individual level (63 datasets), with the rest containing annotations made on the document level (40 datasets).[3]

Size. Of the 63 datasets with individual-level annotations, 23 associated articles described the amount of documents and 62 noted the amount of individuals available. Of the 40 datasets with document-level annotations, 37 associated articles noted the amount of documents and 12 noted the number of unique individuals. The distribution of dataset sizes was right skewed (i.e., long-tailed).

One concerning trend that emerged across the datasets was the presence of a relatively low number of unique individuals. Indeed, these small sample sizes may further inhibit model generalization from platforms that are already

[3] One dataset was annotated at both a document and individual level

demographically-skewed (Smith and Anderson, 2018). The largest datasets, which present the strongest opportunity to mitigate the issues presented by poorly representative online populations, tend to leverage the noisiest annotation mechanisms. For example, datasets that define a mainstream online community as a control group may expect to find approximately 1 in 20 of the labeled individuals are actually living with mental health conditions such as depression (Wolohan et al., 2018a), while regular expressions may fail to distinguish between true and non-genuine disclosures of a mental health disorder up to 10% of the time (Cohan et al., 2018).

Primary Language. Six primary languages were found amongst the 102 datasets – English (85 datasets), Chinese (10 datasets), Japanese (4 datasets), Korean (2 datasets), Spanish (1 dataset), and Portuguese (1 dataset). This is not to say that some of the datasets do not include other languages, but rather that the predominant language found in the datasets occurs with this distribution. While an overwhelming focus on English data is a theme throughout the NLP community, it is a specific concern in this domain where culture often influences the presentation of mental health disorders (De Choudhury et al., 2017; Loveys et al., 2018).

Availability. We were able to identify the availability of only 48 of the 102 unique datasets in our literature search. Of these 48 datasets, 13 were known not to be available for distribution, generally due to limitations defined in the original

collection agreement or removal from the public record (Park et al., 2012; Schwartz et al., 2014). The 35 available datasets were distributed via the following mechanisms: 18 may be reproduced using an API and instructions provided within the associated article (API), 12 require a signed data usage agreement and/or IRB approval (DUA), 3 are available without restriction (FREE), and 2 may be retrieved directly from the author(s) with permission (AUTH). Of the 22 datasets that used clinically-derived annotations (e.g., mental health battery, medical history), 7 were unavailable for distribution due to terms of the original data collection process and 1 was removed from the public record. The remaining 14 had unknown availability. All datasets known to be available for distribution are presented in Table 4.1, while remaining datasets are found our digital directory.

4.6 Challenges and Recommendations

In the previous sections, we introduced and analyzed a standardized directory of social media datasets used by computational scientists to model mental health phenomena. In doing so, we have provided a valuable resource poised to help researchers quickly identify new datasets that support novel research. Moreover, we have provided evidence that affirms conclusions from Chancellor and De Choudhury (2020) and may further encourage researchers to rectify existing gaps in the data landscape. Below,

Reference	Platform(s)	Task(s)	Level	Individuals	Documents	Availability
Coppersmith et al. (2014a)	Twitter	BI, PTSD, SAD, DEP	Ind.	7k	16.7M	DUA
Coppersmith et al. (2014b)	Twitter	PTSD	Ind.	6.3k	-	DUA
Jashinsky et al. (2014)	Twitter	SI	Doc.	594k	733k	API
Lin et al. (2014)	Twitter, Sina/Tencent Weibo	STR, STRS	Ind.	23.3k	490k	API
Coppersmith et al. (2015a)	Twitter	ANX, EAT, OCD, SCHZ, SAD, BI, PTSD, DEP, ADHD	Ind.	4k	7M	DUA
Coppersmith et al. (2015c)	Twitter	PTSD, DEP	Ind.	1.7k	-	DUA
De Choudhury (2015)	Tumblr	EAT, EATR	Ind.	28k	87k	API
Kumar et al. (2015)	Reddit, Wikipedia	SI	Ind.	66k	19.1k	API
Mowery et al. (2015)	Twitter	DEP	Doc.	-	129	AUTH
Chancellor et al. (2016b)	Tumblr	EATR	Ind.	13.3k	67M	API
Coppersmith et al. (2016)	Twitter	SA	Ind.	250	-	DUA
De Choudhury et al. (2016)	Reddit	PSY, EAT, ANXS, SH, BI, PTSD, RS, DEP, PAN, SI, TRA	Ind.	880	-	API
Gkotsis et al. (2016)	Reddit	ANX, BPD, SCHZ, SH, ALC, BI, OPAD, ASP, SI, AUT, OPUS	Ind.	-	-	API
Lin et al. (2016)	Sina Weibo	STR	Doc.	-	2.6k	FREE
Milne et al. (2016)	Reach Out	SH	Doc.	1.2k	-	DUA
Mowery et al. (2016)	Twitter	DEP	Doc.	-	9.3k	AUTH
Bagroy et al. (2017)	Reddit	MHGEN	Doc.	30k	43.5k	API
De Choudhury and Kiciman (2017)	Reddit	SI	Ind.	51k	103k	API
Losada et al. (2017)	Reddit	DEP	Ind.	887	530k	DUA
Saha and De Choudhury (2017)	Reddit	STR	Doc.	-	2k	API
Shen et al. (2017)	Twitter	DEP	Ind.	300M	10B	FREE
Shen and Rudzicz (2017)	Reddit	ANX	Doc.	-	22.8k	API
Yates et al. (2017)	Reddit	DEP	Ind.	116k	-	DUA
Chancellor et al. (2018)	Reddit	EAT	Doc.	-	2.4M	API
Cohan et al. (2018)	Reddit	ANX, EAT, OCD, SCHZ, BI, PTSD, DEP, ADHD, AUT	Ind.	350k	-	DUA
Dutta et al. (2018)	Twitter	ANX	Ind.	200	209k	API
Ireland and Iserman (2018)	Reddit	ANX	Ind.	-	-	API
Li et al. (2018b)	Reddit	MHGEN	Ind.	1.8k	-	API
Losada et al. (2018)	Reddit	EAT, DEP	Ind.	1.5k	1.2M	DUA
Pirina and Çöltekin (2018)	Reddit	DEP	Doc.	-	1.2k	API
Shing et al. (2018)	Reddit	SI	Ind.	1.9k	-	DUA
Sekulic et al. (2018)	Reddit	BI	Ind.	7.4k	-	API
Wolohan et al. (2018a)	Reddit	DEP	Ind.	12.1k	-	API
Turcan and McKeown (2019b)	Reddit	STR	Doc.	-	2.9k	FREE
Zirikly et al. (2019)	Reddit	SI	Ind.	496	32k	DUA

Table 4.1: Characteristics of datasets that meet our study's inclusion criteria and are known to be accessible outside of the original research group.

we summarize additional shortcomings not previously highlighted by Chancellor and De Choudhury (2020).

Non-Standardized Task Definitions. In just 102 datasets, we identified 24 unique annotation mechanisms used to label 36 types of mental health phenomena. This total represents a conservative estimate given that nominally equivalent annotation procedures often varied non-trivially between datasets (e.g., PHQ-9 vs. CES-D assessments, affiliations based on Twitter followers vs. engagement with a subreddit) (Faravelli et al., 1986; Pirina and Çöltekin, 2018). Minor discrepancies in task definition reflect the heterogeneity of how several mental health conditions manifest, but can also introduce difficulty contextualizing results across deployment scenarios. Moreover, many of these definitions may still fall short of capturing the nuances of mental health disorders (Arseniev-Koehler et al., 2018). As researchers look to transition computational models into the clinical setting, it is imperative they have access to standardized benchmarks that inform interpretation of predictive results in a consistent manner (Norgeot et al., 2020).

Sensitive Data Access. Most existing mental health datasets rely on some form of self-reporting or distinctive behavior to assign individuals into task groups, but admittedly fail to meet ideal ground truth standards. The clinically-annotated datasets that do exist are either proprietary or do not provide a clear mechanism for

inquiring about availability. The dearth of large, shareable datasets based on actual clinical diagnoses and medical ground truth is problematic given recent research that calls into question the validity of proxy-based mental health annotations (Ernala et al., 2019). By leveraging privacy-preserving technology (e.g., differential privacy) to share patient-generated data, researchers may ultimately be able to train more robust computational models (Elmisery and Fu, 2010; Zhu et al., 2016; Dwivedi et al., 2019). In lieu of implementing complicated technical approaches to preserve the privacy of human subjects within mental health data, researchers may instead consider establishing secure computational environments that enable collaboration amongst authenticated users (Boebert et al., 1994; Rush et al., 2019).

Selection Bias. There remains more to be done to ensure models trained using these datasets perform consistently irrespective of population. While multiple studies in our review attempted to leverage demographically-matched or activity-based control groups as a comparison to individuals living with a mental health condition (Coppersmith et al., 2015c; Cohan et al., 2018), they made up only a minority of all studies and datasets examined. A recent article found discrepancies between the prevalence of depression and PTSD as measured by the Centers for Disease Control and Prevention and as estimated using a model trained to detect the two conditions (Amir et al., 2019). While the study posits reasons for the difference, it is unable to

confirm any causal relationship. Is the lack of representative data to blame?

Recently, Aguirre et al. (2021) found evidence of demographic (gender and racial/ethnic) bias within datasets from Coppersmith et al. (2015d) and Benton et al. (2017b) that can create fairness issues in downstream tasks. They found poor representation and strong group imbalance in these datasets; however, simple changes in dataset size and balance alone could not fully account for performance disparities between groups. Indeed, common signs of depression recognized in prior linguistic analyses (e.g., differences in distributions for some categories of LIWC) were found not to be equally informative for all demographics. Thus, while performance disparities between demographic groups may certainly arise due to poor representation at training time, disparities may also arise due to an ill-founded assumption that mental health outcomes for all groups can be treated equivalently (Kessler et al., 2003; De Choudhury et al., 2017; Shah et al., 2020). Either way, there exists a need to rethink dataset curation and model evaluation so traditionally underrepresented groups are not further hindered from receiving adequate mental health care. Demographic inference tools may aid in constructing datasets with demographically-representative cohorts (Huang and Carley, 2019; Wood-Doughty et al., 2020). Researchers may also consider expanding the diversity of languages in their datasets to account for variation in mental health presentation that arises due to cultural differences (De

Choudhury et al., 2017; Loveys et al., 2018).

4.7 Discussion

Although this case study focuses on a narrow application area and data source, it highlights many of the data-related challenges that arise more generally across the health domain. These challenges fall broadly into three areas – 1) Human Factors, 2) Domain Expertise Requirements, and 3) Privacy Concerns. Each has a strong potential to introduce selection bias into the dataset curation process, and in turn make NLP models susceptible to being deployed on data distributions that systematically differ from those they were trained on.

Human Factors. Human behavior and physiology are strongly influenced by environmental circumstances and social dynamics. These factors not only introduce heterogeneity into the manner in which clinical conditions arise, present, and are treated, but also shape how individuals interact with the data collection mechanisms that are used by ML and NLP practitioners – e.g., opt-in biases, disclosure biases, lexical variation. Amongst the mental health datasets analyzed in the case study, very few demonstrate an acknowledgement of differences in human experience based on personal characteristics (Coppersmith et al., 2015c; De Choudhury et al., 2017). In the absence of this recognition, models are prone to being trained and evaluated on

datasets that only present a subset of possible life experiences and health outcomes.

Domain Expertise Requirements. Access to ground truth that arises naturally during the course of health-related interactions (e.g., ICD-10 codes, clinical batteries and vitals) is not guaranteed, especially for applications that leverage non-clinical data sources. Unfortunately, acquiring secondary annotations for the purpose of training and evaluating models in the health domain is complicated by the necessity of domain expertise for many applications of interest (e.g., phenotyping, clinical concept extraction, diagnosis). On one hand, individuals with the expertise necessary to perform such annotation efforts (i.e., healthcare providers) are often already time-constrained by their primary responsibilities. On the other hand, many of the annotation efforts (e.g., reading clinical notes, reasoning about diagnoses, span-level labeling) are cognitively taxing in a manner severely limits the amount of annotated data that a practitioner can reasonably expect to obtain.

The heavy presence of mental health datasets curated using regular expressions and other proxy-based labeling mechanisms (e.g., self-disclosed diagnoses) parallels trends in the broader health domain (Ku et al., 2014; Yang et al., 2019a). Because machine learning and NLP models have historically been so data hungry, large sample sizes have been prioritized over sample quality and diversity (Zirikly et al., 2019). As an unintended consequence, large health datasets may not be representative of

their respective target populations (Widner et al., 2023; Majdik et al., 2024). At the same time, small datasets that are diverse and include high quality annotations may not provide enough signal for machine learning and NLP models to effectively learn generalizable relationships (Kaplan et al., 2020). Striking a balance between dataset size, quality, and diversity remains a remarkable challenge in the health domain, and in turn suggests that distribution shift is likely to be a concern in most applications.

Privacy Concerns. Health data is inherently sensitive; knowledge of an individual's health status may be used (unconsciously or consciously) in a manner that negatively affects the individual's livelihood. For example, knowing that an individual lives with a chronic condition may affect decisions related to their employment (Gouvier et al., 2003; Carolan et al., 2020). Alternatively, knowing that an individual has a mental health condition may affect their ability to interact with their community (e.g., due to stigmatization) (Sickel et al., 2014). As a consequence, individuals are generally hesitant to grant ML and NLP practitioners access to their health data due to the risks associated with leakage (Bell et al., 2014; Kalkman et al., 2022). At the same time, when individuals do consent to sharing their health data with ML and NLP practitioners, Health Insurance Portability and Accountability Act (HIPAA) regulations make accessing such data logistically challenging (e.g., security requirements, limits on sharing and length of access) (Shah and Khan, 2020).

The significant gap in the number of datasets identified during the literature search and the number of datasets that were actually accessible highlights the degree to which data privacy remains a concern in the health domain. Datasets that were accessible were often those constructed with less reliable annotation mechanisms (e.g., proxy-based labels as opposed to clinical measures) or leveraged selection criteria that restricted the representativeness of the sample (e.g., participation in select online communities). Moreover, when datasets did use clinically valid annotation mechanisms (i.e., diagnoses documented by a health professional), the individuals who consented to sharing their data were likely to systematically vary from the broader population with that diagnosis (e.g., due to opt-in bias, due to biases related to treatment access).

4.8 Looking Ahead

The use of sub-optimal dataset curation mechanisms can only allow us to speculate about the presence of selection bias. Additional work must be done to quantify and understand how these mechanisms actually affect the data we collect. Techniques for doing so are the focus of the next two chapters. Such methods will likely have broad and lasting utility given that the aforementioned data collection challenges are unlikely to be easily addressed in the short-term.

Part II: Experimental Methods for Measuring and Interpreting Barriers to AI Robustness in Health Data

Chapter 5

Do Models of Mental Health Based on Social Media Generalize?

5.1 Overview

In the previous chapter, we discussed issues plaguing dataset creation in the health domain and explained how they could introduce selection bias that inhibits model generalization. Nevertheless, simply knowing that data is subject to selection bias or that distribution shift has occurred between training and deployment settings is often not enough to take mitigating action. There also exists a need to understand how selection bias has specifically affected a dataset, and whether alternative curation strategies or training-time corrections could ameliorate the issue(s).

In this chapter based on Harrigian et al. (2020), we use rigorous analytic techniques and context-driven experimental designs to measure and interpret sample bias within 5 social media datasets that are widely used in the research community for the purpose of training depression classification models. We focus on datasets curated using non-clinical, proxy-based labels – the status quo method of curating health datasets in the absence of clinical ground truth. Our study provides evidence that proxy-based datasets only represent a subset of the broader population they are interested in modeling and that insufficiently sampled control groups can arbitrarily inflate estimates of model performance.

5.2 Background

Computational methods have been successful at identifying mental health disorders based on an individual's language across multiple modalities and domains, such as social media (Mowery et al., 2016; Morales et al., 2017), speech (Iter et al., 2018) and other writings (Kayi et al., 2017; Just et al., 2019). While early work in this research area leveraged traditional human subject study designs in which individuals with clinically validated psychiatric diagnoses volunteered their language data to train classifiers and perform quantitative analyses (Rude et al., 2004; Jarrold et al., 2010), contemporary work has focused primarily on extracting signal from large, non-clinical datasets annotated via automated mechanisms (Coppersmith et al., 2015b; Winata et al., 2018). Social media has been the main target of such efforts due to its widespread public adoption and, until recently, liberal data collection policies (Davidson et al., 2023; Pfeffer et al., 2023).

Automatic annotation techniques have been prioritized due to the challenges associated with acquiring sensitive health data at a scale that supports the training and evaluation of machine learning models (e.g., Chapter 4). At a high-level, studies leveraging non-clinical, proxy-based annotations of mental health status have supported their design by demonstrating alignment with existing psychological theory regarding language usage by individuals living with a mental health disorder

(Cavazos-Rehg et al., 2016; Vedula and Parthasarathy, 2017). For example, feature analyses of trained NLP models have highlighted higher amounts of negative affect (Park et al., 2012), increased personal pronoun prevalence (De Choudhury et al., 2013b), decreased second-person pronoun usage (Vedula and Parthasarathy, 2017), and non-trivial thematic differences (Cavazos-Rehg et al., 2016) in data generated by individuals living with depression compared to control samples. These types of findings have emerged across studies regardless of computational model complexity (Orabi et al., 2018; Song et al., 2018), formality of language (Coppersmith et al., 2015d; Rathner et al., 2018), and decisions to filter explicit mental health status signal (Wolohan et al., 2018b).

The ultimate goal of these efforts has been threefold – to better personalize psychiatric care, to enable early intervention, and to monitor population-level health outcomes in real time. All of these objectives have been touted as an opportunity to improve health equity. For example, by collecting data about otherwise difficult to reach populations (Saha et al., 2019; Cascalheira et al., 2024), or alternatively, by addressing diagnostic and treatment disparities that arise due to challenges related to access (Jaidka et al., 2021; Henning-Smith et al., 2023) and cultural stigmatization (Pendse et al., 2019). Nonetheless, research has largely trudged forward without stopping to ask one critical question: do models of mental health conditions trained on

automatically annotated social media data actually generalize to new data platforms and populations?

5.3 Motivation and Contribution

Typically, the answer to the aforementioned question is no – or at least not without additional effort. As discussed in Chapter 3, performance loss is to be expected in a variety of scenarios due to underlying distributional shifts (e.g., domain transfer (Shimodaira, 2000; Subbaswamy and Saria, 2020)). It is unclear to what extent factors specific to the intersection of mental health and social media require tailored intervention. In this study, we demonstrate that, at a baseline, proxy-based models of mental health status *do not* transfer well to other datasets annotated via automated mechanisms. Perhaps more importantly, we outline a series of experimental methods that allow practitioners to understand *why* some forms of proxy-based health status annotations can inhibit generalization.

5.4 Related Work

We are not the first group to raise concerns about distribution shift and model robustness in the mental health space. For example, variation in the clinical

presentation of mental health conditions across different populations – e.g., genders, ages, and even personality types – has been cited as a potential concern in existing datasets (Mitchell et al., 2009; Cummins et al., 2015; Preoţiuc-Pietro et al., 2015; Arseniev-Koehler et al., 2018; Shing et al., 2018). Moreover, it has been shown that diverse and sometimes conflicting views humans have regarding mental health conditions can make obtaining reliable gold-standard labels fundamentally challenging and lead to degradation in downstream model performance (Liu et al., 2017b). And lastly, when proxy-based and automatic annotation mechanisms are used to label social media data, there exists a strong risk that the signal used to facilitate the labeling is not representative of the larger population of interest (Lippincott and Carrell, 2018; Amir et al., 2019).

Despite these concerns, research that attempts to quantitatively *measure* and *understand* robustness of mental health status classifiers in social media has been relatively limited. Some researchers have focused on generalization across different communities of the same social media platform. For example, in a within-subject analysis, Ireland and Iserman (2018) examined differences in language usage by Reddit users who had posted in an anxiety support forum within and outside mental health forums. Concurrently, Wolohan et al. (2018b) explored the predictive power of models trained to detect depression within Reddit users as a function of access to text

from explicit mental health related subreddits. Both studies highlighted a mitigation of overt mental health discussion outside of the support forums, but still detected linguistic nuances in individuals with an affiliation to the mental health forums.

Meanwhile, other researchers have focused on generalization across different social media platforms altogether. For example, Shen et al. (2018) attempted to use transfer learning with large amounts of English Twitter data annotated with individual-level depression labels to improve predictive performance of depression classifiers in Chinese Weibo data. Using the English and Chinese versions of the Linguistic Inquiry and Word Count tool (LIWC) (Pennebaker et al., 2001; Huang et al., 2012) in conjunction with other modalities of social data (e.g., profile metadata, images), the authors showed that signal from Twitter was useful for classification on Weibo.

Recent work from Ernala et al. (2019) is perhaps the most similar to our own. Ernala et al. (2019) leverage multiple different annotation mechanisms to train Twitter-based models for identifying schizophrenia and then apply them to Facebook data from an independent population of clinically diagnosed schizophrenia patients. Three different types of proxy signals with varying degrees of manual supervision were each found to generalize poorly to the clinically-verified population. While the authors' analysis suggested the domains were similar enough to justify transfer attempts, only limited post-hoc analysis of the data platform effect was carried out.

Thus, it remains unclear to what extent the annotation methodologies as opposed to platform effects (or other confounds) caused the degradation. One objective of our study is to elucidate this lack of clarity.

5.5 Data

We select depression classification as our task because it is widely studied, has multiple datasets from different platforms, and is of critical importance to society. Estimated to affect 4.4% of the global population, depression presents a significant economic burden and remains the most common psychiatric disorder associated with deaths by suicide (Hawton et al., 2013; Organization et al., 2017). Occupying a lion's share of the computational mental health literature, depression classification is a critical first target for evaluating generalization of mental health models in social media (Chancellor and De Choudhury, 2020).

To quantify and interpret the nature of domain transfer loss, we consider five datasets. Datasets were selected based on their common adoption in the literature (Preoţiuc-Pietro et al., 2015; Gamaarachchige and Inkpen, 2019) and their use of proxy-based annotations (Coppersmith et al., 2014a). We use two Twitter datasets – *CLPsych 2015 Shared Task* (Coppersmith et al., 2015d), *Multi-Task Learning* (Benton et al., 2017b) – and three Reddit datasets – *RSDD* (Yates et al., 2017), *SMHD*

Dataset	Platform	Years	Size (Individuals)	Annotation Mechanism
CLPsych	Twitter	2011-2014	Control: 477 Depression: 477	Regular expressions; Manual verification; Age- & gender-matched controls
Multi-Task Learning	Twitter	2013-2016	Control: 1,400 Depression: 1,400	Regular expressions; Manual verification; Age- & gender-matched controls
RSDD	Reddit	2008-2017	Control: 107,274 Depression: 9,210	Regular expressions; Manual verification; Subreddit-based controls
SMHD	Reddit	2010-2018	Control: 127,251 Depression: 14,139	Regular expressions; Subreddit-based controls
Topic-Restricted Text	Reddit	2014-2020	Control: 7,016 Depression: 6,853	Community participation

Table 5.1: Summary statistics for each dataset considered in our study. All datasets leverage proxy-based annotations of depression diagnoses in lieu of clinically-validated annotations of depression diagnoses. The sample size, class balance, and date range varies significantly between datasets.

(Cohan et al., 2018), and *Topic-Restricted Text* (Wolohan et al., 2018b). Table 5.1 presents summary statistics for the datasets, while additional construction details are described below.

CLPsych 2015 Shared Task. This Twitter dataset was constructed using regular expressions matching phrases similar to "I was diagnosed with depression" by Coppersmith et al. (2015d). The authors manually verified the authenticity of each candidate self-disclosure and then sampled an age- and gender-matched "control" population using tweet-based inferences (Schwartz et al., 2013). To approximate the depression-control pairs in the original dataset, which has since been anonymized, we sampled from the full set of available control group candidates based on their inferred demographics.

Multi-Task Learning. Compiled by Benton et al. (2017b), this Twitter dataset for multiple mental health disorders was constructed in the same manner as the CLPsych 2015 Shared Task.[1] Although depression-control linkages remain in our version of the dataset, we only use them to isolate an appropriate control group for the depression group. Individuals who were annotated as part of both the Multi-Task Learning and the CLPsych 2015 Shared Task data were removed from the CLPsych data (55 depression, 0 control).

RSDD. The Reddit Self-disclosed Depression Diagnosis (RSDD) dataset is a Reddit-based data asset in which individuals who self-disclosed they were living with depression were identified via regular expressions and manually verified much like the two aforementioned Twitter datasets (Yates et al., 2017). Individuals selected for the control group were required not to have posted in a list of 24 mental health related subreddits or to have used any of 19 mental health terms. To align the theme of language generated by individuals across classification groups, each individual in the depression group was greedily matched with 12 individuals from the candidate control pool based on Hellinger distance between each individual's post distribution over subreddits. To preserve

[1] While we were able to reproduce the class distributions of the dataset described in Benton et al. (2017b), we identified discrepancies between the dates that tweets in this version of the dataset were posted relative to the dates that the original component datasets were published.

privacy of individuals within the dataset, usernames were anonymized and post metadata was redacted. Accordingly, linkages between each individual within the depression group and their respective control group pairs could not be recreated.

SMHD. The Self-Reported Mental Health Diagnoses (SMHD) dataset was constructed in a similar manner as RSDD, albeit being expanded to support 9 conditions, leverage more precise regular expressions, and abide by a more conservative term/subreddit filter set (Cohan et al., 2018). As with RSDD, linkages between individuals in the depression group and their controls were not preserved in our version of the dataset nor could they be readily reproduced. A substantial portion of individuals in SMHD are also part of RSDD; for this reason, we refrain from conducting domain transfer experiments between the two datasets.

Topic-Restricted Text. To expand the scope of our analysis, we follow methods described in Wolohan et al. (2018b) to curate an additional Reddit dataset in which annotations are assigned based on community participation and explicit mental health signal is removed (hence "topic-restricted text"). Per the original paper, individuals who initiated one of 10k recent posts in r/depression were considered members of the depression group, while individuals who initiated one of 10k recent posts in r/AskReddit (but not in the recent

r/depression query) were considered to be members of the control group. Due to the anonymous nature of the RSDD and SMHD datasets, we were unable to determine if any individuals found within the Topic-Restricted Text dataset were also in RSDD or SMHD.

5.5.1 Preprocessing

To maintain our ability to interpret results consistently, the same preprocessing pipeline was applied across all datasets. Additional effort was undertaken to limit each proxy-based label mechanism from artificially influencing predictive performance.

Disclosure Artifacts. Each dataset was curated in part by a system of simple rules (e.g., matches to "I was diagnosed with depression," participation in a depression support forum). While these heuristics are useful for identifying candidates to include within each dataset, they also risk introducing bias that may render the modeling task trivial. For example, individuals who disclose a depression diagnosis are likely to also share their experience with other psychiatric conditions (Benton et al., 2017b), while language used in dedicated mental-health subreddits systematically differs from the rest of Reddit (De Choudhury and De, 2014; Ireland and Iserman, 2018).

To encourage our mental health classifiers to learn subtle linguistic nuances that cannot be easily captured using straightforward logic, we make efforts to exclude

unambiguous mental health content from all training and evaluation procedures. In line with prior work, we discard posts that include mentions of clinically-defined psychiatric conditions, adopting the list of mental health terms enumerated by Cohan et al. (2018) as a reference. This list (N=458) extends work from Yates et al. (2017) by including disorders tangential to depression, common misspellings, and colloquial references.

Topical Artifacts. As is standard for mental health modeling, we also discard posts made in subreddits dedicated to providing mental health support (Yates et al., 2017; Cohan et al., 2018; Wolohan et al., 2018b). Since new subreddits are created daily and our version of the Topic-Restricted Text dataset contains posts made after collection of RSDD and SMHD, we create an updated list of mental health support subreddits. To do so, we examine the empirical distribution of posts amongst subreddits within the Topic-Restricted Text dataset and rank each subreddit S based on pointwise mutual information (PMI) for the depression group D, $\log(p(S|D)/p(S))$. We manually examined the top 1000 subreddits based on PMI and identified all subreddits whose description affirmed an association to mental health.

Our list (N=242) expands existing resources from Yates et al. (2017) and Cohan et al. (2018) by providing 162 additional mental health subreddits, many of which were

actually created before the collection of RSDD and SMHD.[2] While this step diminishes the risk of mental health content saturating the Topic-Restricted Text dataset, the list's expansion beyond that of the RSDD and SMHD lists suggests that the former two Reddit datasets may indeed still have overt mental health content. We explore how different degrees of subreddit-based filtering may affect generalization in §5.8.4.

Temporal Artifacts. To constrain models from leveraging time-related artifacts as spurious signal, all datasets were truncated in time so that at least 100 unique data points (e.g., Tweets, Reddit comments) were present in the first and final month across individuals in both classes. Date ranges selected based on this criteria are presented in Table 5.1.

Tokenization. Text within both Tweets and Reddit comments was tokenized using a modified version of the Twokenizer (O'Connor et al., 2010). English contractions were expanded, while specific retweet tokens, username mentions, URLs, and numeric values were replaced by generic tokens. As pronoun usage tends to differ in individuals living with depression (Vedula and Parthasarathy, 2017), we removed any English pronouns from our stop word set.[3] Case was standardized across all tokens, with a single flag included if an entire post was made in uppercase letters.

[2] Subreddits and code are made available to other researchers: https://github.com/kharrigian/emnlp-2020-mental-health-generalization

[3] English Stop Words (nltk.org)

Feature Engineering. Text from all documents for an individual is concatenated together and tokenized using the procedure described above. The vocabulary of each training procedure is fixed to a maximum of 100-thousand unigrams selected based on KL-divergence of the class-unigram distribution with the class-distribution of stop words (Chang et al., 2012a). This reduced bag-of-words representation is then used to generate the following additional feature dimensions: a 50-dimensional LDA topic distribution (Blei et al., 2003), a 64-dimensional LIWC category distribution (Tausczik and Pennebaker, 2010), and a 200-dimensional mean-pooled vector of pretrained GloVe embeddings (Pennington et al., 2014). The reduced bag-of-words representation is transformed using TF-IDF weighting (Ramos et al., 2003).[4]

5.6 Models

We begin by training depression classification models on each dataset. All classification experiments leverage the same training procedure and features (see §5.5.1). As a model architecture, we use ℓ_2-regularized logistic regression. Despite the model's relative simplicity, we are able to achieve respectable classification performance while maintaining an ability to interpret learned parameters. Logistic regression has served

[4] All data-specific feature transformations (e.g., LDA, TF-IDF) are learned without access to development or test data. We use Scikit-learn's implementations of LDA and TF-IDF.

as a difficult benchmark to beat given access to appropriate engineered features for prior mental health studies (Benton et al., 2017b).

5.6.1 Model Validation

To validate our modeling framework against prior work, we first establish within-domain predictive baselines. This step also allows us to contextualize performance by estimating the intrinsic difficulty of modeling each dataset (DeMasi et al., 2017).

Methods. We use train/development/test splits if they have been established by the dataset distributor; otherwise, we sample 20% from the available data to be used as a held-out test set and then create an additional 80/20 train/dev split using the remaining data. For each dataset, we use an independent grid search to select regularization strength C amongst {1e-3, 1e-2, 1-e1, 1, 10, 100, 1e3, 1e4, 1e5} that maximizes F1 in the dataset's development split. We use a binarization threshold of 0.5 (noninclusive) for all datasets.

Results. Our logistic regression models perform on par with prior research for the two Twitter datasets and the Topic-Restricted Text dataset. Our logistic regression models for RSDD and SMHD slightly improve upon previously reported logistic regression scores, but are inferior to neural methods. The latter is likely a

	Train Data	Test Data				
		CLPsych	Multi-Task	RSDD	SMHD	Topic-Restricted Text
Balanced & Downsampled	CLPsych	**77.4 (0.9)**	63.5 (5.4)	16.9 (1.1)	6.4 (0.6)	63.8 (3.4)
	Multi-Task	55.3 (11.1)	**80.2 (1.8)**	14.9 (0.1)	5.4 (0.0)	64.8 (0.7)
	RSDD	24.7 (3.4)	33.8 (4.1)	**33.8** (1.0)	—	48.7 (4.6)
	SMHD	33.5 (4.8)	54.3 (4.0)	—	**18.6 (0.7)**	62.6 (1.1)
	Topic-Restricted Text	62.4 (1.8)	51.6 (6.0)	17.3 (1.7)	10.5 (1.4)	**68.6** (0.7)
Balanced Only	CLPsych	**77.4 (0.9)**	63.5 (5.4)	16.9 (1.1)	6.4 (0.6)	63.8 (3.4)
	Multi-Task	73.9 (0.4)	**83.0 (0.5)**	14.9 (0.1)	5.4 (0.1)	65.5 (1.1)
	RSDD	28.4 (4.6)	40.7 (5.1)	**40.5 (0.3)**	—	43.4 (0.3)
	SMHD	35.5 (2.8)	46.4 (2.8)	—	**21.2 (0.6)**	63.1 (0.7)
	Topic-Restricted Text	66.8 (0.8)	64.8 (2.6)	21.8 (0.4)	10.6 (0.8)	**73.5 (0.2)**
Model Validation (§5.6.1)		77	82	59	38	75

Table 5.2: Mean F1 scores (and standard deviations) for the model validation and transfer experiments. Increasing dataset size (10× in some cases) does *not* unanimously improve transfer. Baselines described in §5.6.1, which preserve any class imbalance during training, are presented in the bottom row. The significant difference in performance between baseline and transfer settings for the RSDD and SMHD datasets suggests that prior shift causes calibration issues.

consequence of differences in feature engineering between our study and prior work (i.e., vocabulary selection, dimensionality reduction).

5.7 Transfer Experiments

We conduct a series of experiments to measure the generalization of models between depression datasets and understand sources of model degradation. Because task formulation and dataset design varies between each dataset (Morales et al., 2017; Chancellor and De Choudhury, 2020), we use these experiments in part as an

opportunity to rule out trivial causes of performance loss.

5.7.1 Baselines

To establish fair baselines, we perform an initial set of transfer experiments that standardize sample size and class balance across datasets. Differences in training sample size could give an unfair advantage to certain models, while variation in class distribution (i.e., prior shift) could promote miscalibration. We hypothesize that there exists a drop in performance when transferring between domains, and further that these data characteristics alone cannot explain the degradation.

Methods. We consider two experimental designs. In the first experiment, we downsample all datasets to have the same training/development size of the smallest class in the smallest dataset (i.e., CLPsych). In the second experiment, we still balance class distributions for each dataset, but allow sample size to vary between datasets. Whereas the former experiment establishes a reasonably fair set of performance baselines, the latter experiment enables us to explore whether access to additional training data ameliorates transfer loss.

For both experiments, we start by combining training and development splits. Then, for each dataset, we sample from the combined splits based on the parameters of the experiment and split the resulting sample into 5 class-stratified folds. We train

5 classifiers per dataset, using 4 folds for training each time, and apply the classifiers to each dataset's held-out test set. Since a substantial portion of individuals in SMHD are part of RSDD, we refrain from conducting experiments between the two datasets.

Results. We report F1 score (mean and standard deviation) for both experiments in Table 5.2. In line with existing research, within-domain training outperforms out-of-domain training in each of our datasets for both sampling settings. While additional samples available for training in the second experiment moderately improve within-domain performance, they are not uniformly helpful for mitigating transfer loss to other datasets. Models generally outperform a random classifier at ranking depression risk in out-of-domain transfer scenarios. Despite class balancing, some models appear to be poorly calibrated for new domains; they obtain low F1 scores that belie the class separation ability measured via AUC. Furthermore, we find that models trained on Twitter data transfer to Reddit data better than models in the reverse direction. Not surprisingly given their overlap in training samples, models trained on the SMHD and RSDD datasets transfer to other domains in an equitable manner, trading improvements with each other across transfer settings. Altogether, these results support our hypothesis that models struggle to generalize and that sample size and class balance are not solely responsible for the performance loss.

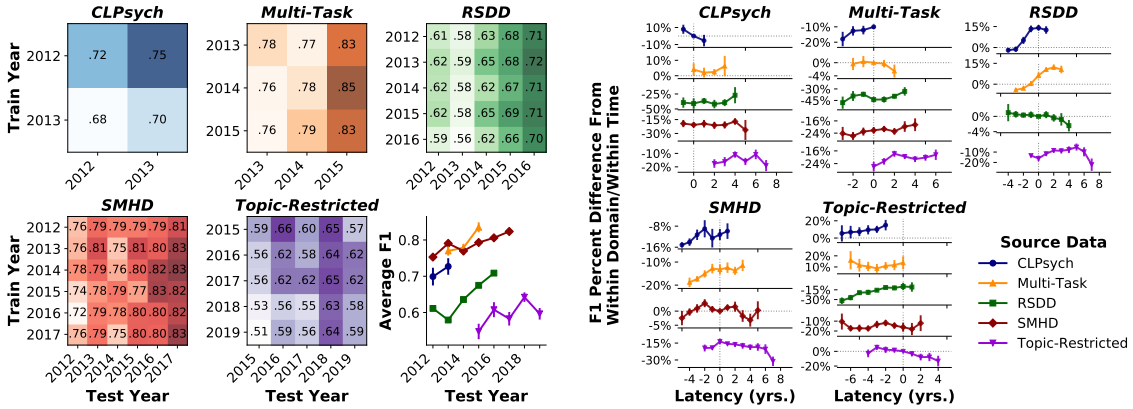


Figure 5.1: Temporal-transfer results. (Left) The mean within-domain F1 score as a function of training and evaluation periods. Predictive performance tends to be better for more recent temporal splits regardless of training period. (Right) The mean percent difference in F1 score relative to each within-domain, no-temporal-misalignment model. Models trained on Twitter data benefit the most from temporal alignment. Performance suffers when applying models trained on more recent data to old data.

5.7.2 Temporality

Prior work has shown that language dynamics over time may hinder models upon deployment (Dredze et al., 2016; Huang and Paul, 2018). In social media, where users adopt new linguistic norms rapidly, performance may be more volatile (Brigadir et al., 2015). We ask whether differences between the time periods covered by each dataset can explain the observed generalization loss.

5.7.2.1 Measuring Temporal Dynamics

As an exercise to understand whether language varies over time within the datasets, we first consider training and evaluating single-domain models with a temporal misalignment between the control and depression groups. By training on mutually-exclusive time periods for each class, we hypothesize the classifier will not only able to learn how to distinguish between groups, but also to distinguish between time periods. If this hypothesis holds true, we expect performance metrics to be artificially inflated when a temporal exclusivity per class exists. This set of experiments is inspired by Ben-David et al. (2006).

Methods. We split each dataset into one-year periods based on the calendar year. For each year, we identify individuals in the Twitter datasets with at least 200 posts and individuals in the Reddit datasets with at least 100 posts.[5] We balance the number of individuals across time periods and groups within each dataset, but allow sample size to vary across datasets. To account for growth in post frequency over time (which increases the number of documents that generate individual feature vectors), we perform additional post-level sampling. We randomly select 200 posts per year in the Twitter datasets and 100 posts per year in the Reddit datasets. Samples of

[5] We use 2x more posts in the Twitter data to account for posts in the Reddit datasets having roughly twice as many words as Tweets do on average.

individuals within each time period are further separated into 5 stratified folds. Folds are established so that individuals in the training data of one time period are never present in the test data of another time period.

To evaluate the degree to which temporal effects are present, we sample groups from all possible combinations of time periods. For example, in one setting, both the control and depression groups are sampled from 2013; in another setting, the control group is sampled from 2013, while the depression group is sampled from 2015. For each combination, we use 4 of the stratified folds for training and use the remaining fold for evaluation, and then repeat the process for all folds. We compare performance when groups are sampled from the same time period against performance when groups are sampled from mutually exclusive time periods.

Results. We achieve a 3-22% average increase in F1 across all datasets when classes are sampled from mutually exclusive time periods instead of being temporally-aligned. The improvement suggests that temporal dynamics are present, as the classifier is able to not only identify signal relevant to classifying depression, but also to classifying data from different periods of time. This result highlights the importance of sampling classes evenly over time.

5.7.2.2 Temporal Effects on Generalization

We now measure the effect temporal dynamics have on out-of-domain performance. We hypothesize model degradation scales as function of the absolute difference in time between training and deployment.

Methods. We use the same data sampling mechanism described in §5.7.2.1. However, we now only consider the case in which control and depression groups are sampled from the same time period. As before, we train a classifier on 4 of the 5 stratified folds for a time period in one dataset. We then evaluate *within-domain* performance using the remaining fold and *out-of-domain* performance using one fold from each time period in the other datasets. We assume ground truth is consistent over multiple time periods.[6]

Results. Examining *within-domain* results in Figure 5.1 (left), predictive performance tends to be better for more recent temporal splits regardless of training period. Classifiers trained on old data (relative to the evaluation period) tend to perform on par with aligned regimens, while classifiers trained on new data show linear losses going back in time. Losses are significant after 2-3 years depending on the dataset.

[6] Given the episodic nature of depression, we recognize this may promote pessimistic results for some periods (Tsakalidis et al., 2018).

Though some trends do emerge, *out-of-domain* performance as a function of the difference in training and deployment time periods is relatively variable. Visualized in Figure 5.1 (right), models trained on the Twitter datasets benefit most from temporal alignment in out-of-domain settings. Models trained on Topic-Restricted Text show significant drop offs in predictive performance when applied to older samples within all Reddit datasets. While models trained on RSDD perform better on Topic-Restricted Text as latency is reduced, models trained on SMHD do not exhibit the same trend.

5.8 Shift Beyond Dataset Design

In the previous section, we identified the degree to which loss occurs under a variety of domain transfer settings, standardizing dataset characteristics (e.g., class balance, temporality) to rule out trivial sources of generalization loss. As expected, differences in the sample characteristics we examined could not account for all performance disparities. In this section, we leverage a suite of clever experiments to identify and interpret selection biases that exist within the language of each of our datasets.

5.8.1 Vocabulary Effects

Differences in vocabulary are one possible barrier to generalization (Serra et al., 2017; Chen and Gomes, 2019; Stojanov et al., 2019). Because users often adopt linguistic conventions that are specific to a given social media platform (e.g., formality, slang) (Baldwin et al., 2013; Liu et al., 2022), or alternatively use the platform as an outlet for conversation on temporally-acute current events (Pramanick et al., 2022), vocabulary differences between the datasets are a natural first area of inspection. We hypothesize that limited feature overlap and poor vocabulary alignment across the datasets hinders out-of-domain generalization.

Methods. We test this hypothesis by computing the Jaccard Similarity (JS) of vocabularies between each dataset. We examine correlations between JS and F1 scores from the out-of-domain transfer experiments presented in §5.7.1.

Results. We find the minimum Jaccard Similarity occurs between the CLPsych and RSDD datasets ($JS = 0.10$) while the maximum occurs between the Topic-Restricted Text and SMHD datasets ($JS = 0.65$).[7] Only a weak correlation between similarity and performance exists (Pearson $\rho < 0.18$), suggesting poor generalization is not solely due to differences in vocabulary. That said, it is

[7] JS is moderately deflated in RSDD due to the dataset's large vocabulary, causing SMHD and Topic-Restricted Text to have the highest similarity instead of SMHD and RSDD.

worth noting that perhaps, unlike some classification tasks, only a few features are informative for depression classification and these do not vary by domain. People may talk about different things on Twitter and Reddit, but the small subset of language relevant to depression is unchanged. If that is true, the correlation between Jaccard Similarity of vocabularies and downstream performance may be overly pessimistic.

5.8.2 Topical and Semantic Effects

Our classification models leverage reduced feature representations in the form of LDA topic-distributions (Blei et al., 2003) and mean-pooled pre-trained GloVe embeddings (Pennington et al., 2014). Designed to capture and reflect semantics, we originally hypothesized these low-dimensional features would mitigate transfer loss that arises due to poor vocabulary alignment. Lacking support from our original out-of-domain transfer results, we now look closer at the themes present within each dataset.

Methods. We identify the unigrams that are most unique to each dataset and group (i.e., Depression vs. Control). For each dataset, we use scores assigned by our KL-divergence-based feature selection method (see §5.5.1) to rank the most informative features per class (Chang et al., 2012a). We jointly examine the top-500 most informative unigrams per class, noting high-level themes common across the datasets.

Results. With respect to similarities, we note that words used in discussion about gender and sexuality are strongly associated with each of the depression groups (e.g., 'cis', 'homophobia', 'masculine'), likely a reflection of marginalized groups being at higher risk of depression (Budge et al., 2013). Also ubiquitous amongst each of the datasets are references to self-injurious behavior (e.g 'wrists', 'self-harm', 'hotline'). Emoji usage, references to athletics ('nbafinals', 'scorer'), and terms reflecting current events are strong indicators of the control group in each dataset.

With respect to differences, associations between word usage and depression are subjectively easier to interpret within the Reddit datasets. For example, discussion of mental-health treatment (e.g., 'counselor', 'therapy', 'wellbutrin') and familial and intimate relationships ('brother-in-law', 'soulmate') are prominent within the Reddit datasets. In contrast, language associated with depression within the Twitter datasets tends to reflect slightly more nuanced elements of the condition – e.g., social inequity ('sexism', '#yesallwomen') and fantasy ('fanfics', 'cosplay', 'villians'). These themes align with empirical findings that women are at a higher risk of depression (Kessler et al., 2003) and depressed individuals often find solace in niche subcultures (Blanco and Barnett, 2014; Bowes et al., 2015). Nonetheless, it is unclear to what extent such discussion captures the core signal related to depression detection; surely, not all individuals who enjoy fantasy are depressed.

Finally, we find several temporally-isolated references within the Twitter datasets (e.g., '#RIPRobinWilliams', '#SDCC'). In the Multi-task Learning dataset, we also observe several "predictive" terms using non-American English (e.g., 'colour', 'favourite') which may represent a geographic bias amongst the sampled individuals. In Chapter 6, we will build a better understanding of these eccentricities.

5.8.3 Lexical Effects

The Linguistic Inquiry and Word Count (LIWC) dictionary has been an effective tool for measuring linguistic-nuances of mental health disorders regardless of textual formality (Mowery et al., 2016; Turcan and McKeown, 2019a). Our version of the dictionary (2007) maps approximately 12k words to 64 dimensions (e.g., negative emotion, leisure) that have been empirically validated to capture an individual's social and psychological states (Tausczik and Pennebaker, 2010).[8] A single LIWC feature value represents the proportion of words used across an individual's post history that match the given LIWC dimension. In the same way that we expect semantic distributions (§5.8.2) to ameliorate transfer loss, we hypothesize that models trained on this representation will be more robust when vocabulary overlap is sparse.

[8] The 2007 version of LIWC has a high similarity with the 2015 version amongst dimensions most strongly associated with depression (Pennebaker et al., 2015).

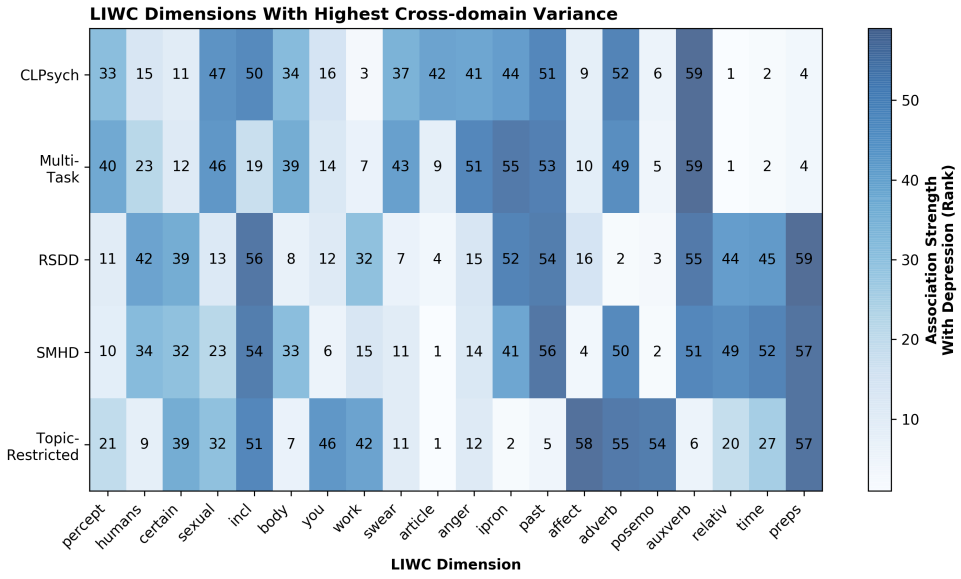


Figure 5.2: The average LIWC-dimension feature rank relative to the Depression group. A higher rank indicates that a LIWC-dimension is more predictive of depression for a given dataset. The 20 dimensions with the most between-dataset variance in feature rank across datasets are presented.

Methods. We explore this hypothesis from three angles: 1) We perform cross-domain transfer experiments using LIWC as the only feature set provided for training and evaluation; 2) We fit LIWC-only classifiers 100 times per dataset using random 70% samples and examine correlations of the learned coefficients; 3) We compute the average feature value of each LIWC dimension per class and measure the difference between classes.

Results. We note that domain-transfer experiments using LIWC as the only feature set maintain high degrees of transfer loss while also sacrificing within-domain performance. Moreover, correlations between coefficients of models between datasets

are relatively low across all comparisons, maxing out at a Spearman R value of 0.338 for the comparison between RSDD and SMHD datasets, which happen to have significant user overlap as is. In general, LIWC coefficients tend to be more correlated within platforms (e.g., CLPsych $\rightarrow$ Multi-Task, SMHD $\rightarrow$ Topic-Restricted Text) than between them (e.g., CLPsych $\rightarrow$ Topic-Restricted Text).

Examination of the underlying class differences provides insight into linguistic differences between each dataset's depression group. In line with prior work, function word use, first-person pronoun use, and cognitive mechanisms are more common within the depression group of each dataset, though their relative prevalence varies. Conversation regarding relativity (i.e., space, motion, time) is strongly associated with the control groups in the Twitter data, but is more associated with the depression groups in the Reddit data. Anger and perceptual topics are more prevalent within the depression groups for Twitter than Reddit. We highlight a subset of the most prominent differences in LIWC's predictive association with depression in Figure 5.2.

5.8.4 Disclosure Effects

Thus far, posts from mental health subreddits and those including mental health terms have been excluded from modeling and analysis. Nonetheless, individuals within each of the depression groups for the Reddit datasets has displayed language

that is unambiguously associated with seeking support or sharing personal experience with mental health issues (e.g., medication, intimate relationships). Accordingly, we hypothesize that existing filters are unable to remove confounds in individuals who disclose a depression diagnosis on Reddit.

Methods. To measure this effect, we examine differences in the distribution of subreddits that individuals in the depression group of the Topic-Restricted Text data post in relative to individuals in the control group. Specifically, we fit a logistic regression model mapping the subreddit distribution of individuals' posts to their mental health status after applying each subreddit filter list (e.g., RSDD, SMHD, Ours). We compare predictive performance of these models and the learned coefficient weights to understand the effect of filtering. As a baseline, we maintain posts from the r/depression subreddit in the feature set. Then, in sequence of coverage from least to most, we apply subreddit filters from RSDD, SMHD, and our study, and measure classification performance. For each filter, we examine the learned coefficient weights to develop a sense for the personality and interests of individuals in the depression group.

Results. The baseline F1 score in the development set maxes out at 0.83 (i.e., when r/depression is included in the feature set), representing the fact that several individuals in the control group had posted in the r/depression subreddit at some point in their history, but were not labeled as having depression due to the sole use of

recent original posts by the automatic annotation procedure. Performance degrades linearly with the expansion of excluded subreddits from each filter, settling at an F1 of 0.72. Coefficients from the model highlight subreddits related to themes of sexuality (r/bisexual, r/actuallesbians), gender (r/ftm), personality (r/introvert, r/INFP), drugs (r/Trees, r/LSD), and relationships (r/MakeNewFriendsHere, r/BreakUps) as being predictive of depression.

The strong classification performance achieved after our filtering measures is evidence that distributional differences in online interaction remain in the "cleaned" Topic-Restricted Text dataset. As our subreddit list is more robust than both the RSDD and SMHD lists, there is reason to believe similar confounds exist in these datasets. The coefficient analysis provides a window into the types of themes and personal identifiers that could incorrectly confuse a classification model during generalization attempts. We could not, for example, apply our model with confidence in a community heavily represented by sexual and gender minorities because it may consistently infer the presence of depression where it is not actually indicated.

5.9 Review of Learnings

In the previous set of experiments, we demonstrated that issues of generalization loss arise in the mental health space, at least for the proxy-based social media datasets

considered in our study. More importantly, we identified *why* this occurs – there exist confounds in the datasets that emerge as a result of each dataset's respective design and sample selection process. So what do we do about it?

Quality Over Quantity. *Practitioners should focus on curating representative and unbiased datasets before considering scale.* Increasing sample size of a mental health dataset does not inherently guarantee better performance in the out-of-domain setting. In our initial domain-transfer experiments, we note that increasing the sample size of the RSDD and SMHD Reddit datasets by over 4x provides no consistent improvement in predictive accuracy when applied to the Twitter datasets (and in some cases even degrades performance). The same holds when we increase the size of the Multi-Task Learning dataset 3x.

Topical Alignment. *Practitioners must account for self-disclosure bias and confounds of personality when curating new datasets.* First discussed in §5.8.2, models trained on the Reddit datasets learn dependencies between support-driven topics, such as medication usage and relationship advice, and depression. In contrast, models trained on the Twitter datasets identify the same correlations between sexuality, gender, and depression that Reddit-based models detect, but also learn about the recreational outlets (i.e., fantasy) and social concerns (i.e., racism, sexism) common in depressed individuals.

We hypothesize that semantic divergences reflect self-disclosure bias and differences in platform interaction patterns (Malik et al., 2015; Shelton et al., 2015). Twitter's status- and reply-based structure serves as a place for individuals to share personal thoughts and experiences in reaction to their daily life. Meanwhile, Reddit's community-based forums require active engagement with specific topics and may silo individuals who wish to discuss their mental health beyond defined areas. The latter gains support from our analysis of subreddit distributions in the Topic-Restricted Text data (§5.8.4).

On one hand, topical nuances in language may appropriately reflect elements of identity associated with mental health disorders (i.e., traumatic experiences, coping mechanisms). However, if not contextualized appropriately during model training, this type of signal has the potential to raise several false alarms upon application to new populations. Accordingly, we urge researchers to minimize the presence of overt topical disparities between classes in their datasets.

Mitigating Temporal Artifacts. *Practitioners must take steps to remove temporal artifacts in new datasets.* Experiments conducted in §5.7.2 reveal that group-based temporal alignment and latency between model training and deployment can have a significant effect on predictive performance. Variability of performance over time is surprising, as there is no clinical evidence to suggest that the underlying

symptoms of depression (on a population level) change over time (APA, 2013).

We hypothesize two reasons for this observation. First, since depression presents in an episodic manner, we may expect data closest to the date of annotation to be the most predictive of an individual's labeled mental status (Melartin et al., 2004). If most posts used for annotation occurred in recent time windows, then it is possible that content in older posts is less relevant to the depressive state of individuals in our data sets. Second, and more problematic, is the possibility that signal used by our classifiers is only a spurious correlation.

At a bare minimum, our results highlight the importance of sampling classification groups so that post volume is equal over time. Discrepancies may wrongly suggest that temporal artifacts are useful for detecting mental health disorders. Going further, researchers should remove temporally-specific references and minimize highly-dynamic language in their datasets. Avenues for accomplishing the latter include using NER to redact n-grams that serve as spurious correlations (Ritter et al., 2011) and leveraging adversarial training to evaluate the degree to which mental health signal may be learned without a notion for time (Tzeng et al., 2017).

5.10 Limitations

Though our study provides a robust perspective towards understanding generalization capabilities of mental health classifiers for social media, we recognize that more learning opportunities exist. Our study only considers a handful of datasets, two platforms, a single mental health disorder, and homogeneous annotation mechanisms. Still unexplored, in large part due to the precautions necessary for securing sensitive mental health data, is how well models trained on data from actual clinical populations generalize to proxy-based datasets and other clinical populations. While high co-morbidity rates between depression and other mental health disorders may allow us to infer model behavior for alternative conditions, we also recognize that presentations of different psychiatric disorders can be quite variable and warrant their own research (Benton et al., 2017b; Arseniev-Koehler et al., 2018).

Another limitation in our work is the lack of depression to control group matches from original reference material. Preoţiuc-Pietro et al. (2015) and De Choudhury et al. (2017) demonstrate that mental health disorders such as depression can have variable presentations based on demographic attributes. The attributes used to construct our Twitter datasets originally were inferred via now-outdated text-based models. Accordingly, demographic inference errors may be propagated to and correlated with depression classification errors. Moreover, these attributes were not considered within

the construction of any of the Reddit datasets we explored. The effect of demographics on generalization remains a valuable insight for future exploration.

Finally, our attempts at domain transfer are constrained. Namely, we do not invoke explicit domain adaptation methods (Peng and Dredze, 2017; Li et al., 2018a; Huang and Paul, 2019). It remains important to test whether existing domain adaptation and generalization methods are sufficient for handling these known instances of distribution shift and selection bias.

5.11 Ethical Considerations

Given the sensitive nature of data containing mental health status of individuals, additional precautions based on guidance from Benton et al. (2017a) were taken during all data collection and analysis procedures. Data sourced from external research groups was retrieved according to each dataset's respective data usage policy. The research was deemed exempt from review by our Institutional Review Board (IRB) under 45 CFR § 46.104.

5.12 Discussion

When working in the health domain, it can be tempting to perform superficial analyses to understand what models learn and then look for connections with clinical knowledge as verification of data quality. Such an approach, however, is potentially subject to confirmation bias. As shown particularly by the feature analyses conducted above (e.g., §5.8.2 and §5.8.4), it's quite easy to come up with reasonable explanations for the signal that a model learns. For example, in isolation, it may appear that discussion about medications or certain niche interests makes clinical sense for people living with depression. However, are these features the core signal we are interested in detecting? Or do they constitute a shortcut for models that evolves due to the manner in which a dataset was collected? Would the same signal have been useful had we started with a purely random sample of social media users?

Multi-dataset experiments such as our own have significant utility in highlighting signal that may or may not be "core" to a modeling problem. At the same time, we fully recognize that access to multiple datasets in the health domain for a given task is not always practical. Accordingly, we note that several experiments we present in the study above do not necessarily require additional data sources. For example, feature analyses of the Twitter datasets revealed temporally-acute events, which a practitioner may reasonably flag as a concern. Likewise, metadata-driven experiments, such as

using subreddit activity to predict mental health status instead of text, revealed sample characteristics that could limit generalization within certain historically marginalized communities. In lieu of multiple datasets, clever, domain-informed experimental design can go a long way in elucidating possible selection bias.

While our study focused on a particular application and only a handful of datasets, we argue that it reveals more general issues with proxy-based annotation mechanisms in the health domain. Automatically assigning health-related labels to data typically requires potentially oversimplified assumptions, which in turn introduce artificial, non-generalizable signal that can be difficult to suppress. At the same time, we recognize that such methods, either used in full or in part to create a dataset, are unlikely to vanish completely due to the previously discussed difficulties acquiring health data. The suite of experiments we present in this study can and should be adopted by practitioners across multiple health applications as a sanity check that informs whether a new sample of data contains selection bias.

5.13 Looking Ahead

To further emphasize the utility of clever experimental designs in identifying selection bias, the next chapter will focus more specifically on the case in which a practitioner only has access to a single dataset. How do we reason about generalization and

selection bias when we do not have a clear-cut out-of-domain sample as a reference point?

Chapter 6

Quantifying and Interpreting the Validity of Self-disclosed Depression Diagnoses for Training Mental Health Models

6.1 Overview

Building off of work in Chapter 5, we will now examine biases introduced into depression classification models specifically when using *self-disclosed diagnoses* as ground truth. We provide a novel finding that temporal latency between a past diagnosis disclosure and future language usage incurs a predictive performance penalty, in turn supplementing current scientific knowledge regarding the temporal specificity of a mental health disclosure. We also expose temporal artifacts and personality-related confounds that arise when using self-disclosed depression diagnoses as labels. We accomplish this by introducing a new counterfactual explanation method to facilitate "train-set debugging". Our method is not only computationally efficient, but also effective at enabling efficient and reliable qualitative analysis. This work was originally presented in Harrigian and Dredze (2022b).

6.2 Background

As alluded to in previous chapters, the most significant advances in computational mental health research have not come from improved modeling architectures (Benton et al., 2017b), but rather from methods for curating large-scale datasets which contain robust and clinically-relevant ground truth annotations of mental health

status (Coppersmith et al., 2014a). Use of regular expressions to identify genuine self-disclosures of a psychiatric diagnosis remains one of the most widely adopted annotation mechanisms by the research community (Chancellor and De Choudhury, 2020; Harrigian et al., 2021), offering a relatively reliable proxy in place of clinical measures which are not only costly to collect, but also often unable to be shared beyond a single institution due to patient privacy policies (Macavaney et al., 2021). Datasets leveraging self-disclosed diagnoses as annotations of mental health status have yielded a variety of insights that align with clinical knowledge and psychological theory (Mowery et al., 2017; Lee et al., 2021). However, a growing body of work, including our own, has raised questions about whether such datasets provide sufficient information to train statistical models that generalize to new populations (Harrigian et al., 2020; Aguirre et al., 2021).

6.3 Motivation and Contribution

Despite the prevalence of datasets dependent on self-disclosure, no analyses have considered how associating a single self-disclosed diagnosis label with data from a variable-length period of time may inhibit the learning of robust statistical relationships. If a user tweets a depression diagnosis in 2015, is their data from 2018 still representative of the condition? Presentation of several mental health

conditions change dynamically and (sometimes) precipitously over time (Collishaw et al., 2004). Yet, it remains common in the computational research community to treat mental health conditions as a static attribute with equal relevance at multiple time points (MacAvaney et al., 2018). In reality, it is likely that only a small fraction of an individual's social media activity is appropriate for training optimal classifiers. Moreover, that a mental health status label may be appropriate for only a subset of time suggests that evaluations of longitudinal model generalization as they are traditionally structured in the community may be insufficient (Sadeque et al., 2018).

We ask: to what extent do mental health diagnosis self-disclosures remain valid over time? We focus specifically on extended durations (i.e., multiple years), a setting which has particular relevance to those who wish to estimate generalization strength of their statistical classifiers for use in longitudinal monitoring applications, as well as those interested in updating existing models with new data to mitigate the effects covariate shift (Agarwal and Nenkova, 2022). In reviewing recent online activity from individuals in the 2015 CLPsych Shared Task dataset who disclosed a depression diagnosis on Twitter over five years ago (Coppersmith et al., 2015d), we not only acquire a new understanding of how presentations of mental health status on social media present over time, but also find new evidence to support prior claims regarding the presence of personality-related confounds in datasets curated using self-disclosures

(Preoţiuc-Pietro et al., 2015; Harrigian et al., 2020; Vukojevic and Šnajder, 2021). Our analysis provides critical guidance to practitioners as they curate mental health social media datasets, while also elucidating factors which inhibit robustness in a dataset that remains one of the most widely adopted by the research community. We also introduce a computationally efficient approach for interpreting the relevance of text segments (e.g., documents, sentences) in a larger input space (e.g., post history).

6.4 Related Work

The majority of mental health research based on social media leverages the same experimental design – assume individuals have a fixed mental health status and attempt to infer this latent attribute using historical online activity traces (e.g., posts, follower network dynamics) (Guntuku et al., 2017; Chancellor and De Choudhury, 2020). This training setting is convenient given the inherent complexities of acquiring temporally-granular psychiatric measures at scale (Canzian and Musolesi, 2015). However, the setting implicitly relies on assumptions that are not supported by clinical knowledge regarding psychiatric dynamics (Johnson and Nowak, 2002; Schoevers et al., 2005). Some work has been done to incorporate time-based priors into mental health models, which allow practitioners to train statistical classifiers using a static label while also explicitly accounting for longitudinal variation in label relevance

(Wongkoblap et al., 2019; Uban et al., 2021). Others have eschewed the use of a static label altogether and instead curated datasets that contain multiple points of ground truth mental health status, albeit still with some element of historical data aggregation (Chancellor et al., 2016a).

Temporally-aware classifiers have achieved better performance benchmarks than their static counterparts in some cases (Rao et al., 2020), though these evaluations remain limited by the dearth of data with mental health status annotations at multiple time points. Meanwhile, datasets that do support dynamic evaluation are curated almost exclusively using protected clinical measures (Reece et al., 2017), cost-intensive interviews (Nobles et al., 2018b), or non-trivial shifts in non-language-based online behavior (De Choudhury et al., 2016).

Computational studies that have focused on self-disclosed diagnoses have not comprehensively reviewed how individual activity evolves over long periods of time (Saha et al., 2021). Our study thus fulfills an important void in the research space by providing a new understanding of long term mental health dynamics in social media, and more particularly, within convenience samples curated using self-disclosed diagnoses.

Dataset	Dates	# Users	# Posts
Original	2012 – 2015	Depression: 477 Control: 872	Depression: 1,121,388 Control: 1,907,508
Updated	2012 – 2021	Depression: 444 Control: 172	Depression: 1,372,868 Control: 546,826

Table 6.1: Summary statistics for the original and updated versions of the 2015 CLPsych Shared Task dataset, further stratified by control and depression groups.

6.5 Data

We support our study using a newly updated version of the 2015 CLPsych Shared Task dataset (Coppersmith et al., 2015d) (see §5.5 for details). In line with guidance from Benton et al. (2017a), individual identifiers in the official version of the CLPsych dataset have been anonymized, with linkages between anonymized and de-anonymized identifiers erased in entirety. However, the original de-anonymized identifiers remain available under explicit permission from Coppersmith et al. (2015d), who provided this information to reverse engineer the original anonymization mapping. To do so, we first query up to 3,200 of the most recent tweets from each de-anonymized user identifier using Twitter's public API and further isolate all relevant tweets found in our institution's cache of Twitter's 1% data stream. We identify candidate pairs of anonymized and de-anonymized accounts based on overlap of raw timestamps within the original dataset's collection window. Normalized text (i.e., punctuation removal, case standardization) from candidate pairs is compared using exact matching to verify

final linkages.

Statistics for the original dataset and its updated counterpart are provided in Table 6.1. We find that a majority of accounts that were unable to be linked had significantly smaller activity traces in the original dataset. These accounts are likely to either have been deleted in entirety or to have tweeted with a small enough frequency such that the 1% stream does not contain any samples. The discrepancy in match rates between individuals in the depression and control groups is unfortunately not fully-understood, though discussions with the dataset's authors suggest this may just be an artifact of the original archival process.

6.5.1 Preprocessing

Twitter's language tags and automatic language identification (Lui and Baldwin, 2012) are used to isolate English text. Retweets are excluded to most acutely highlight personal experiences with depression over time. Unless specified otherwise, keyword-based tweet filtering is applied to preemptively mitigate sampling-induced biases which can artificially inflate estimates of predictive performance.

Overt Bias Filtering. Some biases that we filter have been previously recognized and addressed by the research community (e.g., tweets that include diagnosis disclosures and/or mental health related keywords/hashtags) (De Choudhury and

De, 2014), while others have been traditionally overlooked. A preliminary qualitative analysis of influential n-grams and their source tweets reveals a previously unrecognized surplus of "fan accounts" (e.g., supporters of Harry Styles and Demi Lovato) and tweets containing account statistics (e.g., new followers) within the depression cohort. Meanwhile, daily horoscope tweets were identified with an anomalous frequency within the control group. The latter two sources of noise do not have a clear clinical explanation, while the former (i.e., fan accounts) arises in the context of discussion regarding the mental health of young celebrities. Although some of these motifs represent genuine behavioral correlates of depression, their importance in prediction tends to be inflated due to context of the original collection time period. Altogether, these relationships present the first indication that diagnostic annotations based on self-disclosure may yield datasets that inhibit generalization.

6.6 Quantifying Label Validity

Enabling reliable use of statistical models to evaluate change in mental health status remains a core objective for computational researchers (Choi et al., 2020; Fine et al., 2020). Our success in this task domain critically depends on access to ground truth at multiple time points, not only for evaluating generalization error (DeMasi et al., 2017; Tsakalidis et al., 2018), but also for mitigating the effects of distribution shift.

As discussed above, it is often trivial to update activity traces for individuals with a prior mental health diagnosis disclosure. Nonetheless, clinical knowledge suggests original disclosure-based labels may not be relevant over the course of time, either due to a condition's episodic presentations (Angst et al., 2009) or the effects of psychiatric treatment (Saha et al., 2021). We ask whether the CLPsych Shared Task dataset supports this theory.

6.6.1 Methods

A natural framework for answering this inquiry emerges from computational research regarding label noise (Frenay and Verleysen, 2013). Under such a perspective, we can view changes in mental health status as a stochastic process which blindly alters the correctness of class labels over time. The implications of this mechanism allow us to reason about predictive performance of a statistical classifier within and outside of the time period in which it is trained. Differences in within-time-period performance for two different time periods may be caused by two factors – different levels of label noise and/or different signal-to-noise ratios. Meanwhile, degradation in performance when transferring a classifier from one time period to another may be caused by three possible factors – label noise in the source time period, label noise in the target time period, or distributional shift between the time periods. Although isolated differences

in predictive performance in a longitudinal setting do not implicate a single causal factor, multiple comparisons taken together may allow us to reason about underlying changes in the data.

This logic guides our search for evidence in support of the hypothesis that mental health annotations cannot be treated as fixed attributes. We consider a standard longitudinal domain transfer setup (Huang and Paul, 2019), chunking the CLPsych dataset into three discrete three-year periods[1] (2012–2015, 2015–2018, 2018–2021) and evaluating within- and between-time-period predictive performance for all available pairs. We use Monte Carlo Cross Validation (Xu and Liang, 2001) to obtain estimates of predictive generalization, chosen over alternative protocols that would be unreliable given the limited sample size of the updated CLPsych dataset (Varoquaux, 2018).

Each iteration of the cross validation procedure (1,000 total) begins by randomly splitting individuals into a 60/40 train/test split, with control and depression groups demographically aligned[2] using propensity scores (Imbens and Rubin, 2015). To control for differences in data availability between time periods, we not only constrain the sampling process such that splits have an *equal class balance*, but also that individual-level representations are constructed using an *equal document*

[1] Time periods were chosen to maximize the number of discrete windows while ensuring enough posts were available to construct informative individual-level representations.

[2] We align samples to have a similar joint age and gender distribution.

	Test		
Train	2012-2015	2015-2018	2018-2021
2012-2015	$.71_{(.70,.72)}$	$.66_{(.65,.66)}$	$.69_{(.68,.70)}$
2015-2018	$.66_{(.65,.67)}$	$.66_{(.65,.66)}$	$.68_{(.67,.69)}$
2018-2021	$.65_{(.65,.66)}$	$.67_{(.66,.68)}$	$.68_{(.67,.69)}$

Table 6.2: Mean test-set area under the curve (AUC) and 95% confidence intervals across 1,000 Monte Carlo cross validation iterations. Within-time-period performance is significantly higher for the original diagnosis disclosure window than in subsequent time periods.

history size (250 randomly-sampled posts from each time period). A single binary logistic regression classifier provided with document-term TF-IDF representations (Baeza-Yates, Ribeiro-Neto, et al., 1999) is fit for each time period using data from individuals in the training set. Each classifier is applied to all three time periods, evaluating performance using individuals in the sampled test set.

6.6.2 Results

We report the average test set area under the curve (AUC) and 95% confidence intervals for each discrete time period pairing in Table 6.2. Focusing first on *within-time-period* performance (top left to bottom right diagonal), we find that within-time-period performance is significantly higher in the dataset's original time period (2012-2015) than within subsequent time periods. This holds true even when running experiments only with individuals that have large post histories in the new

time periods, demonstrating that the outcome is not an artifact of survivor bias. At a high level, the differences in within-time-period performance suggest that either label noise has increased or that the signal-to-noise ratio has decreased over time.

Unfortunately, examination of *between-time-period* generalization does not conclusively resolve which of these two factors are responsible for the variation. Focusing first on models trained using data from older time periods (top right triangle), we do not observe any significant difference in predictive performance compared to the benchmarks established by models trained and deployed during the same time period. This serves as a contrast to models deployed on older data (bottom left triangle), where we note that classifiers trained on both of the new time periods incur a loss when being applied to the original CLPsych dataset time period. Interestingly, the absolute differences in performance are minimal. We note that the coefficients of the logistic regression classifiers from each independent time period exhibit significantly positive Pearson correlations, ranging from 0.47 to 0.52, and in turn promote stable performance.

Discussion. Although these experiments have not conclusively answered our primary research question regarding longitudinal label validity, they have provided evidence that not all time periods of data are equally informative for training a robust depression classifier. Critically, these results suggest that practitioners cannot assume

it better to train a depression classifier using new data, which may be more relevant to their deployment scenario, if it means potentially compromising the temporal relevance of the original ground truth annotations.

What remains to be understood is *why* the predictive task appears to become more difficult in the updated time periods at a statistically significant level, but not one that would necessarily raise immediate concerns to a practitioner. Had underlying dynamics significantly changed since the original data collection period, we would have expected to see a more dramatic loss in predictive performance. Has the mental health status for these individuals genuinely remained static, or is there a spurious confound in the data inflating our performance estimates?

6.7 Interpreting Label Validity

We attempt to better understand the variation in predictive performance estimated above by comparing language within the updated dataset to the original CLPsych sample. In particular, we adopt a mixed methods approach that allows us to estimate changes in the proportion of depression labels which remain relevant in the updated dataset, and to qualitatively summarize drivers of model decision-making across time periods. We support our analysis by manually coding content-related motifs within a large sample of document histories in the updated dataset, focusing primarily on

criteria for diagnosing depression as defined within the DSM-5 (APA, 2013). We draw inspiration from the growing literature on "train-set debugging" (Koh and Liang, 2017; Han et al., 2020), which leverages instance attribution and other diagnostics to succinctly interpret the relationship between training data, learned model parameters, and downstream predictions.

6.7.1 Methods

An annotator is presented with up to 30 anonymized tweets made by a single individual during one of the time periods and asked to indicate whether the individual exhibits evidence of depression. The annotator must mark one of four options – Uncertain, No Evidence, Some Evidence (Moderate Confidence), Strong Evidence (High Confidence). Explicit disclosures of a depression diagnosis and references to living with depression are automatically assigned to the Strong Evidence category. Otherwise, the annotator is instructed to indicate their confidence based on the nine DSM-5 criteria for diagnosing depression (APA, 2013) and their prior knowledge regarding the presentation of mental health conditions within social media. If at least some evidence of a depression diagnosis is indicated, the annotator is asked to identify whether the depression appears to be in remission (e.g., discussion of overcoming depression). They are also asked to indicate which DSM-5 criteria and/or prior

knowledge was used to inform their decision, along with any other notable thematic content. My advisor, Mark Dredze, and I worked together to develop a common mental model for identifying DSM-5 criteria and other common linguistic motifs in the text. During a pilot round of coding, 16 thematic patterns were identified within the annotated instances to complement the original DSM-5 criteria. Exemplary tweets (paraphrased non-trivially to preserve anonymity (Ayers et al., 2018)) for each of the DSM-5 criteria and alternative thematic categories are provided in Table 6.3.

Our goal of this analysis is *not* to make diagnostic claims regarding the mental health status of individuals in our dataset, but rather to broadly understand *what* the statistical classifiers are learning. Accordingly, tweets presented to the annotator are those that had the largest positive effect on the classifier's estimated probability of depression. We introduce a new counterfactual explanation measure to facilitate this exercise.[3] Formally, we define the "influence" of a tweet $I(x)$ amongst a set of tweets $x \in X_\tau$ as follows:

$$I(x) = \sum_{k=1}^{K} P_{k,\tau}(y = 1|X_\tau) - P_{k,\tau}(y = 1|X_\tau^{\neg x})$$

where $P_{k,\tau}(\cdot)$ is the probability of depression estimated by a classifier trained on the k-th random sample of data from time period τ, out of K total samples. As was the

[3] Concurrently completed work from Ge et al. (2021) introduced the same measure, which they refer to as a "soft-validity measure."

Evidence	**Exemplary Tweets**
Diagnosis Disclosure	"Bipolar disorder and depression. My doctor finally agrees."
	"I have suffered from depression for several years now"
Depressed & Irritable Mood	"No one ever asks if I'm doing fine."
	"You don't understand what I'm dealing with. Get fucked."
Loss of Interest/Pleasure/Motivation	"...realizing you don't care about the things you used to enjoy"
	"cant get out of bed today"
Weight, Body Image, & Nutrition	"Not that anyone cares, but I'm almost at my goal weight."
	"I bought the dress I've always wanted, but still don't feel pretty."
Sleep Disturbance	"I CANT SLEEP. PAIN. JUST LIKE ALWAYS."
	"Shit! Surviving on only a couple of hours of sleep again :/"
Fatigue	"mentally drained from this pandemic"
	"This should be effortless but I can't work any harder"
Sense of Worthlessness & Guilt	"when you let someone do anything to you..."
	"It truly is always my fault. I probably suck."
Impaired Thought	"I'm failing my classes because I'm depressed."
	"at work. cant focus doe"
Death & Self Harm	"My scars are faded...unless you care to look close"
	"I wish you all never see a loved one fade away."
Cognitive Distortions	"Going to fail this exam. SCREWED."
	"I always think my bf is going to leave me"
Treatment	"Scared to tell a woman that I'm in therapy"
	"Slowly weaning of the prozac."
Gatekeeping	"depression isn't just a bad day. fuck you all."
	"LET ME SHOW YOU WHAT DEPRESSION IS"
Sexuality and Intimacy	"Who wants to come take some pics of me for only fans? ;)"
	"Every girl should watch porn with their bf"
Negative Emotions	"hi sunshine! Too bad no one to spend today with."
	"I feel like no one cares even though I know they do"
Coping Strategies	"Have you talked to anyone about it yet?"
	"Art is always the easiest way to distract me from my anxiety"
Psychiatric Comorbidity & State	"Really stressing today. Lots of built up anger"
	"I am anorexic and cut myself."
Non-psychiatric Comorbidity	"Could use a little bit of aid #DisabilityAid"
	"Lots of back pain ruining what should be a beautiful day."
Substance Use	"I really shouldn't be drunk this early."
	"Weed makes the dreams go away and thats a good thing."
Support & Advocacy	"If I can manage a smile, I believe you can too one day!"
	"RIP Chester. If you're going through pain, reach out to me."
Personality and Identity	"Girls say they love a man in uniform until they do their job"
	"Lol grandma still think I'm bringing a boy home"
Music Culture & Lyrics	"#FallingInReverse :D"
	"Scene doesn't mean emo idiots. I dont want to kill myself."
Familial/Romantic Relationships	"when bae dont answer the phone xx"
	"Mom: You'll never lose weight. Me: Is that why dad left?"
Political & Moral Beliefs	"look in the mirror if you're not upset a cop can murder"
	"Trump will kill us all"
Hobbies	"Missin the old days when eveyone played Pokemon yellow"
	"Boys that watch the Kardashians. Love."
Non-personal Accounts	"My life was about to fall apart until I found the Calm app..."
	"Breaking News: 5-alarm fire just outside Tulsa..."

Table 6.3: Example tweets and phrases (modified to preserve anonymity) for each of the 25 evidence categories that were used to annotate whether an individual's tweets indicated the presence of depression.

case in the classification experiments above, each training sample contains 60% of the available data, with the learned classifiers only being applied to the remaining 40% of individuals at each iteration. We refrain from filtering mental health related tweets and those containing explicit diagnosis disclosures, as the goal in this experiment is not to quantify predictive ability, but rather to identify evidence of depression over time. We control for distributional shift over time by estimating influence using a model trained during the time period in which a tweet was posted.

6.7.2 Data

A total of 300 individuals (574 total instances) were selected randomly for annotation. Three individuals (myself A_1, and two non-authors B_1, B_2) independently generated the annotations used to facilitate our analysis. Statistics presented in the analysis are computed using the author's annotations, while reliability measures were computed using annotations from the non-authors. All annotators have several years of experience modeling language within social media to assess mental health, but do not claim to be experts in clinical psychology. Additionally, all annotators have prior experience with the CLPsych 2015 Shared Task data (Coppersmith et al., 2015d). We include the distribution of instances reviewed by each of our annotators in Table 6.4.

	Time Period			
	2012-2015	2015-2018	2018-2021	**Total**
A_1	298	157	119	574
B_1	103	62	40	205
B_2	26	15	12	53

Table 6.4: The distribution of instances coded by each annotator (A_1, B_1, and B_2) across the three time periods. The set of instances annotated follows the relationship: $B_2 \subseteq B_1 \subseteq A_1$.

6.7.2.1 Annotator Reliability

As a first look into inter-rater reliability, we consider three dimensions of agreement – evidence of depression (four-class and three-class)[4] and remission status (four-class). We present pairwise annotator agreement matrices for each of these dimensions in Figure 6.1. We use Cohen's kappa κ to evaluate pairwise annotator agreement (Cohen, 1960) and Krippendorff's alpha α to evaluate multi-annotator agreement (Krippendorff, 2011).

We observe fair to moderate agreement for the evidence-of-depression task: $\alpha = 0.4376$ and $\alpha = 0.4988$ for the four-class and three-class versions, respectively. Meanwhile, agreement on remission status is poor, reflected by a Krippendorff's α of 0.3561. In isolation, these agreement measures would suggest the results of our analysis should be accepted tentatively at best (Krippendorff, 2004). However, we

[4] Note that the three-class evidence-of-depression grouping simply merges the Some Evidence and Strong Evidence categories of the four-class version.

argue these statistics are perhaps a bit conservative and skewed by the small sample size of annotations generated by B_2. A review of the underlying distributions provides us an opportunity to understand axes of disagreement and, in turn, contextualize the results presented in §6.7.

As shown in Figure 6.1, annotator B_2 exhibits a higher propensity to use the "Uncertain" label in the evidence-of-depression tasks compared to annotators A_1 and B_1. At the same time, while annotator B_2 is more inclined to indicate they are uncertain about an example than annotator A_1, we note that annotator B_1 appears to have a higher baseline threshold of what constitutes evidence of depression than annotator A_1. The latter is demonstrated by the fact that nearly all examples marked in the affirmative by B_1 were also marked as such by A_1, but a large number of examples marked in the affirmative by A_1 were marked as not containing evidence of depression by B_1.

With respect to the remission status task (bottom subplot of Figure 6.1), we note that annotator B_1 is more likely to mark an example as uncertain and more likely to mark an example as being in-remission than annotators A_1 and B_2. Broadly, this distribution highlights the difficulty of distinguishing active cases of clinical depression from prior experiences and lingering effects. It also serves as support for our recommendation in §6.8 that researchers should attempt to include the time a

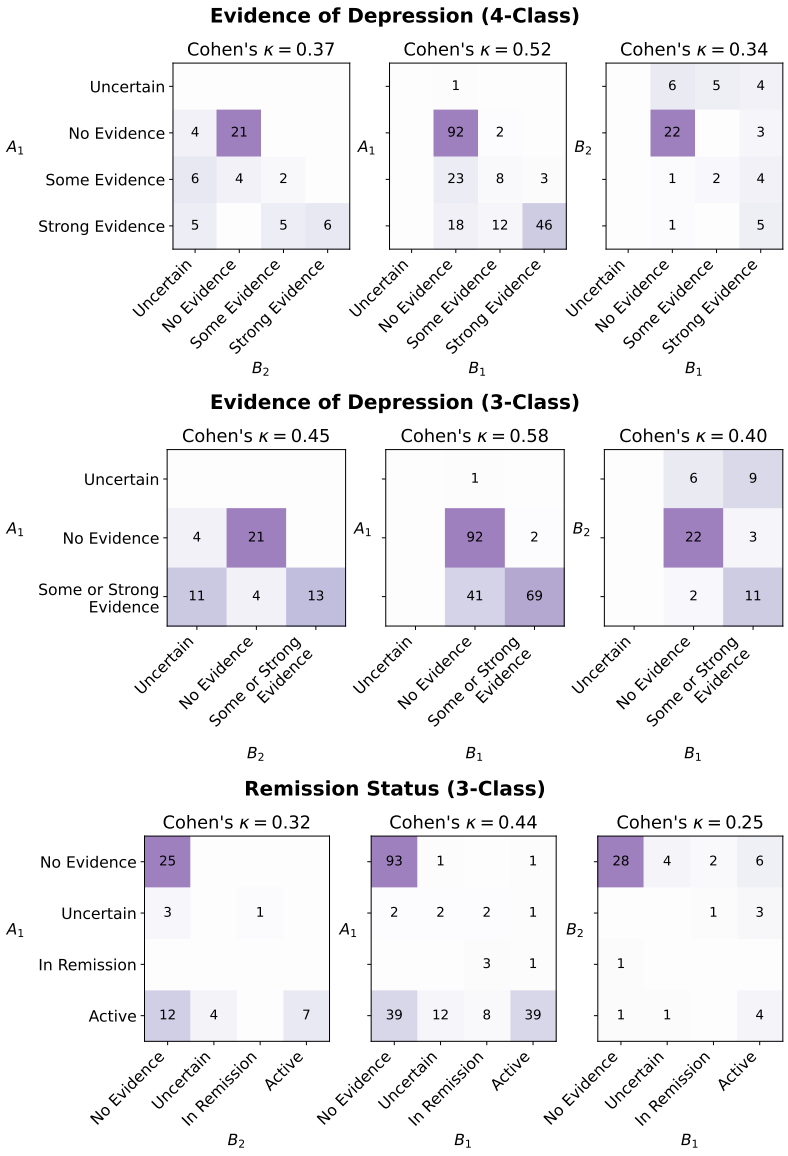


Figure 6.1: Pairwise annotator agreement matrices for the annotation tasks.

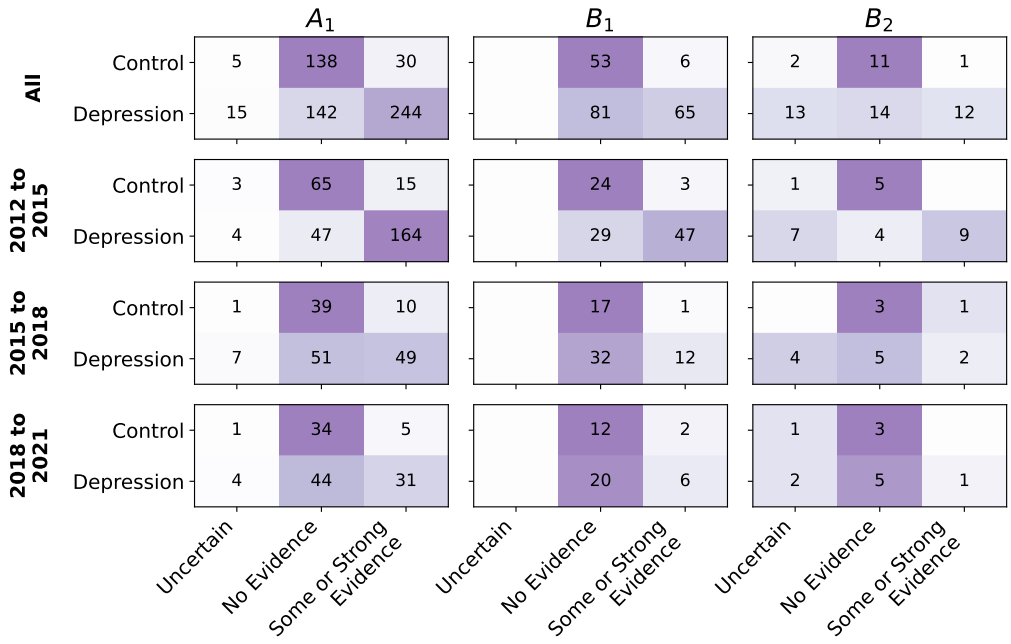


Figure 6.2: The distribution of annotations for the evidence of depression task (three-class) as a function of the original CLPsych labels. Affirmative evidence of depression becomes less prevalent in the new time periods compared to the original time period for each annotator.

diagnosis was received by an individual when curating new datasets.

We acquire additional context for our results by examining the distribution of annotations as a function of the original CLPsych labels. Examining the results visualized in Figure 6.2, we first note that annotator A_1 classifies instances most accurately (under the assumption that ground truth is fixed over time). We believe this outcome to be a result of exposure bias; the annotation task was conducted *after* the completion of several modeling experiments, through which annotator A_1 was

uniquely provided an opportunity to learn more about the presentation of depression by individuals in the 2015 CLPsych Shared Task dataset. We also note the distribution of "Uncertain" decisions from annotator B_2 concentrating within the original depression group. This seems to suggest annotator B_2 adopted a conservative coding approach when presented with instances that contained smaller degrees of evidence, whereas annotators A_1 and B_1 required a lower threshold of evidence to make a decision.

To conclude our reliability analysis, we examine agreement regarding the manner in which each annotator made their decision (i.e., evidence identification). We find that annotators A_1 and B_1 generally identify diagnosis disclosures within the same instances. Annotator B_2 often abstained from making a decision when presented with a disclosure due to uncertainty regarding the subject of the diagnosis. Annotator A_1 also indicated the presence of a depressed and/or irritable mood at a significantly higher rate than the other annotators, seemingly more sensitive to extreme negative emotions than the other annotators.

Discussion. Considering the difficulty of the annotation task, it is perhaps not surprising to have observed less than perfect annotator agreement. Machine learning classifiers often require hundreds of posts to make an accurate estimate of an individual's mental health status, while our annotators were only provided at maximum of 30 posts and encouraged to rely on varying levels of prior knowledge

	Dates	Total	Some Evidence	Strong Evidence	Not Active
Con.	2012-2015	83	15	3	1
	2015-2018	50	10	2	0
	2018-2021	40	5	0	0
Dep.	2012-2015	215	164	136	10
	2015-2018	107	49	28	2
	2018-2021	79	31	16	1

Table 6.5: Breakdown of evidence labels as a function of time period and labels from the original CLPsych dataset. Clinically aligned evidence of a depression diagnosis becomes less prevalent over time.

regarding the presentation of depression in social media. Critically, we emphasize that the goal of the analysis presented in §6.7 is *not* to curate ground truth labels of mental health status or act as clinical experts, but rather to understand biases that may exist in a depression dataset generated using self-disclosed diagnoses. The analysis of inter-rater reliability presented above provides an opportunity to further ground the results discussed in §6.7 and highlight areas that may benefit from future research.

6.7.3 Results

6.7.3.1 Validity Over Time

What proportion of labels in the updated sample remain relevant?

In line with underlying clinical knowledge regarding the dynamic nature of

depression, we observe a significant decrease in linguistic evidence of depression over the course of time. Roughly 76% of individuals in the original depression group displayed at least some clear evidence of a depression diagnosis during the first time period (2012-2015), in comparison to 45% and 39% of individuals in the 2015-2018 and 2018-2021 time periods, respectively. Across all time periods, only a small number of affirmative instances of depression appear to be in remission. That said, the non-zero level of inactive depression annotations in the original time period highlights an important consideration for practitioners who would like to leverage disclosure-based mechanisms to annotate mental health data moving forward.

The presence of evidence for a depression diagnosis in a subset of the original control group is quite striking. Other studies have raised questions regarding the possible risk of introducing such label noise when curating a control group using a random sampling protocol (Wolohan et al., 2018b), though none have provided tangible evidence of this contamination to the best of our knowledge. We see that approximately 4% of individuals in the control group display strong evidence of a depression diagnosis within the original time period. Although relatively small, it is an important reminder of the pitfalls of random control group sampling for health-related social media modeling tasks.

Discussion. The decrease in evidence of a depression diagnosis over time lends

support to the introduction of label noise in the updated dataset. Furthermore, it would explain the decrease in predictive performance observed in our previous classification experiments. However, the proportional drop in evidence of a depression diagnosis over time appears too large given the relatively minor reduction in classification accuracy.

We identify two possible explanations for this inconsistency. First, we recognize the possibility that our annotation procedure is insufficient to provide an annotator with appropriate information and comprehensive criteria for indicating evidence of a depression diagnosis. Only a small subset of an individual's entire post history is displayed to the annotator, a subset chosen using an inherently error-prone statistical ranking method. It is possible that stronger indicators of a depression diagnosis lie outside the 30-tweet sample size window for some individuals. Moreover, the annotator was instructed to rely predominantly on DSM-5 criteria to inform their decision, though several prior computational studies have shown language informative of depression may stray from explicit diagnostic criteria and be difficult for humans to recognize altogether (e.g., increased personal pronoun usage (Holtzman et al., 2017)).

More concerning is the possible presence of non-trivial confounds introduced by the original dataset's sampling/annotation procedure which may artificially inflate predictive performance estimates. Similar types of bias have been identified in prior

work when attempting to transfer statistical mental health models trained using proxy-based annotations to new populations of individuals (e.g., demographics, patient populations) (Ernala et al., 2019; Aguirre et al., 2021). Although sampling-based artifacts may be causally-related to the original diagnosis disclosure (e.g., a coping mechanism that becomes a hobby, heightened levels of neuroticism), they may be serve as a red herring in place of primary indicators of depression.

6.7.3.2 Selection Bias

Do presentations of depression provide evidence of sampling-related confounds?

We present a summary of evidence used by annotator A_1 to justify their decision to mark individuals as having "Some Evidence of Depression" across time periods in 6.3. Personality-related attributes are prominent features in all periods of the updated dataset. For example, indications of a depressed and/or irritable mood were the most common form of evidence in support of an individual having a depression diagnosis. In many cases, anger and irritation were displayed in the form of interpersonal confrontation (passively and actively) with other Twitter profiles. Negative emotions such as loneliness, fear, and existential dread were also displayed readily amongst those showing signs of a depression diagnosis. This result aligns with knowledge

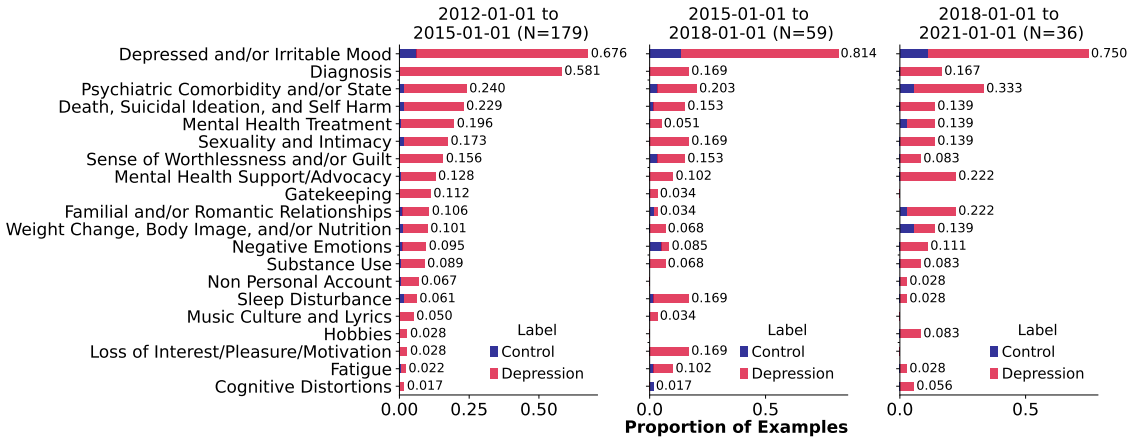


Figure 6.3: Distribution of evidence amongst individuals indicated as displaying at least some evidence of a depression diagnosis. A depressed and/or irritable mood is consistently the most common type of evidence within each of the three time periods.

regarding the relationship between personality and depression, with elevated levels of neuroticism (negative affectivity and vulnerability to stress) being common in those living with depression (Bagby et al., 2008; Lahey, 2009; Bondy et al., 2021). Although etiologically relevant, this heightened level of emotional affect emerges as one possible artifact which may confound displays of depression and serve as a nuisance variable in linguistic models of the condition (Tackman et al., 2019).

We also found it common for individuals to mention comorbid psychiatric conditions – such as obsessive compulsive disorder, bipolar disorder, and general anxiety. Many of these conditions share similar underlying symptoms and causes with depressive disorders (Franklin and Zimmerman, 2001; Goodwin, 2015), but tend to assume a different temporal profile (Schoevers et al., 2005). The significant overlap

often makes it difficult for trained physicians to properly diagnose individuals (Bowden, 2001) and for language-based algorithms to achieve appropriate discriminative sensitivity (Ive et al., 2018). We recognize the possibility that these comorbid conditions are active during the updated time periods for some individuals and may assume a proxy role in place of depression.

Although not captured by any single evidence category in isolation, there emerged a distinct propensity for "oversharing" amongst individuals from the original dataset's depression group. More specifically, we identified ample discussion of topics that are typically considered socially inappropriate in public discourse spaces (e.g., sexual activity, familial conflict, use of controlled substances). On one hand, this is an interesting finding given that individuals living depression often demonstrate lower levels of emotional self-disclosure (Wei et al., 2005; Kahn and Garrison, 2009). On the other hand, we note that prior work in clinical psychology has recognized a similar propensity for depressed and anxious individuals to engage in oversharing within social media (Radovic et al., 2017; Law et al., 2020).

The theory behind the latter is that social media offers an opportunity to discuss the oft stigmatized challenges of mental health (Betton et al., 2015) and increase feelings of connectedness in a less personal environment (Luo and Hancock, 2020). With this in mind, perhaps it is not surprising that those who have openly disclosed

their experience with depression also feel comfortable discussing the aforementioned "taboo" topics. Nonetheless, this personal comfort remains relatively unique amongst the larger social media population. The unfortunate effect of this nuance is that it transforms the primary depression inference task into, essentially, a topic-classification task.

Discussion. Our analysis affirms what other recent studies on proxy-based mental health annotations have claimed – individuals who disclose a mental health condition systematically differ from the larger population of individuals living with that condition (Ernala et al., 2019; Saha et al., 2021). As a research community, we must be careful to disambiguate 1) training a language classifier to identify individuals who live with a mental health condition, and 2) training a language classifier to identify individuals who live with a mental health condition *and* disclose their diagnosis. Inappropriately equating the two creates an opportunity to erroneously estimate population-level dynamics (Amir et al., 2019) and ignore underrepresented voices from communities who tend to possess conservative ideologies regarding mental health (Loveys et al., 2018; Aguirre et al., 2021).

6.8 Recommendations

Demand for computational methods to quantify mental health dynamics within social media data is at an all time high (Galea et al., 2020). However, the potential impact of these methods remains bounded by the robustness of datasets used for their development. Spanning nearly a decade of online activity, our study uniquely identifies evidence of these limitations as they currently manifest in non-clinically derived mental health social media datasets. This evidence leads us to offer three recommendations for enhancing data curation and model evaluation.

Annotate Diagnosis Date & Comorbidities. We identified several instances within our dataset where a diagnosis disclosure was made in reference to a condition that had since entered remission. In other cases, depression diagnoses were either supplanted by or augmented with alternative psychiatric diagnoses. Indicators regarding the time a diagnosis was made, many of which can be identified using inexpensive algorithms (MacAvaney et al., 2018), can provide important signal regarding the temporal relevance of a psychiatric diagnosis. Meanwhile, inclusion of comorbidities may provide researchers an opportunity to model psychiatric heterogeneity (Arseniev-Koehler et al., 2018) and interpret longitudinal generalization.

Sample Control Groups using Propensity Matching. Control group selection is influential in both training and evaluation of statistical models of

mental health (Pirina and Çöltekin, 2018). Prior work has leveraged a myriad of criteria to match individuals who have disclosed a psychiatric diagnosis with suitable counterparts – demographics (Coppersmith et al., 2014a), online behavior (Cohan et al., 2018), and language (De Choudhury et al., 2016). Though use of inconsistent matching criteria is less than ideal, the absence of any protocol is potentially more problematic (Shen et al., 2018; Wolohan et al., 2018b). We recommend practitioners leverage propensity-based matching (Imbens and Rubin, 2015) to reduce the effect of self-disclosure biases (e.g., personality, interests, demographics). In addition to the aforementioned dimensions, researchers may augment their criteria using classifiers to infer relevant latent attributes (Preoţiuc-Pietro et al., 2015) or neural models to derive user-level embeddings (Amir et al., 2017).

Identify and Filter Sampling Biases. Our analysis benefited from context that emerged when attempting to train classifiers that generalize over long time periods. However, access to supplementary data is not necessary to understand whether artifacts may exist in a dataset. Algorithmic approaches, such as those from Le Bras et al. (2020), may be used to identify instances containing spurious correlations. These approaches should be used to augment insights derived from manual annotation and review. We found our counterfactual explanation technique for ranking the influence of individual posts on user-level predictions began yielding

insights after only a few dozen examples, though alternative ranking methodologies are available (Ge et al., 2021; Uban et al., 2021). Outcomes should be used to inform preprocessing decisions, construct fair evaluations (Poliak et al., 2018), and inform the description of a dataset within documentation/datasheets (Gebru et al., 2021).

6.9 Limitations

Though our analysis identified data attributes that may inhibit statistical generalization, we also found evidence in support of the validity of self-disclosed diagnoses for annotating mental health status. The majority of individuals within the CLPsych dataset's original time window showed clear evidence of depression that aligns with clinical criteria. Many of these indicators remained stable over the course of time. Moreoever, the 2015 CLPsych Shared Task dataset is just one of many resources in this research community, all of which are likely to exhibit varying degrees of noise depending on their respective sampling protocols. Conclusive statements regarding the validity of self-disclosed diagnoses require evidence from multiple social media platforms, cultural groups, and time periods.

6.10 Ethical Considerations

Ethical challenges emerging from use of public social media data to analyze an individual's mental health have been examined extensively by members of both computational and clinical/public health communities (Conway and O'Connor, 2016; Chancellor et al., 2019). Privacy-related concerns are the most poignant for our study, which relies both on de-anonymizing records from a vulnerable population and manually reviewing/analyzing individual posts.

Indeed, many individuals who publicly discuss their mental health or disclose a psychiatric condition within social media admit that they worry about harmful repercussions of sharing such sensitive information with the public (Ford et al., 2019; Naslund and Aschbrenner, 2019). Primary fears include risking occupational stability, damaging interpersonal relationships, and being subjected to hostile communications. Whether potential positive outcomes (e.g., development of systems for recommending mental health care, fiduciary aid to address population-level crises) offset these threats remains largely dependent on an individual's personal life experience. For example, psychiatric patients have expressed stronger approval toward analysis of their social media than members of the general public (Mikal et al., 2017). The same holds true amongst younger individuals (Naslund and Aschbrenner, 2019).

Recognizing these viewpoints, we are careful to mitigate privacy-related risks

to the greatest extent possible given our primary research aim. For example, account identifiers distributed within the 2015 CLPsych Shared Task dataset are de-anonymized only temporarily to link updated records with existing post histories. We also redact account handles and URLs from the text analyzed during our manual coding procedure (§6.7). In line with protocols enumerated by Benton et al. (2017a), all data is stored on a remote server and secured using OS-level group permissions. We perform our analysis under the external guidance of clinical psychologists and psychiatrists. Our study is also reviewed by our Institutional Review Board (IRB), obtaining exempt status under 45 CFR §46.104.

Critically, our intention is not to develop a public-facing system for algorithmic analysis of mental health. Rather, our goal is to evaluate the validity of an existing and widely-adopted data curation practice (Chancellor and De Choudhury, 2020; Harrigian et al., 2021). Failure to comprehensively understand biases that arise under this methodology can have severe detrimental effects in downstream systems. In the case of estimating population-level health trends, for instance, we have already seen machine learning classifiers produce outcomes that are inconsistent across computational studies (Wolohan, 2020; Biester et al., 2021; Harrigian and Dredze, 2022a) and in conflict with traditional measurement techniques (Amir et al., 2019). Continuing to pursue this line of research without questioning the validity of its underlying data

has the potential to irreparably damage the public's trust in this domain, and worse, enable ill-informed decision making in highly-sensitive circumstances.

6.11 Discussion

The proliferation of shared tasks (Filannino and Uzuner, 2018) and open data repositories (e.g., Hugging Face's Dataset Hub) (Goben and Sandusky, 2020; Lhoest et al., 2021) has generally been beneficial to the computational community. For example, it has not only allowed practitioners to test and replicate methodological ideas with greater expediency, but also removed a significant barrier to data access that often affected researchers from different institutions disproportionately (Chapman et al., 2011). The latter is especially pertinent with respect to health equity given that differential access to data and computational infrastructure across institutions could exacerbate existing disparities by advantaging well-resourced organizations and leaving under-resourced organizations behind. The downside, however, of these mechanisms is that they discourage practitioners from performing deep, reflective analyses of datasets they did not curate themselves (Parra Escartín et al., 2017). If someone has already done the intensive work of analyzing and preprocessing a dataset, why would a practitioner potentially waste time replicating their efforts?

The main issue with such a mentality is that it is highly unlikely that the

shortcomings of a dataset identified by one practitioner will exactly match the shortcomings identified by a different practitioner. We all have different cognitive biases and experiences that shape the questions we ask about a dataset, the analyses we perform, and the conclusions at which we arrive. For example, one practitioner may perceive predictive signal as causally relevant that a different practitioner perceives as spurious or unstable (Weinberger and Bradley, 2020; Reiss, 2022). When we simply accept the conclusions of those who have analyzed a dataset before us, we risk narrowing our own world view and propagating harmful sample biases (Greenberg, 2009).

The health domain is particularly susceptible to such issues. The people who are most qualified to ask questions about a health-related dataset and interpret results (i.e., health professionals) are often *not* the people who are most qualified to perform the actual analysis (i.e., ML and NLP practitioners). Even when these two groups of people collaborate closely with one another, certain data quality issues are likely to be missed or lost in translation (Mao et al., 2019; Park et al., 2021b). The risks of improperly contextualizing a result are magnified when ML and NLP practitioners work in isolation.

Accordingly, there exists a need for approaches such as our own that facilitate data analysis in a computationally efficient and easy-to-understand manner. The

counterfactual explanation method we introduce in §6.7 for the purpose of measuring the influence of text segments within larger textual predictions requires only a few lines of code and has applicability in nearly any classification setting where constituent linguistic parts are present (e.g., sentences in a document, documents in a history, individuals in a community). While one could use alternative techniques such as attention mechanisms (Li et al., 2016; Jain and Wallace, 2019; Wiegreffe and Pinter, 2019) or influence functions (Han et al., 2020; Schioppa et al., 2022) to examine such relationships, it is arguably more reasonable to expect that a method like ours has a wider user base, potentially beyond that of ML and NLP practitioners.

6.12 Looking Ahead

This chapter and chapter 5 have highlighted several issues that arise when circumventing the challenges associated with collecting health data. The natural follow-up question is: what do we do to address these biases and mitigate issues related to distribution shift? In the next part of this thesis, we will present and analyze methods for accomplishing this feat.

Part III: Counteracting Distribution Shift in Health Language Data to Promote Robustness

Chapter 7

Addressing Semantic Shift in Longitudinal Monitoring of Social Media

7.1 Overview

Once we suspect that our training data distribution is not representative of our target data distribution, we are tasked with counteracting the effects of distribution shift. As discussed in Chapter 3, there exist a variety of methods for doing so (Ramponi and Plank, 2020; Hupkes et al., 2023). However, not all domain adaptation and generalization methods allow practitioners to simultaneously mitigate generalization loss *and* interpret the cause of the loss. The latter is not only useful for guiding a practitioner's approach to addressing distribution shift, but also for instilling trust in a model upon deployment (Katakkar et al., 2022; Madsen et al., 2022).

In this chapter based on Harrigian and Dredze (2022a), we adapt an existing method for measuring semantic shift between two language distributions to act instead as a robust and interpretable feature selection technique (Sun et al., 2019a; Fu et al., 2021). Across 4 datasets, we demonstrate that our original, unsupervised method improves generalization over time while also supporting interpretation of distribution shift. Further, in the context of measuring COVID-19's effect on the prevalence of depression, we show that semantic shift can introduce significant instability in real-world public health surveillance applications. Our study thus calls into question a myriad of contemporary studies that have attempted to measure changes in mental health in the wake of the COVID-19 pandemic, and offers a mechanism to facilitate

sensitivity analyses in future efforts.

7.2 Background

Across multiple disciplines, studies of social media text and metadata have yielded valuable insights into population-level dynamics (e.g., consumer habits (Saura et al., 2019), voting patterns (Beauchamp, 2017), health (Amir et al., 2019; Aiello et al., 2020; Ayers et al., 2024)). In several cases, the outcomes have enabled policy makers to more effectively anticipate and respond to concerns amongst their constituents (Myslín et al., 2013; Burnap and Williams, 2015). Researchers have naturally looked to build upon the utility of these past analyses to inform decision-making in the wake of emerging health crises – e.g., COVID-19 (Schaar et al., 2021), mental health epidemics (Bagroy et al., 2017; Saha and De Choudhury, 2017).

Methods based on machine learning (ML), natural language processing (NLP), and web mining make up the foundation of these efforts, offering an opportunity to answer questions that cannot be easily addressed using traditional mechanisms alone (Paul et al., 2016; Nobles et al., 2018a). While different studies may opt to leverage slightly different techniques (e.g., different model architectures, training data samples), they more or less follow the same formulaic approach. First, acquire ground truth for a target concept within a small sample of data (e.g., regular expressions to

identify medical diagnosis disclosures (Coppersmith et al., 2014a), follower networks indicating political leaning (Al Zamal et al., 2012)). Next, train a statistical classifier on this data with the objective of re-identifying language associated with the target concept. Finally, apply the trained classifier to a new population of individuals across multiple time steps (e.g., annually, weekly). The first two stages of this modeling procedure have been explored extensively (Volkova et al., 2015), but studies validating the final step have been sparse.

7.3 Motivation and Contribution

A lack of analyses of temporal robustness of the models used in these studies belies the seriousness of the problem: language shifts over time – especially on social media (Brigadir et al., 2015; Loureiro et al., 2022) – and statistical classifiers degrade in the presence of distributional changes (Daume III and Marcu, 2006; Huang and Paul, 2019). For example, new terminology could be used to convey existing concepts; existing terminology could be used to convey new concepts; or, the overall usage of terms could change while semantic relationships remain fixed. The latter is of particular concern when major social events cause large-scale shifts in the topic of online conversation (e.g., discussion of healthcare increases during a pandemic, discussion of a political leader increases near an election). Unfortunately, these are

often the types of events we seek to study.

To better understand the gap in the literature, we conduct a case study on estimating changes in depression prevalence during the COVID-19 pandemic, a analysis with value to both the medical and public health communities that has thus far has procured incongruous results across studies (Galea et al., 2020; Bray et al., 2021). We draw inspiration from research on detecting distributional shifts in language over time (Dredze et al., 2010; Huang and Paul, 2018), focusing our attention on a recently-introduced method that leverages word embedding neighborhoods to identify semantic shift between multiple domains (Gonen et al., 2020). We use the semantic stability scores generated from this method, in combination with more traditional feature importance measures, to improve generalization in the presence of semantic shift. More importantly, we provide evidence that semantic shift can introduce undesirable variance in downstream longitudinal monitoring applications, despite having an indistinguishable effect on historical predictive performance. Altogether, our study serves as a cautionary tale to practitioners interested in using social media data and statistical algorithms to derive sensitive population insights.

7.4 Challenges of Public Health Surveillance

When the COVID-19 pandemic began in March 2020, healthcare professionals warned of an impending mental health crisis, with economic uncertainty (Godinic et al., 2020), loss of access to care (Yao et al., 2020), and physical distancing (Galea et al., 2020) expected to reduce mental wellness. Given the inherent difficulties of measuring mental health at scale using traditional monitoring mechanisms, the healthcare community called upon computational scientists to leverage web data to provide evidence for optimizing crisis mitigation strategies (Torous et al., 2020). Computational researchers responded by analyzing search queries regarding anxiety and suicidal ideation (Ayers et al., 2020; 2021), developing novel topic models to gather an understanding of the population's concerns (Koh and Liew, 2020), and applying language-based classifiers to streams of social media text (Wolohan, 2020).

Unfortunately, these inquiries failed to provide unanimous insights that could be used with any confidence to manage the ongoing situation (Zelner et al., 2021). For instance, application of a neural language classifier to the general Reddit population estimated over a 50% increase in depression after the start of the pandemic (Wolohan, 2020), despite an analysis of topic distributions within three mental health support

subreddits finding evidence to suggest the opposite (Biester et al., 2020). Similarly, multiple keyword-based analyses using Google Trends data suggested anxiety increased relative to expected levels (Ayers et al., 2020; Stijelja and Mishara, 2020), while others suggested anxiety levels actually remained stable (Knipe et al., 2020).

While it is easy to criticize the computational community for failing to present a unified narrative regarding COVID-19, more traditional forms of public health surveillance were not necessarily more consistent. For example, in a survey conducted early in the pandemic by the Centers for Disease Control and Prevention (CDC), 10.7% of respondents reported having thoughts of suicide in the previous 30 days (Czeisler et al., 2020) (a 2× increase over the expected rate). Later in 2020, data suggested suicide rates remained stable or even fell after the start of the COVID-19 pandemic (Ahmad and Cisewski, 2021). Some argued this drop was the result of a "pulling together" effect (Ayers et al., 2021), an outcome that had been observed previously during times of crisis (Claassen et al., 2010; Gordon et al., 2011). However, upon closer inspection, it became clear that this trend was confounded by Simpson's paradox (Julious and Mullee, 1994). Reductions in suicide rate were observed amongst White folk, while a significant increase was observed amongst ethnic and racial minorities (McKnight-Eily et al., 2021), with whom stress-inducing factors such as financial instability and food insecurity are more common.

We share these anecdotes for two reasons. First and foremost, they speak to the inherent difficulty of public health surveillance, in particular during unprecedented health events. If this was a solved problem, we would not have seen (and would not still be seeing) a myriad of research articles claiming to have a method that explains what happened to mental health during the COVID-19 pandemic (Kupcova et al., 2023; Ueda et al., 2023; Wolf and Schmitz, 2024). Second, they highlight that what one observes within data on a superficial level is not necessarily what one observes when diving deeper into that data.

7.4.1 Motivating Example

To clarify what we mean by the latter, let us examine usage patterns for a handful of mental-health related n-grams on Reddit and Twitter (i.e., Figure 7.1).[1] On the surface, increases in the incidence of these n-grams seems to indicate periods in time during which the population's collective mental well-being became worse. However, when we look more closely at the posts in which these n-grams are used, we quickly realize they are actually not explicitly reflective of mental health in the manner we would have expected. For example, spikes in usage of the term "suicide" in August

[1] Pointwise Mutual Information (PMI) of each term within historical samples of depressed individuals was used to determine mental health relevance.

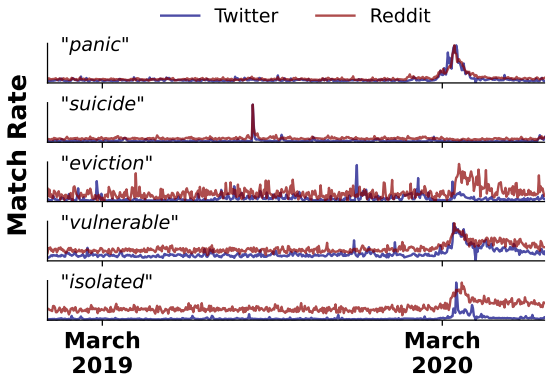


Figure 7.1: The proportion of posts on Twitter and Reddit containing a subset of depression-indicative n-grams over time.

2019 were actually a response to Jeffrey Epstein's death, while increased usage of "panic," "eviction," "vulnerable," and "isolated" in March 2020 primarily correspond to discussion of pandemic-specific circumstances (e.g., toilet paper *panic*, *eviction* moratorium, medically *vulnerable* populations, *isolated* for quarantine).

While studies that leveraged statistical language models (e.g., Biester et al. (2020) and Wolohan (2020)) may have been in a better position to disambiguate between these contexts than those that leveraged purely keyword-based approaches (e.g, Stijelja and Mishara (2020)), there is no guarantee that they were able to handle a shift as large as what was presented by COVID-19 (Dhingra et al., 2022). As such, we find ourselves asking two critical questions:

1. To what extent has semantic shift affected results regarding mental health in the published literature?

2. Is it possible to obtain more reliable longitudinal estimates by explicitly invoking knowledge of semantic shift when training statistical algorithms?

7.5 Measuring Semantic Shift at Scale

Unfortunately, the type of manual lexical analysis discussed in §7.4.1 is not feasible to perform at scale for statistical language models that often have vocabularies with tens of thousands of terms. Fortunately, a substantial pool of prior work has proposed methods for algorithmically quantifying semantic shift between language domains (Dredze et al., 2010; Kutuzov et al., 2018). We choose to leverage a method introduced recently by Gonen et al. (2020), which has not only outperformed several state-of-the-art alternatives in preliminary studies (Hamilton et al., 2016), but also shown promise for use by applied practitioners. Core advantages of this methodology include its interpretability, robustness to stochasticity, ease of implementation, and low computational overhead. We formally introduce the methodology here and will refer back to it throughout the study.

Gonen et al. (2020)'s method assumes that semantically stable language has similar sets of neighboring lexical units within word embedding spaces of different data samples. More formally, for two data samples $\mathcal{P}$ and $\mathcal{Q}$, the semantic stability

S of a lexical unit (e.g., word, n-gram) w can be measured as:

$$S(w; \mathcal{P}, \mathcal{Q}) = \frac{\text{nb}_{\mathcal{P}}^{(k)}(w) \cap \text{nb}_{\mathcal{Q}}^{(k)}(w)}{k},$$

where $\text{nb}_{\mathcal{X}}^{(k)}(w)$ (i.e., the neighborhood of w in $\mathcal{X}$) denotes the top-k set of lexical units nearest to lexical unit w in word-embedding vector space X based on a vector distance metric of the modeler's choosing. Hyperparameters include the neighborhood size k, the minimum frequency of n-grams used for building each neighborhood cf_{nb}, the minimum frequency of n-grams input to the semantic shift calculation cf_{shift}, the distance function used for measuring a word's neighborhood, and the embedding model architecture. For the purpose of measuring semantic shift longitudinally, we can think of independent, discrete time periods as the data samples $\mathcal{P}$ and $\mathcal{Q}$.

7.6 Data

To comprehensively understand how semantic shift may influence downstream longitudinal analyses, we leverage datasets which come from multiple social media platforms, span a wide range of time periods, and leverage different annotation/sampling mechanisms. As mentioned before (§7.4), we focus on the the task of estimating depression prevalence, an important undertaking within the public health community (Gelenberg, 2010) due to the substantial burden on individuals,

Dataset	Platform	Dates	# Users
CLPsych (Coppersmith et al., 2015d)	Twitter	2012 - 2014	Control: 477 Depression: 477
Multi-Task Learning (Benton et al., 2017b)	Twitter	2012 - 2016	Control: 1,400 Depression: 1,400
SMHD (Cohan et al., 2018)	Reddit	2013 - 2018	Control: 127,251 Depression: 14,139
Topic-Restricted Text (Wolohan et al., 2018b)	Reddit	2016 - 2020	Control: 107,274 Depression: 9,210
1% Stream (Wang et al., 2015)	Twitter	Jan. 2019 - July 2020	All: 25,379
Pushshift.io (Baumgartner et al., 2020)	Reddit	Jan. 2019 - July 2020	All: 40,671

Table 7.1: Summary statistics for labeled and unlabeled datasets. Labeled dataset statistics are further broken out as a function of control and depression groups.

communities, and society (Dressler, 1991; Lépine and Briley, 2011; Chang et al., 2012b). While we consider this specific use case, our analyses are generally applicable to longitudinal monitoring of social media.

7.6.1 Data Sources

7.6.1.1 Labeled Data

To numerically quantify the effect semantic shift has on predictive generalization, we consider four widely adopted datasets containing ground truth annotations of individual-level depression status. To diversify our data sample and understand platform-specific differences, we consider two Twitter datasets – 2015 CLPsych Shared Task (Coppersmith et al., 2015d), Multi-Task Learning (Benton et al., 2017b) – and

two Reddit datasets – Topic-restricted Text (Wolohan et al., 2018b), Self-reported Mental Health Diagnoses (SMHD) (Cohan et al., 2018). Each dataset relies on a form of distant supervision; the Topic-restricted Text dataset assumes original posts made in the r/depression subreddit serve as a proxy for a depression diagnosis, while the remaining three datasets use regular expressions to identify self-disclosures of a depression diagnosis. This annotation procedure remains widely used to train classifiers for monitoring population-level trends due to challenges inherent in acquiring sufficient samples of annotated data (De Choudhury et al., 2013a; Chancellor and De Choudhury, 2020), but remains prone to introducing label noise and other sampling artifacts (Ernala et al., 2019). Additional detail about the manner in which these datasets were constructed is provided in §5.5.

7.6.1.2 Unlabeled Data

Our primary interest is understanding the practical effects of semantic shift in longitudinal monitoring applications. Accordingly, we also collect large samples of text data from both Twitter and Reddit to use for extrinsic model evaluation. Our sampling procedures are inspired by those used in prior COVID-19 related work (Saha et al., 2020; Wolohan, 2020).

Twitter. We acquire raw Twitter data from the platform's streaming API, a

random 1% sample of all public tweets available for non-commercial research use. We isolate all original tweets (i.e., no retweets) that include an 'en' language metadata attribute and are further classified as being written in English based on automatic language identification (Lui and Baldwin, 2012). To facilitate application of our statistical classifiers, which require multiple documents from each individual to make accurate inferences, we further isolate individuals with at least 400 posts across the entire study time period (January 1, 2019 through July 1, 2020).

Reddit. We sample Reddit data within the same time period using the Pushshift.io archive (Baumgartner et al., 2020), which, unlike the Twitter streaming API, provides access to nearly all historical Reddit data (Gaffney and Matias, 2018). We begin data collection by identifying all users who posted a comment in one of the 50 most popular subreddits[2] between May 25, 2020 and June 1, 2020. Of the 1.2 million unique users identified by this query, roughly 200k were identified to have posted at least once per week during January 2019 and to not exhibit clear indicators of bot activity (e.g., repeated comments, username indicators, abnormal activity volume). We collect the entire public comment history from January 1, 2019 through July 1, 2020 for a random sample of 50k users in this cohort and perform additional filtering to isolate English data and users who have at least 200 posts across the study

[2] Popularity is based on the total number of subscribers as of 6/1/2020. Statistics were sourced from https://subredditstats.com

time period. Summary statistics for all datasets are provided in Table 7.1.

7.6.2 Preprocessing

To promote consistent analysis, we use automatic language identification (Lui and Baldwin, 2012) to isolate English text in each dataset (labeled and unlabeled). Additionally, per the recommendations of De Choudhury and De (2014), we exclude posts in the labeled datasets that either contain a match to a mental health related n-gram or are drawn from a subreddit explicitly dedicated to providing mental health support. This filtering is designed to encourage statistical models to learn robust linguistic relationships with depression as opposed to those introduced by sampling-related artifacts.

7.7 Improving Generalization

Our ultimate goal is to understand how the presence of semantic shift affects downstream outcomes obtained from longitudinal analyses of social media data. However, critical to the success of this goal is a methodology for controlling a statistical classifier's access to semantically unstable features when making inferences on unseen data. In this initial experiment, we demonstrate that Gonen et al. (2020)'s

method for measuring semantic shift can be adapted with minimal effort to curate vocabularies with constrained levels of semantic stability. Further, we demonstrate that these vocabularies often improve predictive generalization and outperform alternative feature selection methods despite lacking an explicit awareness for the target classification outcome.

7.7.1 Methods

We design our experiment with the intention of replicating a standard deployment paradigm seen within longitudinal analyses. Broadly, language classifiers are fit on historical accumulations of annotated data and evaluated iteratively within future one-year-long time windows (see Table 7.2). The influence of semantic shift on generalization is measured by comparing predictive performance (F1 score) of classifiers trained using a subset of semantically-stable terms to performance of classifiers trained using alternative feature selection methods which lack awareness of semantic shift altogether. Specific details are enumerated as a courtesy to the reader below before diving into the results.

Sampling. Each instance of an experimental run consists of multiple stages. In the first stage, users are randomly allocated into a 80/20 train/test split, with data from those in the training group used to learn word embeddings for each time

period within a dataset (i.e., historical accumulations and future one-year windows). During the second stage, users in both the training and test splits are independently resampled to form subsets for training and evaluating the depression classifiers. This secondary sampling step is constrained such that all users meet a minimum post threshold within each time period,[3] the number of users within each training time period is equivalent, and that classes are balanced in both training and test subsets. The first and second stages are repeated 100 and 10 times, respectively, providing us with 1,000 experimental samples for each dataset. Classifiers are evaluated using data from users in the 20% test split sampled during the first stage of the experiment.

Model Setup. Vocabularies of a maximum 500k n-grams (unigrams, bigrams, and trigrams only) are learned using Gensim's Phraser module, parameterized using a PMI threshold of 10 and minimum frequency of 5 (Mikolov et al., 2013b). For classification experiments, all posts sampled for a user from a given time period are tokenized using a Python implementation of Twokenizer (O'Connor et al., 2010), rephrased using the time period's specific Phraser model, concatenated together, and translated into a single document-term vector. Representations are transformed using TF-IDF weights learned at training time before being ℓ_2-normalized. As a classification architecture, we use ℓ_2-regularized logistic regression, optimizing parameters using limited-memory

[3] 200 for Twitter, 100 for Reddit; Reddit comments contain 2x as many tokens on average.

BFGS as implemented in Scikit-learn (Pedregosa et al., 2011). Inverse regularization strength C is selected independently for each training sample such that F1 score is maximized within the development splits of a 10-fold cross validation run.

Measuring Semantic Shift. To measure semantic shift, the same vocabularies and Phrase detection models introduced above serve as the foundation for training unique Word2Vec models (Mikolov et al., 2013b) for each dataset and time period. We leverage Gensim's implementation of the continuous bag of words (CBOW) formulation of Word2Vec to learn 100-dimensional embeddings, training each model for 20 iterations using the default window and negative sampling sizes (Mikolov et al., 2013a). We obtain semantic neighborhoods for each n-gram w using $k = 500$, $\text{cf}_{\text{nb}} = 50$, and $\text{cf}_{\text{shift}} = 50$. Alternative neighborhood sizes ($k = 250, 1000$) and frequency thresholds ($\text{cf}_{\text{nb}} = 10, 25, 100$; $\text{cf}_{\text{shift}} = 25, 100$) did not have a significant effect on downstream outcomes. In line with Gonen et al. (2020), we measure vector similarity using cosine distance.

Feature Selection. Semantic stability scores S are computed for each source (training) and target (evaluation) time period combination using Gonen et al. (2020)'s method. We vary vocabulary size in linear, 10-percentile intervals until all available tokens are used for training the language classifier. All vocabulary selection methods are enumerated below, chosen to encompass a variety of common strategies (naïve

and statistical) for reducing dimensionality and enhancing model performance.

Cumulative: Frequency > 50 in the source time period

Intersection: Frequency > 50 in the source & target time periods

Frequency: Top $p\%$ of n-grams with highest frequency

Random: Randomly selected terms, $p\%$ of the total available vocabulary

Chi-Squared: Top $p\%$ of n-grams with highest chi-squared test statistic (Pedregosa et al., 2011)

Coefficient: Top $p\%$ of n-grams with highest absolute logistic regression weight within the training data

Overlap: Top $p\%$ of n-grams with the highest semantic stability score S

Weighted (Overlap): Top $p\%$ of n-grams with the highest 50/50 weighted combination of Coefficient and Overlap scores

All feature selection methods below Frequency (inclusive) are a subset of the Intersection method. We introduce Weighted (Overlap) approach to balance the predictive value of a given feature and its semantic stability, theorizing that a vocabulary based solely on semantic stability may come at the cost of significant

				Naïve			Statistical		Semantic	
	Dataset	**Train**	**Test**	**Cumulative**	**Intersection**	**Frequency**	**Chi-Squared**	**Coefficient**	**Overlap**	**Weighted**
Twitter	CLPysch	2012-2013	2013-2014	65.6	67.7	67.6	68.7	67.7	**71.5***	69.6
	Multi-Task Learning	2012-2013	2013-2014	74.6	75.9	75.9	76.1	75.7	**77.9***	77.2
			2014-2015	70.3	76.0	76.0	76.2	75.8	**77.8***	76.5
			2015-2016	69.9	77.5	77.3	77.7	77.2	**78.3***	77.2
		2012-2014	2014-2015	77.8	77.9	77.8	78.1	78.3	78.1	**78.6**
			2015-2016	78.8	78.7	78.7	78.9	**79.2**	78.9	79.1
		2012-2015	2015-2016	79.9	80.0	80.0	80.0	**80.6**	80.2	**80.6**
Reddit	Topic Restricted Text	2016-2017	2017-2018	65.9	**66.2**	66.1	66.0	66.0	66.1	66.1
			2018-2019	**67.0**	66.9	66.8	66.8	66.8	66.6	66.8
			2019-2020	**67.0***	66.5	66.3	66.5	66.7	66.6	66.7
		2016-2018	2018-2019	66.7	67.2	67.1	67.1	66.9	**67.4**	66.9
			2019-2020	67.2	67.4	67.4	67.4	67.4	**67.5**	67.4
		2016-2019	2019-2020	66.7	66.8	66.9	66.8	66.8	**67.4***	67.0
	SMHD	2013-2014	2014-2015	79.9	79.8	**80.3**	79.9	79.9	79.9	79.9
			2015-2016	80.1	80.0	80.0	**80.5**	80.1	80.2	80.2
			2016-2017	79.2	79.2	79.3	79.8	79.7	79.2	**79.9**
			2017-2018	79.9	80.0	80.0	80.3	80.4	80.4	**80.8**
		2013-2015	2015-2016	79.7	79.5	79.8	79.9	79.8	**80.1**	79.9
			2016-2017	78.6	78.5	78.7	79.0	79.0	78.8	**79.1**
			2017-2018	79.6	79.6	80.2	79.9	80.4	80.4	**80.7**
		2013-2016	2016-2017	79.0	79.0	79.1	79.2	79.3	79.2	**79.4**
			2017-2018	79.8	79.6	80.4	79.8	80.4	80.6	**80.8**
		2013-2017	2017-2018	79.9	79.7	80.4	80.0	80.3	80.8	**81.0**

Table 7.2: Mean F1 score for the best performing vocabulary size of each feature selection method (oracle setting). Bolded values indicate top performers within each test set, while asterisks (*) indicate significant improvement over alternative classes of feature selection (i.e., Naive vs. Statistical vs. Semantic). Semantically-informed vocabulary selection matches or outperforms alternatives in nearly all instances, despite lacking knowledge of target outcome.

predictive power, while a vocabulary solely based on within-domain predictive power will be vulnerable to generalization issues.

7.7.2 Results

7.7.2.1 Quality of the Semantic Shift Measure

To validate our implementation of Gonen et al. (2020)'s method and build context for our classification results, we first manually inspect a sample of learned semantic stability scores for each dataset. On a distribution level, we see that stability scores tend to decrease as the gap between training and evaluation time periods increases – evidence of increased semantic shift over time. Additionally, we note that semantic stability scores within the Twitter datasets are generally lower than scores within than the Reddit datasets. These platform-specific differences align with our prior understanding of each platform's design, with Twitter tending to foster conversations motivated by current events (i.e., personal and global conflict) and Reddit offering individuals an opportunity to connect through shared interests that evolve over longer time periods (Noble et al., 2021).

For all datasets, common nouns and verbs make up the majority of terms with the highest semantic stability scores (e.g., eat, bring, give, city, room, pain). These types of tokens arise only infrequently within the lower tier of semantic stability scores, typically a result of isolated conflation with current events/pop culture – names of video games (e.g., blackout, warzone), television characters and celebrities (e.g.,

sandy, gore, rose), and athletic organizations (e.g., twins, braves, cal). Hashtags are frequently found in the lower semantic stability tier for the Twitter datasets, a reflection of the diversity of conversations in which they are used. Broadly, most of the observed semantic shift can be described as changes in the popularity of different word senses (Haase et al., 2021). Although this suggests that contextual language models (Devlin et al., 2019) would be well-suited for mitigating the effect of semantic shift in longitudinal analyses, emerging research suggests this is not necessarily true in the absence of additional tuning (Dhingra et al., 2022; Loureiro et al., 2022).

7.7.2.2 Effect on Generalization

We find that classifiers trained using vocabularies derived with a knowledge of semantic stability achieve equal or better predictive performance than alternative feature selection techniques in the majority of classification settings (Table 7.2).[4] Semantic stability tends to be more useful for generalization within the Twitter datasets than the Reddit datasets, likely due to the aforementioned platform-specific nuances. In all cases, joint use of semantic stability and coefficient weights to derive feature selection scores (i.e., Weighted (Overlap)) matches or moderately improves performance over use of coefficient weights in isolation.

[4] We exclude the Random method to save space, but note that it performed significantly worse across all settings as expected.

Finally, we note that the semantically-informed vocabulary selection methods not only offer reasonably wide operating windows (usually 20 to 50% of the total vocabulary size), but also tend to correlate with performance within source time periods. This latter detail suggests that semantically-stable vocabulary selection can be adequately performed in the absence of validation samples from a target time period, a necessity for most longitudinal analyses. We leave hyperparameter optimization for this methodology as an area for future exploration.

7.8 Practical Effects of Semantic Shift

Having demonstrated that semantically-aware vocabulary selection methods achieve comparable performance to alternative techniques using a fraction of features and can even improve predictive generalization outright, we turn our attention to understanding the practical effects semantic shift has in longitudinal modeling applications. Specifically, we leverage our ability to systematically constrain a language classifier's access to semantically volatile terms to evaluate how estimates of depression prevalence vary in the presence of semantic shift. Ultimately, we find that small changes in the vocabulary of a language classifier can promote large deviations in downstream outcomes, despite offering little to no indication of concern within historical data samples.

7.8.1 Methods

We leverage a similar experimental design to that from §7.7, making small methodological changes to acutely focus on understanding the practical effect of semantic shift in a deployment scenario. The full list of modifications is enumerated below; we release our code to support future research and allow others to reproduce our analysis.[5]

Measurement. To control for seasonal effects, we focus on estimating the year-over-year change in the prevalence of depression-indicative language amongst individuals in each of the unlabeled social media samples. Each unlabeled data sample is split into two distinct time periods – March 1, 2019 to July 1, 2019 (Pre-Pandemic) and March 1, 2020 to July 1, 2020 (During-Pandemic). Classifiers are applied to each individual in the unlabeled temporal samples who meets a minimum post volume criteria – 200 for Twitter and 100 for Reddit. We compute the prevalence of depression as the proportion of users in the unlabeled sample who have a predicted probability of depression greater than 0.5. We then measure the difference in estimated prevalence between the two time periods as a function of the underlying model vocabulary.

Experimental Design. In the first stage of the experimental procedure, we fit 10 embedding models for each of the labeled datasets – using randomly sampled

[5] https://github.com/kharrigian/semantic-shift-websci-2022

subsets (80% size) of the complete dataset – and three embedding models for each of the unlabeled data samples – one using data from the entire Jan. 1, 2019 to July 1, 2020 time period, one using data from March 1, 2019 to July 1, 2019, and one using data from March 1, 2020 to July 1, 2020. The latter two unlabeled data models are used to qualitatively identify language which has undergone semantic shift since the start of the COVID-19 pandemic, while the former model is used in conjunction with the labeled dataset models to identify semantically stable vocabularies for training classifiers. To reduce computational expense, we randomly sample 20% of posts to train each of the unlabeled data embedding models, but otherwise maintain the same hyperparameters and training settings enumerated in §7.7. The second stage of the experiment proceeds as before, resampling amongst the users allocated to the training split of the annotated data to derive semantically stable vocabularies and train language classifiers. We perform the second stage 10 times, providing us with 100 classifiers for each labeled dataset.

Term	2019 Context	2020 Context
Panic	Emotion (i.e., Fear)	Panic Buying, Misinformation
Cuts	Physical	Economic
Isolated	Feeling Detached	Quarantine
Strain	Discomfort/Pressure	Virus
Vulnerable	Emotion	At-risk Populations
Doctors	Personal Experience	Frontline Workers, PPE

Table 7.3: Change in the most prevalent context from 2019 to 2020 for a handful of terms which historically over-indexed in usage amongst individuals living with depression.

7.8.2 Results

7.8.2.1 Qualitative Analysis

Semantic stability scores for each of the raw data samples (pre/during COVID-19) align with intuition regarding sociopolitical events of the era. Many of the n-grams with the lowest semantic stability are related to the pandemic: "viral," "masks," "transmission," "isolation," "zoom," "lockdown." Of terms that over-index in historical usage amongst individuals with depression, the least semantically stable include: "panic," "cuts," "isolated," "strain," "vulnerable," and "doctors." Each of these terms becomes closely aligned with pandemic-related phenomena that is not explicitly linked to mental-health. We provide an overview of the changes in Table 7.3 and supporting examples (i.e., embedding neighborhoods) in Table 7.4.

	Token	2019 Context	2020 Context
Twitter	Panic	rage, meltdown, anxiety, anger, barrage, migraine, phobia, outrage, manic, rush, asthma	hysteria, chaos, fear, misinformation, confusion, frenzy, paranoia, mayhem, insanity, fearmongering
	Eviction	deportation, emergency, cancellation, negligence, reservation, injunction, immediate, incarceration, inaction, cancellations, termination	evictions, repayment, foreclosure, indefinite, immediate, injunction, moratorium, visitation, bankruptcy, deportation
	Crisis	humanitarian crisis, crises, catastrophe, epidemic, disaster, threat, emergency, issue, income inequality	pandemic, crises, crisi, catastrophe, #pandemic, cris, pand, disaster, epidemic, outbreak, pandem, pandemi, downturn
	Corona	bourbon, margarita, mesa, distillery, carmel, lager, grove, hut, riverside, fireball, mayo, scottsdale, tempe	carona, covid, coro, virus, #corona, rona, #covid_19, #coronavirus, ebola, #covid, vir, #covid2019, #wuhanvirus
	Doctors	psychiatrists, medically, clinic, cps, accountants, police, miscarriages, malpractice, abortions, prescribe, counseling, procedure	midwives, #nurses, #doctors, epidemiologists, emts, front-line, #coronawarriors, frontliners, virologists, masks, ppe, respirators, docs, heroes
	Zoom	zooming, zooms, tap, log, zoomed, hop, hover, dial, jump, slide, camera, plugged, optical, infrared, roll, nikon	skype, webex, #zoom, hangouts, #microsoftteams, webcam, facetime, telephone, livestream, classroom
Reddit	Floyd	dwight, wesley, sheldon, eddie, andy ruiz, mayfield, wade, ronnie, willie, holloway, george, henry, albert	floyd's, floyd's, #georgefloyd, floyds, ahmaud arbery, #georgefloyd's, #ahmaudarbery, arbery, zimmerman, carlin, geor, pell
	Panic	rage, anger, despair, desperation, reflex, terror, adrenaline, silence, anxiety, dread, paranoia, laughter	hysteria, fear, panicking, paranoia, civil unrest, toilet paper, adrenaline, diarrhea, virus, corona, anxiety, shutdowns, #panicbuying
	Cuts	cut, jumps, runs, cutting, pulls, moves, bounces, falls, turns, burns, drags, dips, breaks, bursts, rips, goes, bumps	cut, cutting, subsidies, budgets, deductions, revenues, checks, payments, breaks, deals, figures, loans, deposits, gains
	Isolated	unpleasant, unstable, detached, unsafe, populated, invasive, unknown, confined, endangered, absent, vulnerable, insulated	quarantined, isolating, separated, enclosed, insulated, infectious, confined, active, populated, autonomous, vulnerable, detached
	Vulnerable	susceptible, dangerous, prone, unstable, aggressive, hostile, disruptive, detrimental, receptive, fragile, damaging, sensitive	susceptible, dangerous, immunocompromised, infectious, isolating, elderly, disadvantaged, contagious, tolerant, likely, isolated, symptomatic
	Looting	spawning, loot, raiding, farming, grinding, camping, sniping, reloading, spawn, mobs, afk, fighting, hunting	rioting, looters, rioters, riots, arson, vandalism, protesting, blm, protest, peaceful, protestors, protesters, riot, protests, violence
	Relief	satisfaction, sadness, fatigue, disappointment, warmth, despair, desperation, payoff, excitement, dread, nausea, accomplishment	assistance, stimulus, aid, bailout, funding, compensation, temporary, medicaid, fund, bailouts, cheque, unemployment, benefits, donations, payout
	Strain	inflammation, deficiency, dose, stress, pressure, calcium, medication, concentration, tissue, nausea, receptors, doses, acne, effect, discomfort	disease, illness, infections, symptom, mutation, virus, outbreak, pneumonia, infection, strains, influenza, epidemic, dependency, diseases, allergy, flu, coronaviruses
	Testing	test, tests, qa, certification, training, scans, screening, monitoring, research, filtering, tested, imaging, treatments, coding, experiments	tests, contact tracing, screening, test, containment, infections, transmission, tracing, infection, ppe, ventilators, tested, lockdowns

Table 7.4: Examples of embedding neighborhoods for terms which experienced significant semantic shift from 2019 to 2020 according to our semantic shift measure.

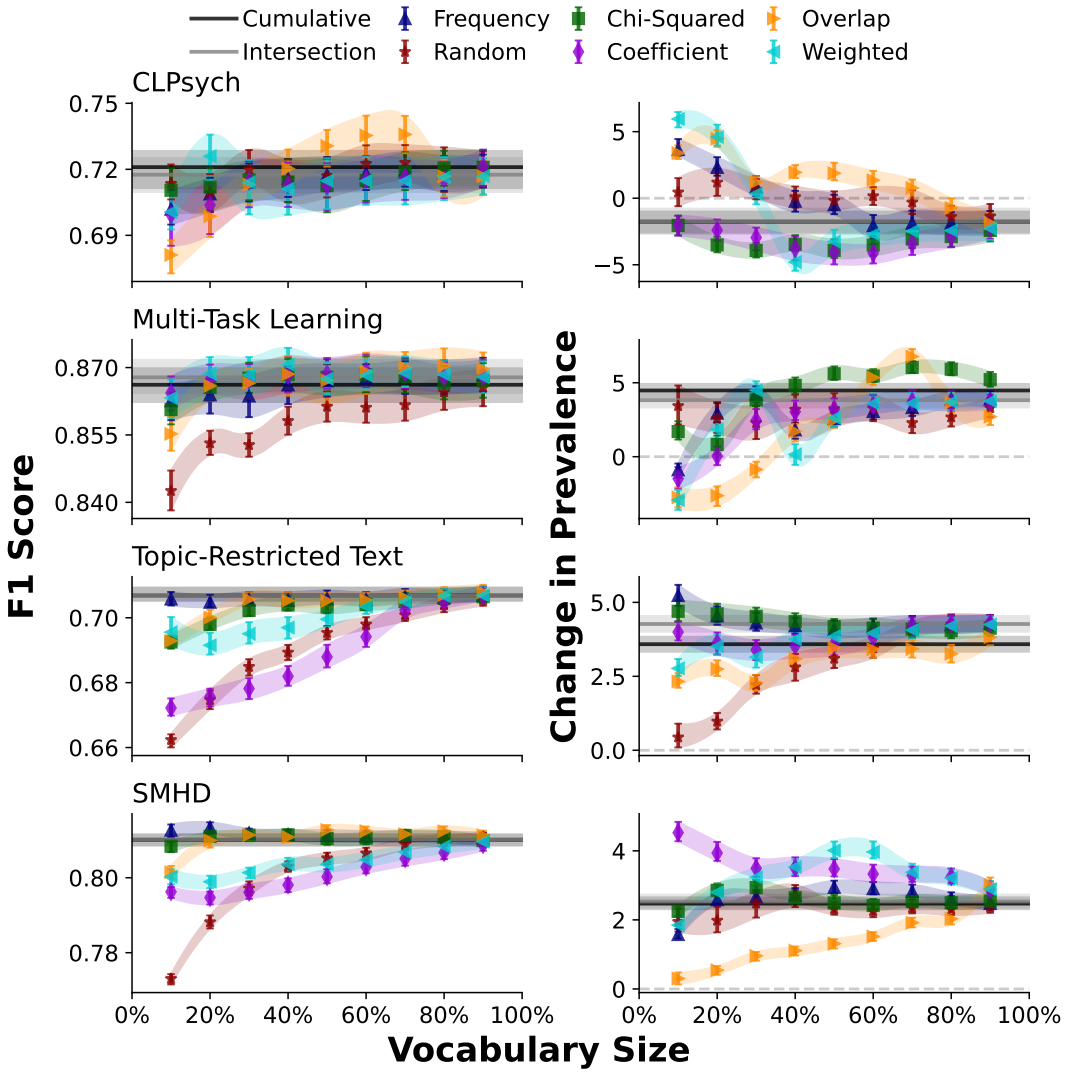


Figure 7.2: Horizontal bars denote each dataset's estimate under the naïve, Intersection baseline. Curves denote performance over varying sizes of vocabulary selected based on semantic stability S relative to the unlabeled datasets. (Left) Mean F1 score within held-out samples drawn from each dataset's complete time period. Performance is largely indistinguishable for several of the vocabulary sizes. (Right) Estimated change in depression prevalence as a function of vocabulary.

7.8.2.2 Quantitative Analysis

Turning our attention to the statistical classifiers, we observe that predictive performance as a function of the underlying vocabulary is nearly indistinguishable for vocabularies of size 40% and higher. However, as shown in Figure 7.2, we identify significant differences in the estimated change in population-level depression prevalence as a function of the model's underlying vocabulary. In some cases, these differences are relatively minor and lead to the same general conclusions. In other cases, we arrive at entirely different statements regarding the directional change of depression prevalence (i.e., increase instead of decrease) and absolute change (i.e., nearly 10% in the case of the minimum and maximum Multi-Task Learning estimates).

7.9 Review of Learnings

In this study, we demonstrated that semantic shift can be problematic in longitudinal monitoring applications, both in terms of pure predictive performance and our ability to estimate population-level outcomes. The method for measuring semantic stability introduced by Gonen et al. (2020) and adapted by us for use as a feature selection method here is promising for reducing domain divergence and improving generalization over time. However, more research must be done to understand which deployment

scenarios may obtain the most significant benefit from its use.

7.10 Limitations

The outcomes of this study are designed to spur conversation amongst practitioners, not necessarily to provide a panacea for addressing semantic noise in deployment scenarios. Indeed, we recognize our quantitative experiments are limited by the annotated data itself. For example, it remains to be seen whether semantically stable vocabularies are most useful over the course of certain time frames (e.g., decades instead of years), within a subset of social media platforms, or in the context of specific modeling tasks. Moreover, the labeled datasets may not be entirely conducive to the longitudinal classification experiments we performed in §7.7 (DeMasi et al., 2017; Tsakalidis et al., 2018), with depression known to present episodically within individuals (Collishaw et al., 2004; Angst et al., 2009). Given these dataset constraints, we urge researchers to consider replicating our analysis using new datasets which feature different underlying temporal dynamics.

Additionally, we foresee substantial value in continued exploration of semantic stability's effect on predictive generalization under alternative technical perspectives. For example, we note that our current study focuses solely on discrete time windows, an abstraction that is useful for simple monitoring applications, but too constraining

for others. It would be of significant value to the longitudinal monitoring community to evaluate whether continuous time and diachronic embeddings offer advantages over their discretized counterparts (Hamilton et al., 2016; Huang and Paul, 2019). We also recognize that our implementation operates in two distinct stages (i.e., feature selection, model training), a setup which may inhibit performance. A better approach may involve leveraging knowledge of semantic shift to explicitly regularize coefficients at training time.

7.11 Ethical Considerations

Institutional Oversight. This research was deemed exempt from review by our Institutional Review Board (IRB) under 45 CFR §46.104. All datasets in this study are either publicly accessible (via authenticated application programming interfaces) or available through secure distribution mechanisms (i.e., non-commercial data usage agreements). Given the sensitive nature of mental health data, we abide by additional protocols enumerated in Benton et al. (2017a) to govern all data collection, storage, and analysis. Nonetheless, we would be remiss not to acknowledge that population monitoring at scale warrants an additional ethical discussion. We discuss some trade-offs between risk and reward, specifically when studying sensitive characteristics (e.g., mental health status) from social media. We direct the reader to work from

Conway and O'Connor (2016) and Golder et al. (2017) for an expanded review.

Risks. Two serious risks arise from measuring personal characteristics using social media data: 1) discrimination, and 2) measurement error. The former is a challenge associated with any approach used to acquire information about human characteristics or behavior, whether inferred by an algorithm or not. Knowledge of personal attributes could be used by educational institutions to make biased admissions decisions, by law enforcement to track individuals without cause, or by political/government entities to target vulnerable individuals. These concerns are particularly poignant with regards to stigmatized characteristics, such as mental illness. Discriminatory actions based on these characteristics could have long-lasting financial and social consequences – e.g., difficulty obtaining loans, increased insurance premiums, and exclusion from certain communities. While statistical models are not the only method for gathering this information, they can be used in some situations where other approaches are infeasible (Paul et al., 2016). With respect to the second challenge, we draw the reader's attention to substantial evidence that demonstrates language models trained on social media datasets perform disproportionately amongst different demographic groups (Aguirre et al., 2021) and maintain historical social biases (Brunet et al., 2019). These systematic errors in models of mental health may further exacerbate social stratification in a opaque and elusive manner (Bender et al., 2021).

Rewards. We must also take care not to ignore the tremendous need for these methods and the benefits they bring. The same technology used to ostracize vulnerable individuals could also be used to provide those individuals with social services. Likewise, access to reasonably accurate classifiers with well-defined bounds of uncertainty could help small organizations acquire sufficient data to optimize resource allocation without needing to invest in the cost-prohibitive infrastructure necessary to execute traditional monitoring at scale (e.g., random digit dialing, online surveys) (Vaske, 2011; Shaver et al., 2019). These opportunities come with a variety of additional advantages over traditional population monitoring mechanisms – social media monitoring preempts the need to use downstream outcomes that are not useful in situations that require immediate decisions (e.g., latent changes in suicide rate), addresses certain forms of sample bias (e.g., selection bias introduced when individuals opt into a survey, disclosure bias that emerges when individuals are hesitant to discuss stigmatized topics with an interviewer), and provides the opportunity to make comparisons against retrospective baselines. Moreover, a significant amount of work focuses on methods to mitigate risk of discrimination (Zhao et al., 2018b) and adequately correct for sampling biases specific to social media data (Giorgi et al., 2021). The large body of literature on social media monitoring in public health, for example, evidences the tremendous need for these technologies (Paul and Dredze, 2017). It is

our responsibility to develop and deploy them in an ethically responsible manner.

Discussion. Practitioners must weigh these trade-offs in the context of their particular use case. In our use case, we note the goal of this study is *not* to make claims about a particular longitudinal trend or even demonstrate the prowess of a statistical modeling approach. Rather, our intention is to understand whether existing models can be trusted for measuring longitudinal trends at all in the presence of semantic shift, and if not, identify potential opportunities for practitioners to improve reliability of their models. The utility of such an exploration would be questionable if these types of models had not already been deployed in academia and beyond. However, one need only to look at research published within the last year regarding COVID-19 to see that machine learning classifiers are actively being used to understand a variety of social dynamics, ranging from mental health outcomes (Fine et al., 2020; Tabak and Purver, 2020) to transportation usage (Morshed et al., 2021). These analyses will form a foundation for public policy in the coming post-pandemic years. It is critical that we answer: are these results reliable?

7.12 Discussion

When we began this project in July 2020, we had the same intention as many other ML and NLP practitioners – to leverage our knowledge and myriad of existing

computational tools in support of individuals making critical pandemic-related decisions. After spending several months trying to understand why the models of mental health status that we had previously developed were yielding highly-variable and unreasonable results, we finally looked more closely at the predictions being made and realized that semantic shift was likely to blame. At that point, we made it our goal to rigorously quantify the post-pandemic semantic shift and understand how it could affect the emerging set of results being published by ML and NLP practitioners.

In this regard, our study succeeded; we identified a numerical method to measure semantic shift and then adapted it to test our hypothesis that semantic shift was driving the variability we were observing. That the feature selection method we developed to test our hypothesis also benefited generalization was an added bonus. We might as well have developed a method that sacrificed a small amount of predictive performance to have less variability in new domains and data distributions. In fact, under alternative settings, practitioners may find that our method actually has this effect. Indeed, for the purpose of measuring longitudinal change, we note that such an outcome isn't necessarily a deal breaker. A higher error rate that is stable across time periods would likely be more useful in this application setting than a error rate

that changes over time.[6]

All this said, we would be remiss to suggest that the feature selection method we propose is perfect. In its current construction, the method does not support transfer between more than two distributions, nor does it readily accommodate shifts that occur along a continuous dimension (e.g., geographic, continuous time). We would also be remiss to claim that the method is entirely novel. While it is true that efforts to validate and extend existing methods are often just as important to the broader research community as the original methods themselves, much of the credit for our feature selection method should still be attributed to Gonen et al. (2020), who proposed the semantic shift measurement upon which the feature selection method is predicated.

7.13 Looking Ahead

Contemporary language models built using transformers (Vaswani et al., 2017) have been a notable omission from this chapter. Since this study was originally published, the NLP community has moved even more firmly away from traditional feature engineering approaches to instead leverage end-to-end neural language models that implicitly learn low-dimensional feature representations of text (e.g., BERT (Devlin

[6] It remains an open question whether average accuracy and worst-case accuracy are reconcilable (Tsipras et al., 2019; Michel et al., 2021; Robey et al., 2022). For now, the tolerances and objectives of a system are likely to influence which performance measure guides the model selection process.

et al., 2019), T5 (Raffel et al., 2020), GPT (Radford et al., 2018)). While the feature selection method we proposed may still have utility in health applications that require certain degrees of interpretability or are compute-limited, it would be unwise not to acknowledge the significant potential (and challenges) presented by the contemporary language models. In the next chapter, we focus on the adapting contemporary language models to new distributions, and offer evidence that affirms domain adaptation as a preferable approach over domain generalization in the health domain.

Chapter 8

Do Clinical Language Models Generalize?

8.1 Overview

Our efforts so far have focused primarily on non-clinical data (i.e., social media) in the context of health applications. Nonetheless, as discussed in Chapter 4, clinical data is likely to face similar challenges – the datasets are often small, collected using non-trivial enrichment mechanisms, and may not necessarily produce generalizable models across populations. Language models (LMs) trained on large amounts of unsupervised text have been cast as a potential mitigation strategy for training downstream models on small, task-focused clinical datasets. However, choosing an LM that is appropriate for downstream clinical tasks is not necessarily straightforward.

Contemporary work argues that LMs pretrained on generic web data are inferior to LMs pretrained on clinical data (Lehman et al., 2023). In this chapter derived from Harrigian et al. (2023a), we provide novel evidence using a newly curated Ophthalmology phenotyping dataset that LMs trained on generic web data actually perform on par with LMs trained on out-of-distribution clinical data. Our results emphasize the heterogeneity of clinical language data and temper existing claims regarding the necessity of clinical LMs and the manner in which they should be trained.

8.2 Background

Diabetic eye disease (e.g., diabetic retinopathy, diabetic macular edema) is a major cause of blindness worldwide (Steinmetz et al., 2021; Wykoff et al., 2021). These conditions can develop in patients with diabetes, whereby elevated sugar levels in the blood can cause damage to the retinal blood vessels. Management of diabetic eye disease relies not only on a patient's glycemic control, but also on regular ophthalmic screening for the early detection and treatment of vision threatening complications (Solomon et al., 2017; Flaxel et al., 2020). The ability to monitor clinical trajectories and efficiently detect lapses in care is critical to achieving the latter objectives. Unfortunately, structured data in the electronic health record (EHR) remains ill-suited for describing many ophthalmic conditions at the granularity necessary to support these goals (Cai et al., 2021). Much of the critical information is found only in the free text of the EHR.

Provided sufficient support at training time, supervised machine learning models can extract useful clinical information from free text in the EHR to augment structured metadata (Voorham and Denig, 2007; McCoy Jr et al., 2017; Koleck et al., 2019). However, annotated clinical datasets are typically small by contemporary standards due to the inherent bottleneck imposed by the necessity of involving highly trained domain experts (i.e., physicians and other healthcare professionals) (Spasic, Nenadic,

et al., 2020). This limitation is compounded in the diabetic eye disease use case, where the number of clinical concepts needed to effectively monitor the condition is large (Pearce et al., 2019; Gale et al., 2021) and their associated attributes (e.g., severity, temporality) are heavily class imbalanced (Yau et al., 2012; Yang et al., 2019b).

8.3 Motivation and Contribution

Language models (LMs) pretrained on massive text corpora are a powerful tool for representing language across a variety of downstream modeling tasks (Howard and Ruder, 2018; Wei et al., 2021). In low-resource settings particularly, pretraining can inject useful knowledge for differentiating linguistic instances in context (Gao et al., 2021). This fact has inspired the training and release of several models trained on biomedical and clinical text over the last several years (Lee et al., 2020b; Gu et al., 2021). Nonetheless, the success of LMs as encoders in the clinical domain and beyond is typically correlated with the degree of alignment between pretraining and task-specific language distributions (Roberts, 2016; Gururangan et al., 2020; Talmor et al., 2020). The field of ophthalmology serves as a departure from the domains on which existing clinical language models have been trained (Alsentzer et al., 2019; Yang et al., 2022), requiring highly detailed knowledge of a single anatomical system. In developing a phenotyping system for diabetic eye disease, we find ourselves in a

unique position to answer an important question: Do clinical LMs perform consistently better than non-clinical LMs on out-of-domain clinical data?

In the remainder of this chapter, we conduct an empirical investigation of multiple BERT encoders and training paradigms, allowing us to evaluate the sufficiency of existing BERT language models in a specialized clinical domain. In contrast to common perceptions about clinical language models, we find that the LMs trained on out-of-domain clinical data provide little-to-no benefit in our domain compared to the LMs trained on non-clinical data. Furthermore, advantages derived from an initial pretraining phase can be nullified almost entirely via tailored in-domain pretraining. Given the ubiquity of distribution shift and scarcity of data across clinical NLP use cases, our results suggest that the research community may benefit from focusing on adapting language models to low-resource clinical settings instead of training "general" clinical language models from scratch.

8.4 Related Work

8.4.1 NLP in Ophthalmology

Diabetic eye disease refers to a collection of eye problems that can result from diabetes, including diabetic retinopathy (DR) and diabetic macular edema (DME) (Solomon et

al., 2017; Flaxel et al., 2020). DR is a progressive disease caused by insufficient blood flow to the retina which, in its most severe state, sees the growth of abnormal blood vessels around the retina. This process, referred to generally as neovascularization, may lead to vision-threatening complications such as vitreous hemorrhage, retinal detachment, and blindness (Steinmetz et al., 2021).

Although advanced stages of diabetic eye disease cannot be reversed, treatments can prevent the condition from worsening and even return some visual fidelity if delivered in a timely manner (Duh et al., 2017). Follow-up timelines depend on patient-specific trajectories which are specified only with the free text of the EHR (Cai et al., 2021). Extraction and synthesis of this information has the potential to dramatically reduce the rate at which patients are lost to follow up, for example by introducing automatic notifications regarding delayed treatment (Gale et al., 2021).

There exists a brief, albeit rich, history of artificial intelligence systems targeting problems in the field of ophthalmology (Grewal et al., 2018; Ting et al., 2019). The majority of effort has been allocated to improving imaging diagnostics via computer vision (Teikari et al., 2019) and building clinical decision support systems using structured EHR data (Jacoba et al., 2021; Ogunyemi et al., 2021). Systems leveraging natural language processing (NLP) techniques to improve ophthalmic care make up the minority of these efforts and have typically focused on narrow concept extraction

objectives (Liu et al., 2017a; Mao et al., 2017). However, the rise of large LMs, such as GPT-3, has drawn increased attention to the NLP research community from ophthalmologists interested in better synthesizing free text in the EHR (Yang et al., 2021b; Nath et al., 2022; Antaki et al., 2023).

The most similar work to our own comes from Yu et al. (2022). Although our work shares a common objective in extracting concepts related to diabetic retinopathy and linking their associated attributes, there are multiple key differences. First, Yu et al. (2022) focus on imaging reports from patients already diagnosed with diabetic retinopathy, whereas we focus on progress notes and problem lists from a general ophthalmology patient population. Second, our ontology of clinical concepts is larger and more diverse (e.g., we include comorbidities and treatments). Finally, they approach the assignment of attributes to clinical concepts as a relation extraction task, which assumes overt evidence of each attribute in the free text. In contrast, we assume some attributes are not explicitly stated in the text, but can be inferred based on context and reasoning.

8.4.2 Clinical Language Modeling

Our understanding of the value of clinical language models has evolved significantly over time, but remains far from complete. Early work in neural language modeling

demonstrated that word embeddings learned using clinical and biomedical text can improve performance in downstream clinical tasks compared to embeddings trained on general web data (Wu et al., 2015; Dingwall and Potts, 2018). Similar results have emerged for contextual language models (Alsentzer et al., 2019; Khattak et al., 2019; Lee et al., 2020b).

Why do these clinical language models typically outperform their generic counterparts on clinical tasks? The primary hypothesis is that pretraining on clinical data is necessary to address the distributional shift that occurs from non-clinical to clinical settings (Naik et al., 2021; Lamproudis et al., 2022). Common examples include changes in the distribution of word senses (e.g., aggressive treatment regimen, aggressive behavior) and the introduction of medicine-specific terminology (e.g., abbreviations, diagnoses, etc.) (Wu and Liu, 2011; Liu et al., 2012). Research from Lewis et al. (2020) and Lehman et al. (2023) has suggested that domain-specific vocabularies are invaluable for allowing clinical language models to learn semantics in a more parameter-efficient manner. Recently, these tenets have motivated the development of domain-specific GPT-style models outside of the clinical space (Taylor et al., 2022; Venigalla et al., 2022; Wu et al., 2023).

At the same time, other researchers have shown that non-clinical language models, provided sufficient size, are still able to perform remarkably well in clinical and

clinical-adjacent tasks (Agrawal et al., 2022; Harrigian et al., 2023b; Singhal et al., 2023a). Moreover, domain-specific vocabularies tailored for clinical tasks do not consistently provide comprehensive performance improvements (Gutiérrez et al., 2023). Are these negative results an anomaly? Or do they more accurately represent the capabilities of clinical language models?

Unfortunately, challenges with sharing sensitive clinical data have thus far limited the strength of conclusions we can draw regarding the value of clinical language models. Many of the public datasets that are used to evaluate clinical language models are drawn from the same public datasets used to train the language models and generally cover a narrow range of clinical tasks (Thirunavukarasu et al., 2023; Wornow et al., 2023). It is unclear whether available clinical language models are better on clinical data broadly, or on the specific medical speciality for which they are trained. In this study, we leverage a unique clinical dataset to evaluate the sufficiency of available clinical language models on a new clinical domain. Since not all clinical datasets are drawn from the same domain, it is important to determine the advantages available clinical models provide for work on the diverse range of clinical domains.

8.5 Data

To the best of our knowledge, only one EHR dataset containing diabetic eye disease annotations exists (Yu et al., 2022). In addition to not being publicly available, this dataset has several shortcomings that make it suboptimal for our use case (e.g., enriched patient population, imaging report focus, extractive modeling setting, §8.4.1). To ensure that we can cover the breadth of concepts and attributes required to monitor diabetic eye disease in our patient population, we curate a new clinical note dataset from scratch.

8.5.1 Inclusion Criteria

All ophthalmology-related visits to our institute's hospital system from January 1, 2013 through April 1, 2022 were considered candidates for the study. Visits for imaging services and visits which lacked either a progress note or problem list (Weed, 1968) were excluded, leaving a total of 692,486 visits by 91,097 patients. Notes were processed adhering to our institution's privacy policy after approval by our Institutional Review Board (IRB).

8.5.2 Concept Ontology

We developed a multi-level ontology for 19 clinical concepts with significance in the management and treatment of diabetic eye disease. Concepts with similar clinical relevance were grouped together into higher-level semantic categories (e.g., Retina Conditions, Complications of Diabetes Mellitus). Each concept was further associated with modifiers within up to 3 attribute categories (i.e., Laterality, Temporality, and Severity/Type). The ontology was fine-tuned over multiple iterations of pilot annotation experiments to balance label utility with the cognitive load required by annotators to apply the ontology consistently. We present the ontology in Table 8.1, with a list of relevant abbreviations in Table 8.8 (see §8.12).

8.5.3 Annotation Strategy

Ophthalmology notes commonly refer to the same clinical concept multiple times, albeit with different attributes, thus rendering note-level application of our ontology inappropriate. Span-level annotation was necessary, but non-trivial. Within pilot experiments, our domain experts found it challenging to consistently identify spans across the relatively wide label space. Moreover, data privacy and technical limitations made it infeasible to deploy existing annotation software capable of both span identification and labeling.

Group	Concept	Laterality	Temporality	Severity/Type
Retina Conditions	A1 - DR (General)	OS, OD, OU	Active, History of	–
	A2 - NPDR	OS, OD, OU	Active, History of	Mild, Mild-Moderate, Moderate, Moderate-Severe, Severe
	A3 - PDR	OS, OD, OU	Active, History of	NHR-PDR, HR-PDR
	A4 - NV	OS, OD, OU	Active, Resolved	Iris, Iris + NVD and/or NVE, NVD, NVE, NVD/NVE, AMD, Other
	B1 - ME	OS, OD, OU	Active, History of	DME, CI-DME, Non-CI-DME, CS-DME, Non-CS-DME, CME, AMD, Other
Complications of Retina Conditions	C1 - VH	OS, OD, OU	Active, History of, Resolving, Resolved	–
	C2 - RD	OS, OD, OU	Active, History of	RRD, TRD, Serous, Combined RRD/TRD
	C3 - NVG	OS, OD, OU	Present, Not Present	–
Treatment (Procedure)	D1 - Anti-VEGF	OS, OD, OU	History of, Performed Today, Recommended, Considering	–
	D2 - PRP	OS, OD, OU	History of, Performed Today, Recommended, Considering	–
	D3 - Focal Grid Laser	OS, OD, OU	History of, Performed Today, Recommended, Considering	–
	D4 - Other Intravitreal Injections	OS, OD, OU	History of, Performed Today, Recommended, Considering	–
Treatment (Surgery)	E1 - Retina Surgery	OS, OD, OU	History of, Performed Today, Recommended, Considering	Indication VH, Indication RD
	E2 - NVG Surgery	OS, OD, OU	History of, Performed Today, Recommended, Considering	Tube, Trab, MIGS
Comorbidities	F1 - Diabetes Mellitus	–	Active, Resolved	Type I, Type II, Gestational, Other
Complications of Diabetes Mellitus	G1 - Nephropathy	–	Present, Not Present	–
	G2 - Neuropathy	–	Present, Not Present	–
	G3 - Heart Attack	–	History of, No History of	–
	G4 - Stroke	–	History of, No History of	–

Table 8.1: Ontology of concepts related to diabetic eye disease. We include definitions for all abbreviations in Table 8.8. Where appropriate, temporality classes can be negated (e.g., Negation + History of = No History of).

Document ID: ad1fc53fe509fdea65d2099d8b5b3c57a8b5d1978f9bb8f0fa5eb1c427015aaf
Encounter Date: 2015-06-21

[[[ENCOUNTER ICD-10 CODES]]]
[[E11.319: Diabetic retinopathy]]
[[E11.311: Diabetic macular edema, both eyes]]

[[[PROBLEM LIST]]]
[[E11.3313: Diabetic macular edema of both eyes with moderate nonproliferative diabetic retinopathy associated with Type 2 diabetes mellitus]]
[OVERVIEW]
Eylea initiated right eye 6/2015 and and left eye 5/2014. No progression to PDR.
[ASSESSMENT & PLAN]
Right eye has foveal edema. Eylea #1 given. She will return in 2 weeks for eylea left eye after vacation to MI.

Start	End	Concept	Text Span	Context	Laterality	Severity/Type	Temporality	Negated	Incorrect
31	38	DR (General)	E11.319	[[«E11.319»: Diabetic Retinopathy]]	▾ OU	–	▾ Active	▾	▾
31	38	DM	E11.319	[[«E11.319»: Diabetic Retinopathy]]	–	▾ Type 2	▾ Active	▾	▾
267	270	PDR	PDR	left eye 5/2014. No progression to «PDR».	▾ OU	▾	▾ Active	▾ Negated	▾
				...					
525	537	ME	foveal edema	Right eye has «foveal edema». Eylea #1	▾ OD	▾ CI-DME	▾ Active	▾	▾
601	603	Heart Attack	MI	eylea left eye after vacation to «MI».	–	–	▾	▾	▾ Incorrect

Figure 8.1: Interface displayed to annotators in Microsoft Excel. Drop-down data validation cells provide possible attribute labels conditioned on each row's clinical concept, with irrelevant attributes denoted using a '–' symbol. If an attribute is not clearly specified within a note or not inferable via context (e.g., severity of PDR), the annotator is instructed to leave the cell blank; we treat these instances as missing data at training time.

As an alternative strategy, we curated a set of regular expressions to identify candidate concept spans which could then be shown to annotators to validate correctness and assign appropriate attribute labels. Expressions were applied to the free text of the note first, and then to diagnostic codes (i.e., International Classification of Diseases, Tenth Revision (ICD-10)) contained in the problem list and note metadata. To limit redundant annotation efforts, concepts found in the diagnostic codes were excluded if already found in the free text. Patterns for the free-text were developed iteratively with domain experts, while relevant diagnostic codes were identified using

an online database.[1] Concept matches were organized by encounter and displayed in context to facilitate span-level attribute annotation (see Figure 8.1).

Two annotators with clinical expertise in diabetic eye disease (a PGY-4 ophthalmology resident and a licensed optometrist) were responsible for all annotations. Efforts were overseen by a board-certified ophthalmologist, an author of this study, who designed the concept ontology. Both annotators had access to an annotation guide containing edge-case examples and other rules for applying the concept ontology to the free text notes. Annotation was completed in a secure remote desktop environment using a Microsoft Excel workbook outfitted with conditional data validation cells and custom text formatting (see Figure 8.1). Notes were shared with us in sanitized CSV files, thus necessitating the use of additional rules to format the text to be more readable (e.g., line-breaks, removal of tables containing lab results). Metadata for the encounter (i.e., ICD-10 codes) was included in the formatted note and annotated in the same manner as free text.

Annotation was completed during two phases, each consisting of multiple rounds.

Phase 1. During the first phase, a random sample of 236 notes from 139 patients containing at least one concept span were labeled. During the first round of the first phase, annotators labeled 3,013 concept-spans independently. During

[1] https://www.icd10data.com

the second round, annotators were asked to independently relabel all concept spans which contained disagreement during the first round (869 spans). During the third and final round, annotators resolved any remaining disagreements via discussion (501 spans).

Phase 2. A review of the label distribution after the first phase of annotation suggested that more annotation would be necessary to address severe class and concept imbalance. For the second phase of annotation, a random sample of 500 notes from 209 patients (different from those in the first phase) where labeled. During the first round of the second phase, annotators labeled 3,552 concept-spans independently. Rather than including another round of independent review as was done during the first phase, annotators met immediately after the first round to resolve any disagreement via discussion (790 spans).

In total, annotators independently reviewed notes for 736 clinical encounters from a random sample of 348 patients. A total of 12,723 attribute labels were generated from 6,565 spans (see Table 8.2).

8.5.4 Concept Extraction

We ran multiple pilot experiments to optimize the annotation procedure described above. At the beginning of the study, our institution hosted internally-developed, HIPAA-compliant software which supported span-level extraction and annotation.[2] Unfortunately, this service was discontinued after a single pilot experiment, in turn requiring us to seek alternative annotation protocols. External software could not be deployed easily while maintaining patient privacy. At the same time, our pilot experiment suggested that annotators would struggle to identify relevant concept spans over the large label space in an efficient manner without compromising accuracy. Together, these circumstances motivated us to develop regular expressions which could be used to automatically identify relevant concept spans which would then be manually reviewed and further annotated by annotators.

We used labels generated from the HIPAA-compliant annotation software during the aforementioned pilot experiment in combination with domain knowledge from members of our team to construct the base set of regular expressions. We then augmented this set with ICD-10 codes based on data from an online resource.[3] ICD-10 code matches were only included for annotation if their associated clinical concepts

[2] https://github.com/JHUAPL/PINE

[3] https://www.icd10data.com

Concept	Valid	Invalid	Laterality	Severity/ Type	Temporality
A1 - DR (Generic)	455	3	446	0	425
A2 - NPDR	194	0	189	165	187
A3 - PDR	209	0	203	170	189
A4 - NV	360	1	354	332	355
B1 - ME	896	12	848	845	777
C1 - VH	266	1	264	0	253
C2 - RD	275	1	258	145	263
C3 - NVG	369	0	356	0	355
D1 - Anti-VEGF	508	0	501	0	465
D2 - PRP	211	0	205	0	207
D3 - Focal Laser	16	0	16	0	16
D4 - Other Injections	16	0	16	0	16
E1 - Retina Surgery	226	37	212	118	198
E2 - NVG Surgery	70	4	63	63	65
F1 - Diabetes	1,045	0	0	785	1,029
G1 - Nephropathy	340	0	0	0	337
G2 - Neuropathy	351	0	0	0	345
G3 - Heart Attack	342	1	0	0	336
G4 - Stroke	355	1	0	0	351
Total	6,504	61	3,931	2,623	6,169

Table 8.2: The distribution of spans and attribute labels for the 19 clinical concepts in our ontology.

were not already found in the note free text; this decision was made to limit redundant annotation efforts. In general, regular expressions were designed to capture the most generic manifestation of each concept (e.g., "NPDR" instead of "Severe NPDR"). We provide our expressions and parsing logic in the supplemental material.[4]

The distribution of concept-spans and attribute labels is provided in Table 8.2. False positives are those which were marked as "Invalid" spans by annotators. An example of an invalid span is provided in Figure 8.1 – "eylea left eye after vacation

[4] https://github.com/kharrigian/ml4h-clinical-bert

to «MI»" – where the proposed span "«MI»" refers to the geographic location of Michigan, not Myocardial Infarction. As seen in the table, precision of the regular expressions varies as a function of clinical concept. The lowest level of precision (0.86) occurs for Retina Surgery, with mentions of 'laser' being the largest source of error. The overall precision across all concepts spans was greater than 0.99.

We initially planned to train classifiers to infer span validity for all clinical concepts. However, as shown in Table 8.2, only two concepts – Macular Edema, Retina Surgery – had enough invalid spans to support/warrant this modeling. Spans which were marked as invalid by annotators were not used for training the attribute classification models; they were only used for training the Macular Edema and Retina Surgery span validity classifiers.

The aforementioned limitations regarding span-level annotation made it difficult to empirically estimate recall. However, we note that this system's goal is to improve recall of clinical concepts over what is possible through diagnostic codes alone. Accordingly, in Table 8.3, we show the distribution of concept matches in the entire 692,486 note dataset broken down by location of the match. The right-most column is most important, indicating the number of additional notes which were identified as relevant to our clinical use case by examining the free-text. Some relevant concepts

Clinical Concept	ICD-10	∩	Text
A1 - DR (General)	37,743	85,878	55,848
A2 - NPDR	24,970	28,203	14,594
A3 - PDR	14,217	32,212	13,038
A4 - NV	7,624	14,945	58,970
B1 - ME	37,081	83,852	50,719
C1 - VH	3,569	4,882	28,691
C2 - RD	4,337	15,279	70,781
C3 - NVG	155,429	6,127	3,418
D1 - Anti-VEGF	0	0	93,038
D2 - PRP	0	0	41,531
D3 - Focal Grid Laser	0	0	7,339
D4 - Other Injections	0	0	8,230
E1 - Retina Surgery	0	0	90,680
E2 - NVG Surgery	19	11	34,980
F1 - Diabetes Mellitus	44,019	165,813	54,499
G1 - Nephropathy	3,073	33	1,941
G2 - Neuropathy	5,615	363	5,997
G3 - Heart Attack	56	3	4,796
G4 - Stroke	783	921	10,452

Table 8.3: The number of notes containing at least one regular expression match for each clinical concept in our ontology, faceted by location of the match. Free-text search improves recall of relevant clinical concepts over using ICD-10 codes alone.

(e.g., surgical procedures) cannot be found by examining ICD-10 codes,[5] while others receive significantly higher coverage by looking at the free text. Besides improving recall as desired, the free-text search allows us to independently characterize multiple instances of the same concept (e.g., Mild NPDR → Moderate NPDR).

[5] At our institution, procedures are entered as structured metadata using Current Procedural Terminology (CPT) codes.

8.5.5 Task Consolidation

The annotated dataset exhibits sparsity for some of the clinical concepts and heavy class imbalance for the majority of attributes. Rather than train independent models for each concept-attribute pair in our dataset, we draw statistical strength from the overlapping label space by modeling multiple concepts jointly and merging attribute labels with high semantic similarity in the diabetic eye disease use case. For example, we model temporality for the retina conditions and their complications together, while also remapping the original temporality attribute labels into a binary output set indicating whether the concept is present (or relevant) at the time of the encounter.

The consolidated summary of classification tasks is presented in Table 8.4. In addition to the remapped labels, the reader may also note the exclusion of certain concept-attribute pairs altogether. We decide not to model temporality for `F1 - Diabetes Mellitus` due to extreme label imbalance – all but 9 spans are Active, and these 9 spans are concentrated in only 5 patients. We also decide to model span validity (i.e., whether the regular expression match was correct) for only `B1 - Macular Edema` and `E1 - Retina Surgery` because the number of false positives for the remaining concepts was small and unlikely to yield a useful model. The label distribution for each of the 14 tasks in provided in Table 8.5.

Attribute	Concepts	Concept IDs	Consolidated	Original
Temporality	Retina Conditions	A1, A2, A3, A4, B1, C1, C2, C3	Present	Active, ¬ Resolved, Resolving, Present
			Not Present	History of, ¬ Active, ¬ History of, Resolved
	Complications of Diabetes Mellitus	G1, G2, G3, G4	Present	Present, History of, Active, ¬ Resolved
			Not Present	Not Present, No History of, ¬ Active, Resolved
	Treatments	D1, D2, D3, D4, E1, E2	History of	–
			Performed Today	–
			Discussed	Recommended, Considering, ¬ History of, ¬ Performed Today, ¬ Recommended, ¬ Considering
Laterality	All	A1, A2, A3, A4, B1, C1, C2, C3, D1, D2, D3, D4, E1, E2	OS	–
			OD	–
			OU	–
Type	Neovascularization	A4	NVD and/or NVE	NVD, NVE, NVD/NVE
			Iris + NVD and/or NVE	–
			Iris	–
			AMD	–
			Other	–
	Macular Edema	B1	DME	DME, CI-DME, CS-DME, Non-CI-DME, Non-CS-DME
			Other	CME, AMD, Other
	Retinal Detachment	C2	RRD	–
			TRD	–
			Combined RRD/TRD	–
			Serous	–
	Diabetes Mellitus	F1	Type I	–
			Type II	–
			Other	Gestational, Other
	Retina Surgery	E1	Indication VH	–
			Indication RD	–
	NVG Surgery	E2	Tube	–
			Trab	–
			MIGS	–
Severity	NPDR	A2	Mild	Mild
			Moderate	Mild-Moderate, Moderate
			Severe	Moderate-Severe, Severe
	PDR	A3	HR-PDR	–
			NHR-PDR	–
Span Validity	Macular Edema	B1	Valid	–
			Invalid	–
	Retina Surgery	E1	Valid	–
			Invalid	–

Table 8.4: Consolidation of our concept ontology into 14 classification tasks. The (¬) symbol denotes negation.

		Concept ID ⟶	*A1*	*A2*	*A3*	*A4*	*B1*	*C1*	*C2*	*C3*	*D1*	*D2*	*D3*	*D4*	*E1*	*E2*	*F1*	*G1*	*G2*	*G3*	*G4*
Temporality	Retina	Not Present	309	7	65	270	385	133	235	326											
		Present	116	180	124	85	392	120	28	29											
	DM Complications	Not Present																334	331	334	331
		Present																3	14	2	20
	Treatment	History of									265	180	14	7	139	50					
		No Action									145	22	2	8	58	15					
		Performed									55	5	0	1	1	0					
Laterality	All	OD	20	25	49	99	236	97	87	27	217	79	1	3	119	24					
		OS	11	24	32	163	251	152	61	12	212	95	11	10	61	15					
		OU	415	140	122	92	361	15	110	317	72	31	4	3	32	24					
Type	ME	DME					672														
		Other					173														
	RD	RRD							97												
		TRD							48												
	NV	AMD				27															
		Iris				31															
		Iris + NVD/NVE				39															
		NVD/NVE				234															
		Other				1															
	DM	Other															9				
		Type 1															173				
		Type 2															603				
	NVG Surgery	MIGS														2					
		Trab														36					
		Tube														25					
	Retina Surgery	Indication RD													67						
		Indication VH													51						
Severity	NPDR	Mild		62																	
		Moderate		47																	
		Severe		56																	
	PDR	HR			161																
		NHR			9																
Span Validity	ME	False					12														
		True					896														
	Retina Surgery	False													37						
		True													226						

Table 8.5: The distribution of attribute labels for each of the consolidated tasks, broken down further by clinical concept.

8.6 Quantifying the Importance of Domain Adaptation

Although regular expressions and hand-crafted rules can be used to extract clinical concepts with moderately high precision and better recall than diagnostic codes, they are poorly suited for inferring attributes for the extracted concepts. As an example, consider the task of inferring laterality or negation for the PDR span in Figure 8.1. Contextual language models (CLMs) on the other hand have the ability to learn inter-token dependencies and, when explicit evidence is not available in the text, leverage prior knowledge to infer latent attributes (Devlin et al., 2019; Liu et al., 2019). However, training a CLM from scratch typically requires a substantial amount of data, a constraint that is difficult to satisfy in many clinical NLP settings.

To address data scarcity issues, we can instead use models pretrained on out-of-distribution data as a starting point and then fine-tune them for our target domain (Alsentzer et al., 2019). Nonetheless, maximizing performance under this regime is non-trivial. Which language model do we use as our foundation? Does the pretraining distribution matter? What about the vocabulary? We investigate these issues in the context of building our phenotyping system for diabetic eye disease.

8.6.1 Do clinical LMs outperform non-clinical LMs in the presence of clinical data distribution shift?

Prior studies have shown that LMs trained on clinical data may achieve better performance in downstream clinical tasks compared to LMs trained on non-clinical data (Lamproudis et al., 2022). However, this effect is not consistent across tasks and datasets (Lewis et al., 2020; Yang et al., 2022), potentially due to differences between the pretraining and target distributions. In this first experiment, we ask whether an LM trained on data from a significantly different clinical setting than ophthalmology still transfers better to ophthalmology-related tasks than an LM trained with non-clinical data.

Methods. Our primary task models consist of an encoder-style LM with a dense MLP output layer (see Figure 8.4 in §8.12). As a naive baseline, we consider a majority classifier conditioned jointly on the target concept and token span (see §8.12.4). We conduct a 2 x 2 factorial experiment. As the first factor, we compare downstream task performance achieved using a general purpose LM – BERT Base (Cased) (Devlin et al., 2019) – with the performance achieved using a clinical LM trained on notes from an ICU setting – Clinical BERT (Alsentzer et al., 2019). As

			BERT Base				Clinical BERT			
			w/o Continued Pretraining		w/ Continued Pretraining		w/o Continued Pretraining		w/ Continued Pretraining	
Attribute	**Concepts (# Classes)**	**Majority**	❄	🔥	❄	🔥	❄	🔥	❄	🔥
Temporality	Retina (K=2)	.76 (.75,.77)	.81 (.79,.82)	.83 (.82,.84)	.84 (.83,.86)	.87 (.85,.89)	.83 (.82,.84)	.84 (.83,.86)	.85 (.84,.87)	.87 (.85,.88)
	DM Complications (K=2)	.81 (.72,.87)	.81 (.73,.88)	.80 (.70,.89)	.81 (.71,.88)	.80 (.71,.88)	.80 (.70,.89)	.84 (.76,.90)	.84 (.73,.93)	.85 (.77,.92)
	Treatment (K=3)	.35 (.34,.36)	.59 (.55,.64)	.79 (.75,.82)	.81 (.79,.83)	.84 (.81,.86)	.69 (.65,.73)	.81 (.78,.84)	.81 (.77,.83)	.82 (.76,.85)
Laterality	All (K=3)	.53 (.52,.55)	.54 (.51,.57)	.84 (.83,.86)	.60 (.58,.61)	.92 (.90,.93)	.56 (.54,.58)	.84 (.81,.87)	.60 (.57,.63)	.90 (.89,.92)
Type	ME (K=2)	.83 (.77,.88)	.83 (.79,.88)	.87 (.79,.93)	.86 (.79,.91)	.88 (.82,.94)	.82 (.78,.86)	.87 (.80,.93)	.85 (.82,.89)	.90 (.85,.94)
	RD (K=4)	.68 (.52,.85)	.75 (.57,.89)	.79 (.58,.95)	.70 (.55,.83)	.82 (.76,.88)	.78 (.59,.94)	.81 (.61,.98)	.72 (.50,.95)	.87 (.78,.95)
	NV (K=5)	.79 (.63,.91)	.66 (.53,.80)	.75 (.60,.89)	.81 (.71,.91)	.82 (.72,.93)	.71 (.56,.87)	.78 (.65,.92)	.81 (.73,.90)	.77 (.66,.86)
	DM (K=3)	.41 (.30,.55)	.39 (.31,.53)	.40 (.32,.52)	.37 (.29,.52)	.57 (.43,.79)	.31 (.28,.35)	.40 (.31,.53)	.33 (.29,.38)	.54 (.40,.74)
	NVG Surgery (K=3)	.91 (.76,1.0)	.85 (.69,1.0)	.85 (.69,1.0)	.74 (.63,.88)	.85 (.69,1.0)	.79 (.54,1.0)	.85 (.69,1.0)	.79 (.63,.96)	.85 (.69,1.0)
	Retina Surgery (K=2)	.58 (.47,.68)	.51 (.38,.64)	.66 (.52,.79)	.60 (.44,.71)	.71 (.59,.83)	.59 (.46,.70)	.52 (.46,.59)	.64 (.53,.75)	.76 (.65,.85)
Severity	NPDR (K=3)	.54 (.48,.61)	.56 (.50,.62)	.83 (.73,.90)	.69 (.57,.84)	.89 (.77,.98)	.58 (.54,.61)	.91 (.87,.96)	.70 (.60,.83)	.95 (.90,.99)
	PDR (K=2)	.48 (.47,.49)	.45 (.43,.48)	.71 (.53,.89)	.39 (.29,.47)	.81 (.64,.93)	.43 (.37,.47)	.53 (.36,.77)	.45 (.43,.48)	.82 (.63,.98)
Span Validity	ME (K=2)	.60 (.49,.80)	.60 (.49,.72)	.55 (.49,.62)	.77 (.57,.97)	.81 (.64,.97)	.65 (.52,.78)	.56 (.50,.64)	.66 (.49,.86)	.83 (.65,.98)
	Retina Surgery (K=2)	.46 (.44,.47)	.77 (.72,.81)	.77 (.68,.87)	.83 (.77,.89)	.82 (.75,.90)	.76 (.69,.84)	.80 (.74,.86)	.83 (.76,.90)	.80 (.75,.88)
Average (All Tasks)		.62 (.58,.67)	.65 (.61,.69)	.75 (.70,.79)	.70 (.65,.75)	.82 (.78,.85)	.67 (.62,.71)	.74 (.69,.79)	.71 (.66,.75)	.82 (.79,.86)

Table 8.6: Mean test-set macro F1 score (and 95% C.I.) across 5-fold cross validation. We compare BERT Base and Clinical BERT task models with a frozen (❄) and unfrozen (🔥) encoder. We also compare BERT Base and Clinical BERT task models with and without continued pretraining.

the second factor, we compare performance achieved with the encoder parameters frozen and unfrozen. Our evaluation metric is task-level macro F1 score. So as not to overwhelm the reader, we allocate many of the specific details (e.g., cross validation, hyperparameters, model architectures) for this experiment and those that follow to a supplement at the end of this chapter (§8.12).

Results. As shown in Table 8.6, we do not observe any significant difference in performance between task models using BERT Base and Clinical BERT.[6] Clinical BERT models achieve slightly higher average performance than BERT Base models

[6] Our discussion and test statistics are based on average task performance. We use paired t-tests with a significance level of 0.05.

when the encoder is frozen, but actually fall below BERT Base models when the encoder is unfrozen. This trend may suggest that advantages of pretraining with out-of-distribution clinical data are nullified once tuning models to a new clinical data distribution. On average, when the encoder is frozen, neither the Clinical BERT ($t(69)=1.858$, $p=0.067$) nor BERT Base ($t(69)=1.530$, $p=0.131$) task models significantly outperform the majority classifier baseline. In comparison, when the encoder is unfrozen, both Clinical BERT and BERT Base task models significantly outperform the majority classifier baseline and their frozen counterparts ($p<0.001$). Putting these results together, we note that task fine-tuning mitigates issues related to clinical data distribution shift.

8.6.2 Is task fine-tuning sufficient for adapting LMs to a new clinical data distribution?

In low-resource settings, supervised task fine-tuning can be a sub-optimal method of LM transfer due to overfitting (Grießhaber et al., 2020; Tinn et al., 2023). Provided a sufficient corpus of text from a downstream task's domain, continued pretraining of the LM in a self-supervised manner can mitigate this risk (Gururangan et al., 2020; Dery et al., 2021). In this experiment, we ask whether continued pretraining improves

generalization beyond what is achieved via task fine-tuning.

Methods. We use notes from all patients not in the annotated dataset to continue pretraining the BERT Base and Clinical BERT language models. Using each model's respective tokenizer, this amounts to approximately 192M tokens over 1.8M sequences (128 token max length).[7] We train the language model for 16,500 steps using the standard masked language modeling objective (Devlin et al., 2019), allowing for early stopping if validation loss starts increasing, and use the final checkpoint as the encoder in our task models. As before, we compare task performance with the encoder frozen and unfrozen. Additional training details are included in §8.12.5.

Results. Continued pretraining leads to an additional, significant improvement in downstream task performance in each of the four settings considered in §8.6.1 (i.e., {BERT Base, Clinical BERT} × {Frozen Encoder, Unfrozen Encoder}). As shown in the bottom row of Table 8.6, continued pretraining improves average macro F1 score for the BERT Base and Clinical BERT task models with a frozen encoder by 0.05 and 0.04, respectively. The effect of continued pretraining is more pronounced when we unfreeze the encoder, with the BERT Base and Clinical BERT task models achieving an average increase of 0.07 and 0.08 in macro F1 score, respectively, over their counterparts that did not undergo continued pretraining. Overall, our results indicate

[7] BERT Base (Cased) and Clinical BERT use the same vocabulary, but their tokenizers have learned slightly different word-piece splitting criteria.

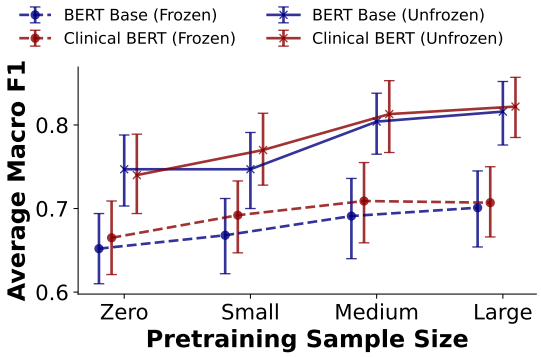


Figure 8.2: Mean task performance (i.e., macro F1 score) as a function of pretraining sample size. Clinical BERT performs slightly better than BERT Base with little to no pretraining.

that both task fine-tuning *and* continued pretraining are critical for maximizing downstream task performance. Furthermore, for our dataset, whether these procedures are applied to an encoder pretrained initially on clinical or non-clinical data does not make a difference in downstream performance.

8.6.3 Are LMs pretrained on clinical data more efficient than LMs trained on non-clinical data in low-data regimes?

We find ourselves in a fortunate position, having access to a non-trivial amount of electronic medical records from our target clinical domain and reasonable compute

infrastructure. If clinical language models don't significantly outperform non-clinical language models in the presence of these resources, then perhaps they require fewer resources to achieve equivalent levels of performance.

Methods. We continue pretraining the BERT Base and Clinical BERT models using the same procedure outlined in §8.6.2. However, we now consider 2 smaller subsets of the pretraining dataset – N=1,024 (`Small`) and N=16,384 (`Medium`). As before, we continue pretraining for a maximum of 16,500 steps and use early stopping to prevent overfitting. We evaluate overall efficiency of the two LMs on the basis of sample efficiency – downstream performance as a function of pretraining dataset size – and compute efficiency – the number of updates required before the stopping criteria is met. We contextualize performance against prior results – no continued pretraining (`Zero`) and pretraining with all available data (`Large`).

Results. Average task performance as a function of the pretraining dataset size is visualized in Figure 8.2 and task-specific performance is included in Table 8.9 of §8.12.7. For 3 out of 4 settings, continued pretraining does not significantly increase downstream performance until utilizing the `Medium` subset. The unfrozen, Clinical BERT task model is the sole exception, with continued pretraining on the `Small` subset causing a statistically significant improvement in performance ($t(69)=2.582$, $p=0.012$). That said, the frozen, Clinical BERT task model also seems to take

advantage of the `Small` pretraining sample slightly better than the frozen, BERT Base model. The frozen, Clinical BERT task model nearly achieves a significant average improvement in macro F1 of 0.02 with the `Small` subset (t(69)=1.954, p=0.055), compared to the frozen, BERT Base task model's improvement in macro F1 of 0.01 with the `Small` subset (t(69)=1.136, p=0.260).

We note that both models approximately match the `Large` (full-dataset) pretraining performance using a fraction of the data (`Medium`). Both models trained on the `Small` subset reach stopping criteria at the same point – 2300 steps – while the Clinical BERT model reaches its stopping criteria 600 steps earlier than BERT Base on the `Medium` subset (6800 vs. 6200). Directionally, the Clinical BERT models appear to be more efficient in low-data regimes than BERT Base models. However, the differences are still relatively small, limiting the strength of conclusions we can draw.

8.6.4 Can we ignore out-of-domain pretraining entirely?

We've seen that continued pretraining is critical to maximizing downstream performance. At the same time, the difference in performance when initializing with a clinical LM instead of a non-clinical LM has been marginal at best. It is natural

to ask whether the initial pretraining phase was even necessary. Can we achieve the same downstream performance when training an LM on our dataset entirely from scratch? Doing so would remove an external dependency that could create non-trivial legal and robustness challenges upon deployment.

Methods. We run the same pretraining procedure as above, but we now initialize the BERT encoder randomly instead of from an external checkpoint. We train one model using the BERT Base tokenizer, and one model using a tokenizer trained on our pretraining dataset. We compare downstream task performance achieved by the two models to each other and to the models initialized from an existing checkpoint.

Results. We report average task performance in Table 8.7 and include task-specific performance in Table 8.10 of §8.12.7. Task models using encoders initialized with random weights prior to continued pretraining achieve roughly equivalent performance as those initialized from existing checkpoints. The domain-specific vocabulary magnifies the efficacy of task fine-tuning, enabling the randomly initialized LM to outperform the LMs pretrained first on other data distributions for some tasks. Once again, we observe that out-of-distribution clinical pretraining provides little-to-no benefit for our target domain.

Initialization	Tokenizer	❄	🔥
BERT Base	BERT Base	.70 (.66,.75)	.82 (.78,.85)
Random	BERT Base	.71 (.66,.75)	.77 (.73,.81)
Random	Learned	.71 (.67,.76)	.81 (.78,.84)

Table 8.7: Mean task performance (i.e., macro F1 score) for each model after pretraining on our ophthalmology dataset. We compare models without (❄) and with task fine-tuning (🔥).

8.7 Review of Learnings

Within each of the experiments above, we find our clinical language models only match (or in some cases fall behind) their non-clinical counterparts. Moreover, we see that both task fine-tuning and continued pretraining using data from our specialized clinical domain are necessary to maximize downstream performance. These results align with and augment existing work that highlights shortcomings of out-of-domain clinical pretraining (Lin et al., 2020; Ranti et al., 2020; Ji et al., 2021; Harrigian et al., 2023b). At the same time, they temper the generality of the claim from Lehman et al. (2023) that language models pretrained on clinical data are necessary for clinical NLP tasks. Our results suggest it is more appropriate to say *domain-specific* clinical language models are still necessary for clinical NLP tasks.

8.8 Limitations

Data. There are four data-related limitations in our study. First, our dataset is drawn from a single, academic hospital system in a mid-sized U.S. city. Our patient population differs from other geographic locations, as does the hospital system's policies, documentation practices, and clinical priorities. Second, although annotations in our dataset were agreed upon by two domain experts, it is still possible that application of our ontology is imperfect. Third, we note that attributes which could not be confidently labeled by annotators were treated as missing data; it is possible that our dataset is biased such that it contains "easier" examples. And fourth, our dataset is relatively small by contemporary standards and exhibits significant class/concept imbalance. Confidence intervals in Table 8.6 reflect uncertainty that arises due to this limitation.

Experimental Design. There are three major limitations with our experimental design to remain cognizant of when interpreting results. First, the concept and class imbalance issues mentioned above necessitated that we consolidate certain ontology attributes and classes. Although this was accomplished with guidance from domain experts, it is possible that alternative groupings would have maintained more (or less) statistical signal. Second, as is common in empirical studies, computational constraints limited the breadth of hyperparameters which were explored to optimize

performance. This is especially pertinent for the pretraining results, in which only a single LM could be trained due to computational expense. Finally, we note that our study only focused on two base LMs which have the same transformer architecture and vocabulary. Alternative language models with different architectures (Lehman et al., 2023) and pretraining data (Lee et al., 2020b; Yang et al., 2022) may have yielded different outcomes.

With respect to the latter, while it is true that BERT language models are comparatively small in the modern language modeling landscape, we argue they still are capable of providing insight regarding the value of out-of-domain clinical pretraining. BERT-style models have outperformed alternative architectures in several clinical NLP tasks and across datasets, even those which are significantly larger (Agrawal et al., 2022; Gutiérrez et al., 2022; Lu et al., 2022; Lehman et al., 2023; Rehana et al., 2023; Labrak et al., 2024). Furthermore, clinical datasets such as ours are generally orders of magnitude smaller than non-clinical datasets and may not even support the training of larger architectures (Spasic, Nenadic, et al., 2020; Touvron et al., 2023; Wornow et al., 2023). Our intention is not to make sweeping claims regarding clinical LMs, but rather to encourage further exploration of clinical LM shortcomings and the heterogeneous nature of clinical language domains.

8.9 Ethical Considerations

Our study involves the analysis of sensitive medical information from real patients. As such, our work is subject to appropriate Health Insurance Portability and Accountability Act (HIPAA) regulations and additional privacy policies set forth by our institution (e.g., compute environment restrictions, patient limits during annotation). We are currently unable to release models due to the risk of personal health information (PHI) leakage (Lukas et al., 2023). All research was approved by our institutional review board (IRB) before its start and adhered to the tenets of the Declarations of Helsinki.

8.10 Discussion

Accurately portraying the limits of clinical language models can have a meaningful impact when designing systems for novel clinical use cases such as our own. Cost, compute, and data privacy constraints already limit the breadth of parameters we can consider when developing a new clinical system. Choosing a sub-optimal pretrained language model as the system's foundation can introduce a performance ceiling before the first experiment is even executed. Notably, this decision is becoming increasingly difficult, with it now common for language models boasting improved performance

along various task axes to be released daily (Alaga and Schuett, 2023; Future of Life Institute, 2023).

Where clinical language models will ultimately reside amongst this deluge of resources remains uncertain. Most state of the art LMs draw performance not only from their complexity (i.e., # parameters), but also their massive training datasets (Chung et al., 2022; Touvron et al., 2023). Even if clinical data becomes easier to obtain, general text corpora from the internet will remain orders of magnitude larger than clinical corpora. Moreover, our experimental results suggest that in the presence of distributional shift, some form of domain adaptation (e.g., continued pretraining) is necessary to maximize an underlying clinical LM's utility, regardless of its pretraining data source.

Together, these observations raise an important and timely question: should we focus on training general-purpose clinical LMs, or extracting performance from larger, non-clinical LMs via domain adaptation? This study has only evaluated two, albeit two widely-used, LMs, and is unable to answer this question. However, various other studies have already demonstrated that large, non-clinical LMs such as GPT-3 contain non-trivial amounts of medical knowledge and can perform well across several biomedical tasks (Agrawal et al., 2022; Nori et al., 2023; Singhal et al., 2023a). What remains to be seen is whether these levels of performance extend beyond the relatively

small and homogeneous pool of clinical NLP benchmarks to novel domains.

8.11 Looking Ahead

In this thesis thus far, we have introduced methods to identify, understand, and mitigate *statistical bias* in clinical and non-clinical data. These contributions have primarily targeted the defensive angle for promoting health equity – i.e., training models that generalize across distributions (e.g., demographics, temporal). In the next chapter, we will shift focus to the proactive angle for promoting health equity – i.e., identifying and combating health disparities directly. How can we address *social bias* using ML and NLP?

8.12 Supplement

8.12.1 Abbreviations

Enumerated in Table 8.8 is a list of useful clinical abbreviations used throughout our study.

Abbreviation	Clinical Concept
DR	Diabetic Retinopathy
PDR	Proliferative Diabetic Retinopathy
NPDR	Non-proliferative Diabetic Retinopathy
HR-PDR	High-Risk PDR
NHR-PDR	Non-High-Risk PDR
NV	Neovascularization
NVD	Neovascularization of the Disc
NVE	Neovascularization of the Retina Elsewhere
AMD	Age-related Macular Degnereration
NVI	Neovascularization of the Iris
ME	Macular Edema
DME	Diabetic Macular Edema
CI-DME	Center Involved DME
CS-DME	Clinically Significant DME
CME	Cystoid Macular Edema
VH	Vitreous Hemorrhage
NVG	Neovascular Glaucoma
RD	Retinal Detachment
TRD	Traction Retinal Detachment
RRD	Rhegmatogenous Retinal Detachment
Anti-VEGF	Anti–vascular Endothelial Growth Factor Therapy
PRP	Panretinal Photocoagulation
DM	Diabetes Mellitus
OS	Oculus Sinister (Left Eye)
OD	Oculus Dexter (Right Eye)
OU	Oculus Uterque (Both Eyes)

Table 8.8: A list of clinical abbreviations used throughout the study.

8.12.2 Stratified Multi-task, Multi-label Cross Validation

We use a variation of stratified cross-validation (see Algorithm 1) for task model experiments, making two modifications to the standard evaluation protocol to minimize information leakage and address concept-attribute class imbalance. First, stratification is done with respect to patients instead of documents (i.e., encounters) or spans. This is done to mitigate the risk of overestimating generalization performance due to copy-forward and other near-duplication that is likely to occur across multiple encounters for the same patient. Pilot experiments confirm that performance estimates

are higher without this splitting condition.

Second, assignments to each fold are made in an iterative fashion using a (potentially) different label criteria during each iteration. This is done to account for the multi-task, multi-label nature of the dataset and address concept-attribute imbalance. Unique combinations of target classes are too sparse to use as stratification, while a purely random stratification approach could lead to some folds not having any instances of certain concepts or attribute classes. Although stratification could be done for each classification task independently, this would preclude us from making direct performance comparisons in the event we trained the task models in a multi-task fashion. Our approach is inspired by prior work in multi-label stratification (Sechidis et al., 2011). We provide an implementation in the supplemental material.

8.12.3 Experimental Setup

All results in this paper are reported using our stratified cross-validation approach, with $K = 5$ folds. Within each fold, 3 patient subsets are used for training, 1 subset is used for parameter tuning and model selection, and 1 subset is used for evaluation. Although the number of unique patients in each subset is roughly equivalent, the number of encounters and concepts spans is not. This is due to the non-uniform concentration of encounters and spans per patient. To limit any single patient from

Input: K # of Folds

Data: g group IDs for each document, Y one-hot labels for each document

Output: s split assignments for each group ID

begin

$g' \leftarrow$ unique group IDs;

$Y' \leftarrow$ group-level one-hot encoding of labels;

$s \leftarrow$ list of length $|g'|$;

$C[K, |\mathcal{L}|] \leftarrow$ Count of group IDs with label $\ell \in \mathcal{L}$ assigned to fold $k \in K$;

while *not all group IDs assigned in* s **do**

$m \leftarrow$ unassigned group ID mask;

$z \leftarrow \sum_{\ell \in \mathcal{L}} Y'[m]$;

$\ell^* \leftarrow \arg\min_{\ell \in \mathcal{L}} z$;

$p \leftarrow$ sort folds $1 \dots K$ ascending based on $C[k, \ell^*]$ with ties broken by $C[k, :]$;

for *fold* i *in folds* $1 \dots K$ **do**

sample $g^* \ni: (g^* \in m) \wedge (Y'[g^*, \ell^*] = 1)$;

assign $s[g^*] \leftarrow i$;

update counts C;

end

end

end

Algorithm 1: Multi-label, Multi-task Stratification

contributing too strongly to the training or evaluation process, we sample a maximum of 10 (Concept, Attribute, Label) tuples from each patient.

8.12.4 Majority Classifier

Let an input X to one of our task classifiers consist of two mutually exclusive groups – 1) a text span T_c which indicates the potential mention of a clinical concept c (e.g., NPDR, Retinal Detachment) and 2) a context window T_z of text surrounding T_c. T_z can be further decomposed as $T_z^{(\text{pre})}$ and $T_z^{(\text{post})}$, such that input X is the ordered concatenation of the components $X = \langle T_z^{(\text{pre})}, T_c, T_z^{(\text{post})} \rangle$.

A majority classifier M for a classification task with K classes outputs the class $k \in K$ which was seen most frequently during training, regardless of input X. In §8.6, we consider a variation of a traditional majority classifier, M' which outputs the class $k \in K$ which was seen most frequently during training amongst training instances associated with the same clinical concept c and token span T_c as the input instance X.

Due to the small, imbalanced nature of our dataset, we cannot assume that all combinations (c, T_c) will have been seen at training time. As such, we adopt the logic enumerated below to back-off to a solution depending on the presence of c and T_c in our training data. There are four cases we must consider.

1. (c seen, T_c seen): M' outputs the most common class label k seen amongst all

training inputs X having c and T_c.

2. (c seen, T_c not seen): M' outputs the most common class label k seen amongst all training inputs X having c.

3. (c not seen, T_c seen): Our expressions are set up such that $T_c \Rightarrow c$. Therefore, this case does not occur in practice. However, M' would output the most common class label k seen amongst all training inputs X having T_c.

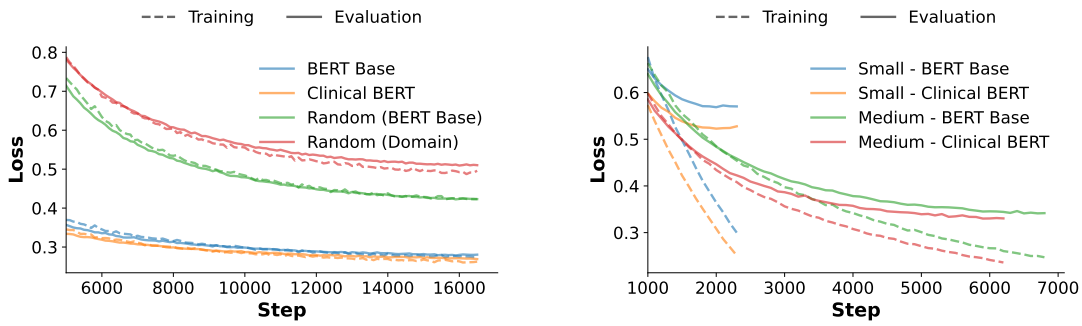
4. (c not seen, T_c not seen): M' outputs the most common class label k seen amongst all training inputs X, regardless of c or T_c. Accordingly, $M' \equiv M$.

The primary purpose of the majority classifier is to provide an appropriate reference for performance given the imbalanced nature of many tasks in the study. That said, the majority model M' conditioned on c and T_c will outperform a simple majority model M without additional conditioning when the extracted text spans can be directly associated with a downstream attribute (e.g., T2DM = Type 2 Diabetes Mellitus, NVI = Neovascularization of the Iris, Trabeculectomy = NVG Surgery Type) or when there are different concept-level differences in the class distribution within a task.

8.12.5 Language Models

We use Hugging Face's implementation of BERT via the `transformers` Python package. Base BERT (Devlin et al., 2019) and Clinical BERT (Alsentzer et al., 2019) are initialized from the Hugging Face hub. Although we use a context window of 128-tokens, we do not re-initialize the positional embeddings.

All patients not included in the annotated dataset are considered candidates for continued pretraining. We use notes from a randomly sampled 5% of this cohort as an evaluation set to monitor convergence, training on notes from the remaining 95% of patients. We continue pretraining of BERT Base and Clinical BERT for a maximum of 16,500 steps using the AdamW optimizer (Loshchilov and Hutter, 2018), an initial learning rate of 5e-5, linear learning rate decay with 5,000 warmup steps, weight decay of 0.01, and an effective batch size of 1,024 (via distributed data parallelism and gradient accumulation). We also apply early stopping based on evaluation loss with a tolerance of 0.01 and patience of 3 (evaluated every 100 steps). The original tokenizer and vocabulary of two BERT models is not modified. We use the masked language modeling objective with a masking probability of 0.15 (up to maximally 20 tokens / sequence) and binary cross entropy loss. All sequences are a maximum 128 tokens long. All other parameters in the `Trainer` and `TrainingArguments` modules of the `transformers` package are kept at their defaults. Loss curves as a function of



(a) Full Dataset **(b)** Downsampled Datasets

Figure 8.3: Training and validation loss curves for continued pretraining on the full (a) and downsampled (b) datasets as a function of initialization strategy and tokenizer. We start the x-axis after a warmup period for visual clarity.

the initialization weights and tokenizer are provided in Figure 8.3a. Loss curves for BERT Base and Clinical BERT pretrained on subsets of the full dataset are provided in Figure 8.3b.

8.12.6 Task Models

The reader should recall two important aspects of the annotated data and concept ontology. First, there may be multiple instances of a clinical concept in a single encounter, not all of which share the same attributes (i.e., laterality, severity, temporality). Second, regular expressions may identify overlapping text spans for *different* clinical concepts. These overlaps may be partial – e.g., [[diabetic] retinopathy] – or full – e.g., [E11.319] is a match for both diabetic retinopathy and diabetes mellitus.

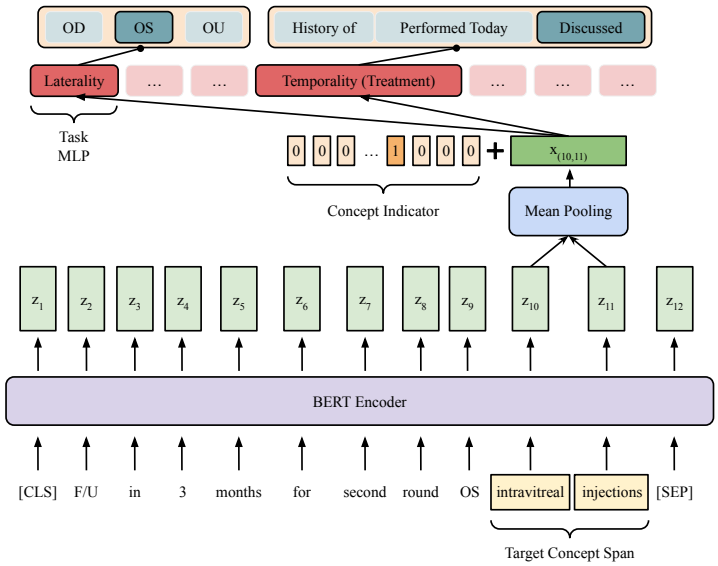


Figure 8.4: Overview of our task model architecture. In practice, we center the context window around each target concept span and train each task model independently.

Together, these aspects inspire us to treat the task of inferring attributes associated with extracted clinical concepts as a span classification problem as opposed to either a document-classification or token-classification problem.

We provide an overview of our model architecture in Figure 8.4. A maximum of 128 tokens centered around the target concept span are passed through the BERT encoder. Centering is accomplished by iteratively expanding the context window on the left and right of the target concept span until the maximum context size or a boundary of the note has been reached. Embeddings for tokens in the target concept span are mean-pooled and then concatenated with a one-hot indicator vector denoting

the clinical concept being classified. The 19-dimensional concept-indicator (one dimension per clinical concept) is included to account for cases of fully-overlapping concept spans for multiple target concepts and explicitly model concept-specific class priors. Finally, the concatenated vectors are passed through a dense, one-layer MLP. The MLP has an input dimensionality of 787 (768 dimension BERT output layer + 19 dimension concept-indicator) and a hidden dimension of either 0 (i.e., linear map to output), 256, or 512 (see discussion below regarding hyperparameter tuning).

All task models are trained independently. That is, a separate backbone LM encoder and MLP classification head are trained for each task. Multi-task learning (i.e., training a shared LM encoder with independent MLP classification heads for each task) is out of this project's scope. We opted to focus on single-task learning because multi-task learning can introduce additional optimization challenges and result in negative task transfer (Lee et al., 2016; Wu et al., 2020). That said, future work which explores the interaction between pretraining domain and the number of fine-tuning tasks is warranted.

Both training and evaluation are implemented in `torch` and leverage language models from the `transformers` Python package. We use the same training setup for all tasks. We minimize the cross entropy loss, inversely weighted based on training class proportions (King and Zeng, 2001). We use the AdamW optimizer and a

step-wise learning rate scheduler configured to reduce the learning rate 10% every 50 steps after a warmup of 100 steps. We hold the weight decay factor (0.1), dropout rate (0.1), and gradient clipping max norm (1.0) constant.

Models are trained for a minimum of 50 steps and a maximum of 500 steps, with early stopping configured to preempt training if the validation loss or macro F1 score have not improved by 1% over 5 evaluation subroutines. We evaluate performance every 5 steps and select the checkpoint which maximizes macro F1 score in the validation set.

We run a hyperparameter grid search over the Cartesian product of learning rates {1e-5, 1e-4, 1e-3}, batch sizes {8, 32, 64}, and MLP hidden dimensions {Null, 256, 512}.[8] The search is run independently for each cross validation fold. The hyperparameter configuration which maximizes macro F1 score in a fold's associated development set is used for evaluation on the test set. Configurations which the training data does not support (i.e., a batch size larger than the number of training instances) are ignored. Higher learning rates were better for task models with frozen BERT encoders, while lower learning rates were better for task models with unfrozen BERT encoders. No other clear trends where observed.

[8] A null hidden dimension denotes a single linear layer from the pooled embeddings to the output layer.

8.12.7 Task-specific Outcomes

We provide task-specific performance breakdowns for the pretraining dataset size (§8.6.3) and weight initialization (§8.6.4) experiments in Tables 8.9 and 8.10, respectively.

		Frozen Encoder (❄)							
		`Zero`		`Small`		`Medium`		`Large`	
Attribute	**Concepts**	**Base**	**Clinical**	**Base**	**Clinical**	**Base**	**Clinical**	**Base**	**Clinical**
Temporality	Retina	.81 (.79,.82)	.83 (.82,.84)	.82 (.80,.83)	.85 (.84,.87)	.84 (.82,.85)	.85 (.84,.87)	.84 (.83,.86)	.85 (.84,.87)
	DM Complications	.81 (.73,.88)	.80 (.70,.89)	.87 (.84,.91)	.86 (.83,.88)	.82 (.71,.89)	.82 (.70,.93)	.81 (.71,.88)	.84 (.73,.93)
	Treatment	.59 (.55,.64)	.69 (.65,.73)	.64 (.57,.71)	.76 (.71,.81)	.78 (.73,.82)	.77 (.74,.79)	.81 (.79,.83)	.81 (.77,.83)
Laterality	All	.54 (.51,.57)	.56 (.54,.58)	.56 (.54,.58)	.56 (.53,.60)	.59 (.56,.62)	.59 (.57,.62)	.60 (.58,.61)	.60 (.57,.63)
Type	ME	.83 (.79,.88)	.82 (.78,.86)	.83 (.80,.86)	.81 (.75,.86)	.88 (.84,.92)	.82 (.79,.86)	.86 (.79,.91)	.85 (.82,.89)
	RD	.75 (.57,.89)	.78 (.59,.94)	.79 (.66,.90)	.79 (.60,.94)	.77 (.64,.88)	.72 (.49,.94)	.70 (.55,.83)	.72 (.50,.95)
	NV	.66 (.53,.80)	.71 (.56,.87)	.79 (.71,.89)	.80 (.66,.93)	.82 (.73,.91)	.85 (.78,.92)	.81 (.71,.91)	.81 (.73,.90)
	DM	.39 (.31,.53)	.31 (.28,.35)	.31 (.30,.32)	.38 (.29,.52)	.39 (.31,.53)	.32 (.28,.37)	.37 (.29,.52)	.33 (.29,.38)
	NVG Surgery	.85 (.69,1.0)	.79 (.54,1.0)	.82 (.62,1.0)	.79 (.61,.96)	.71 (.51,.93)	.79 (.51,1.0)	.74 (.63,.88)	.79 (.63,.96)
	Retina Surgery	.51 (.38,.64)	.59 (.46,.70)	.63 (.45,.78)	.52 (.40,.65)	.50 (.26,.67)	.52 (.37,.66)	.60 (.44,.71)	.64 (.53,.75)
Severity	NPDR	.56 (.50,.62)	.58 (.54,.61)	.55 (.51,.61)	.60 (.55,.65)	.64 (.53,.79)	.74 (.64,.84)	.69 (.57,.84)	.70 (.60,.83)
	PDR	.45 (.43,.48)	.43 (.37,.47)	.38 (.32,.43)	.45 (.42,.47)	.44 (.42,.46)	.53 (.44,.68)	.39 (.29,.47)	.45 (.43,.48)
Span Validity	ME	.60 (.49,.72)	.65 (.52,.78)	.56 (.49,.64)	.74 (.56,.91)	.71 (.55,.87)	.76 (.56,.97)	.77 (.57,.97)	.66 (.49,.86)
	Retina Surgery	.77 (.72,.81)	.76 (.69,.84)	.81 (.72,.89)	.77 (.69,.84)	.80 (.73,.88)	.84 (.78,.89)	.83 (.77,.89)	.83 (.76,.90)
Average (All Tasks)		.65 (.61,.69)	.67 (.62,.71)	.67 (.62,.71)	.69 (.65,.73)	.69 (.64,.74)	.71 (.66,.75)	.70 (.65,.75)	.71 (.66,.75)

		Unfrozen Encoder (🔥)							
		`Zero`		`Small`		`Medium`		`Large`	
Attribute	**Concepts**	**Base**	**Clinical**	**Base**	**Clinical**	**Base**	**Clinical**	**Base**	**Clinical**
Temporality	Retina	.83 (.82,.84)	.84 (.83,.86)	.85 (.83,.87)	.85 (.84,.86)	.87 (.85,.88)	.87 (.85,.88)	.87 (.85,.89)	.87 (.85,.88)
	DM Complications	.80 (.70,.89)	.84 (.76,.90)	.79 (.70,.87)	.86 (.82,.90)	.89 (.85,.93)	.88 (.84,.92)	.80 (.71,.88)	.85 (.77,.92)
	Treatment	.79 (.75,.82)	.81 (.78,.84)	.80 (.76,.84)	.84 (.81,.86)	.83 (.79,.86)	.85 (.81,.87)	.84 (.81,.86)	.82 (.76,.85)
Laterality	All	.84 (.83,.86)	.84 (.81,.87)	.87 (.86,.87)	.87 (.85,.89)	.90 (.89,.92)	.89 (.89,.90)	.92 (.90,.93)	.90 (.89,.92)
Type	ME	.87 (.79,.93)	.87 (.80,.93)	.86 (.80,.93)	.89 (.86,.92)	.90 (.84,.95)	.91 (.89,.94)	.88 (.82,.94)	.90 (.85,.94)
	RD	.79 (.58,.95)	.81 (.61,.98)	.79 (.66,.91)	.83 (.67,.95)	.80 (.69,.90)	.81 (.61,.99)	.82 (.76,.88)	.87 (.78,.95)
	NV	.75 (.60,.89)	.78 (.65,.92)	.77 (.62,.89)	.76 (.63,.89)	.78 (.64,.92)	.78 (.66,.89)	.82 (.72,.93)	.77 (.66,.86)
	DM	.40 (.32,.52)	.40 (.31,.53)	.35 (.32,.38)	.40 (.31,.53)	.59 (.48,.77)	.52 (.37,.76)	.57 (.43,.79)	.54 (.40,.74)
	NVG Surgery	.85 (.69,1.0)	.85 (.69,1.0)	.85 (.69,1.0)	.85 (.69,1.0)	.85 (.69,1.0)	.85 (.69,1.0)	.85 (.69,1.0)	.85 (.69,1.0)
	Retina Surgery	.66 (.52,.79)	.52 (.46,.59)	.62 (.50,.73)	.61 (.53,.68)	.70 (.58,.79)	.70 (.63,.75)	.71 (.59,.83)	.76 (.65,.85)
Severity	NPDR	.83 (.73,.90)	.91 (.87,.96)	.92 (.86,.98)	.91 (.85,.96)	.91 (.86,.97)	.96 (.90,1.0)	.89 (.77,.98)	.95 (.90,.99)
	PDR	.71 (.53,.89)	.53 (.36,.77)	.67 (.45,.90)	.69 (.51,.86)	.73 (.55,.90)	.69 (.47,.90)	.81 (.64,.93)	.82 (.63,.98)
Span Validity	ME	.55 (.49,.62)	.56 (.50,.64)	.57 (.49,.73)	.63 (.50,.78)	.73 (.60,.88)	.82 (.65,.97)	.81 (.64,.97)	.83 (.65,.98)
	Retina Surgery	.77 (.68,.87)	.80 (.74,.86)	.75 (.68,.83)	.81 (.76,.86)	.76 (.68,.83)	.87 (.82,.90)	.82 (.75,.90)	.80 (.75,.88)
Average (All Tasks)		.75 (.70,.79)	.74 (.69,.79)	.75 (.70,.79)	.77 (.73,.81)	.80 (.76,.84)	.81 (.77,.85)	.82 (.78,.85)	.82 (.79,.86)

Table 8.9: Task-specific performance (i.e., macro F1 score) as a function of the pretraining dataset size. Performance with a frozen encoder and an unfrozen encoder is shown in the top and bottom tables, respectively. We observe gradual increases in performance for both BERT Base and Clinical BERT task models as the pretraining dataset grows. The Clinical BERT model is able to take advantage of the `Small` pretraining dataset slightly better than BERT Base.

		Frozen Encoder (❄)			Unfrozen Encoder (🔥)		
		BERT Base	Random		BERT Base	Random	
Concept	**Attribute**	**BERT Base**	**BERT Base**	**Learned**	**BERT Base**	**BERT Base**	**Learned**
Temporality	Retina	.84 (.83,.86)	.83 (.82,.84)	.86 (.84,.87)	.87 (.85,.89)	.85 (.82,.87)	.85 (.83,.87)
	DM Complications	.81 (.71,.88)	.79 (.65,.89)	.80 (.75,.84)	.80 (.71,.88)	.83 (.73,.92)	.90 (.87,.93)
	Treatment	.81 (.79,.83)	.72 (.70,.74)	.80 (.77,.83)	.84 (.81,.86)	.80 (.77,.84)	.79 (.72,.85)
Laterality	All	.60 (.58,.61)	.58 (.56,.60)	.61 (.60,.62)	.92 (.90,.93)	.87 (.85,.89)	.89 (.87,.91)
Type	ME	.86 (.79,.91)	.87 (.82,.91)	.89 (.85,.93)	.88 (.82,.94)	.89 (.85,.92)	.90 (.87,.94)
	RD	.70 (.55,.83)	.80 (.60,.97)	.84 (.77,.91)	.82 (.76,.88)	.76 (.58,.92)	.74 (.65,.83)
	NV	.81 (.71,.91)	.83 (.74,.93)	.81 (.74,.89)	.82 (.72,.93)	.79 (.71,.87)	.75 (.69,.83)
	DM	.37 (.29,.52)	.42 (.32,.58)	.40 (.31,.54)	.57 (.43,.79)	.40 (.29,.54)	.53 (.38,.75)
	NVG Surgery	.74 (.63,.88)	.72 (.43,.97)	.75 (.58,.93)	.85 (.69,1.0)	.85 (.69,1.0)	.85 (.69,1.0)
	Retina Surgery	.60 (.44,.71)	.67 (.53,.80)	.53 (.42,.63)	.71 (.59,.83)	.71 (.63,.79)	.75 (.63,.87)
Severity	NPDR	.69 (.57,.84)	.73 (.65,.84)	.67 (.46,.86)	.89 (.77,.98)	.84 (.73,.91)	.91 (.88,.94)
	PDR	.39 (.29,.47)	.43 (.35,.48)	.48 (.47,.48)	.81 (.64,.93)	.66 (.43,.90)	.89 (.79,.98)
Span Validity	ME	.77 (.57,.97)	.74 (.54,.94)	.72 (.53,.90)	.81 (.64,.97)	.75 (.60,.90)	.82 (.66,.95)
	Retina Surgery	.83 (.77,.89)	.77 (.59,.89)	.83 (.77,.90)	.82 (.75,.90)	.84 (.81,.88)	.81 (.74,.87)
Average (All Tasks)		.70 (.65,.75)	.71 (.66,.75)	.71 (.67,.76)	.82 (.78,.85)	.77 (.73,.81)	.81 (.78,.84)

Table 8.10: A comparison of task-specific performance (i.e., macro F1 score) when pretraining from scratch instead of continuing pretraining from an existing checkpoint. With the domain-specific (learned) vocabulary, we are able to achieve the same level of performance when pretraining from scratch as we do when pretraining from the existing BERT Base checkpoint.

Part IV: Using Machine Learning to Proactively Address Health Disparities

Chapter 9

Characterizing Stigmatizing Language in Medical Records

9.1 Overview

Our focus up to this point in the thesis has been primarily on measuring and addressing statistical bias that arises in training data for health-related applications. The objective of this effort has been to mitigate the risk of introducing or re-introducing health disparities via models that have different levels of quality for different populations (e.g., demographic groups). What do we do about the health disparities that already exist? How can we take a more proactive approach to promoting health equity? In this chapter, we offer an answer to these important questions.

We begin by formalizing a new natural language processing task – characterizing stigmatizing language in electronic medical records.[1] We then theoretically and empirically embed this task within the existing harmful language research domain, and introduce two new datasets to facilitate emerging research in the area. We leverage several analytical techniques to measure and interpret statistical biases intrinsic to these datasets and, in turn, provide additional evidence of the heterogeneity in clinical language. Finally, we apply our system to a public clinical dataset and demonstrate that it not only contains significant usage of stigmatizing language, but also that this

[1] The detection and understanding of social bias (e.g., racism, sexism, stigmatization) in language has become a major area of focus in the NLP community over the past several years (Breitfeller et al., 2019; Ali et al., 2020; Banko et al., 2020). Nonetheless, most of this research has focused on language found online (e.g., social media). It is not clear *a priori* which principles from this research are relevant for the clinical domain.

language usage differs across racial groups. Broadly, our study yields insights that have implications both on how healthcare providers interact with patients, and how ML and NLP practitioners use existing clinical data to train language models.

9.2 Background

Widespread and well-documented disparities in healthcare outcomes between demographic groups exist within the United States (Baciu et al., 2017b; Zavala et al., 2021). The sources of these disparities are diverse and complex, with numerous interacting factors contributing to worse outcomes for minority patients (Bell and Lee, 2011; Williams et al., 2019). One source of disparities may stem from latent biases of healthcare providers (Hall et al., 2015). Multiple studies have highlighted the tendency for providers to prescribe different treatment plans to Black patients compared to White patients despite having similar clinical dispositions (Nelson, 2002; Green et al., 2007; Hoffman et al., 2016). Elevated implicit bias scores have been associated with these decisions and have been further linked with decreased levels of patient-provider communication (Van Ryn et al., 2011; Cooper et al., 2012). A major challenge with these biases is that they are invoked unconsciously.

A new line of work in medical sociology has explored this issue through the lens of clinical documentation (Beach et al., 2021), in which bias may be exhibited in how

medical providers describe and document patient interactions in the medical record. In particular, studies have shown physicians often use language that has subtle, stigmatizing connotations (Wolsiefer et al., 2021). This documentation practice may not only negatively frame patients to future providers and thus influence their quality of care, but also discourage patients from seeking treatment altogether (Goddu et al., 2018; Werder et al., 2022). The latter is especially pertinent given the passage of the 21st Century Cures Act that mandates clinical notes are freely accessible by patients in the US (Blease et al., 2021; Harris et al., 2022).

9.3 Motivation and Contribution

How is stigmatizing language in medical records different from other forms of abusive language? Prior studies of stigmatizing language in clinical notes have relied on qualitative methods (Park et al., 2021a) or refrained from analyzing computational nuances of the problem domain (Sun et al., 2022). Modeling tasks such as hate-speech detection (Garg et al., 2022; Jahan and Oussalah, 2023) and analyses of social bias encoded within language models (Liang et al., 2021) share many similarities with characterizing stigmatizing language in medical records. However, it is not clear *a priori* where the task of characterizing stigmatizing language in medical records falls within the broader abusive language landscape.

In this chapter, we demonstrate that characterization of stigmatizing language in medical records most strongly parallels the characterization of linguistic microaggressions (Sue et al., 2007). However, unlike traditional microaggressions, harmful and biased language in the clinical domain is concentrated in unremarkable phrases and lacks any indication of the targeted identity group. Our analysis establishes a foundation for a novel task that has high importance to both patients and clinicians. We demonstrate this value by applying our models to a public clinical dataset and showing that stigmatizing language is used disproportionately for certain demographic groups.

9.4 Grounding Clinical Stigmatizing Language

9.4.1 Harmful Language Taxonomy

We use the typology of abusive and harmful language introduced by Waseem et al. (2017) to contextualize clinical stigmatizing language in the broader research landscape. The typology consists of two axes – 1) whether harm is explicit or implicit, and 2) whether harm is directed at a specific target. We claim that clinical stigmatizing

language lies in the *implicit* and *directed* quadrant of the typology.

Explicit vs. Implicit. With respect to the first dimension, we note that physicians (generally) do not use racial slurs or other forms of language which have unambiguous negative connotations. Instead, they use a consistent vocabulary of terms and phrases which has negative implications when interpreted in certain contexts or by other physicians (Beach et al., 2021; Valdez, 2021). For instance, if the physician describes the patient as non-compliant and that they decline to take medications, this might affect their future treatment plan. Here, the necessity of context is what distinguishes stigmatizing language in medical records as *implicit* instead of explicit according to the abusive language typology. In §9.7.1, we demonstrate that neither the ambiguous "anchor" terms or their context are sufficient in isolation to characterize stigmatizing language in the clinical domain.

Directed vs. Undirected Because the primary purpose of clinical documentation remains facilitating provider-to-provider communication, stigmatizing comments are rarely written to the patient as the intended recipient.[2] That said, a single patient is still the subject of the stigmatizing language as opposed to a general population. In fact, we show in §9.7.2 that semantic representations of stigmatizing language in the clinical domain do not encode any information about latent demographic groups.

[2] Not including e.g., patient instructions written prior to discharge events.

Thus, by the definition proposed in Waseem et al. (2017), stigmatizing language in clinical records tends to be a form of *directed* abuse.

9.4.2 Broader Connections

Considering the harmful language landscape more broadly, the characteristics of stigmatizing language in medical records are quite similar to those found in linguistic microaggressions. In both cases, the language typically reflects an unconscious bias internalized by the speaker and materialized through thinly veiled innuendo (Sue et al., 2007; Raney et al., 2021). This innuendo is not necessarily confined to being negative in affect. Positive language, either used disproportionately across demographics or with an underlying implication of inferiority, can also inflict harm (Glick and Fiske, 2001; McMahon and Kahn, 2016). This subtlety is part of what makes characterizing this type of stigmatizing language difficult.

One major difference between stigmatizing language in the clinical domain and other forms of harmful language is the notion of *necessity*. With the latter, it's not unreasonable to say the harmful language would have been better left unsaid altogether; the proverb, "If you don't have anything nice to say, don't say anything at all" comes to mind. Clinicians, on the other hand, do not usually have this luxury. Clinicians have a responsibility to comprehensively document their interaction

with patients (Shanley et al., 2009). Often, this requires that they describe and interpret socially-stigmatized circumstances (e.g., adolescent pregnancy, substance use disorders) and medically-relevant patient eccentricities (e.g., reluctance to accept medical advice, unfounded social histories). As a consequence, minor differences in phrasing and word choice have a large impact on whether a statement is stigmatizing to patients.

9.5 Related Work

Very few studies of stigmatizing language in medical records exist to date. A qualitative review by Goddu et al. (2018) was first to provide evidence of negative language in the medical record. Beach et al. (2021) and Himmelstein et al. (2022) later analyzed the frequency of words having a stigmatizing connotation within a large sample of notes. Their respective studies revealed an elevated prevalence of stigmatizing language within records of Black patients than White patients at two different institutions.

Recent work from Sun et al. (2022) was the first to use machine learning to analyze stigmatizing language in medical records. The authors used a manually-curated word list to identify sentences with possible bias and then annotated whether each of the candidate matches was positive, negative, or taken out-of-context. They then

trained a logistic regression classifier provided with a bag-of-words representation of the sentence to predict the polarity/relevance of each match. Although Sun et al. (2022) boast their model achieving a macro F1 score of 0.935, they do not provide any details regarding the distribution of seed terms in their dataset, nor do they provide a reference baseline to indicate how valuable context around the seed terms is for classification. Indeed, we will show in §9.7.1 that a naive baseline model using only the seed terms as features can achieve strong discriminatory accuracy. We also expand upon their annotation taxonomy to better capture different types of stigmatizing language in the medical domain.

The broader task of identifying biased and abusive language has garnered much attention from the NLP community in recent years (Schmidt and Wiegand, 2017; Yin and Zubiaga, 2021; Jahan and Oussalah, 2023). Breitfeller et al. (2019) was the first to computationally analyze microaggressions, showcasing the importance of understanding context when characterizing covert forms of linguistic bias. The majority of research built on this foundation has remained confined to using web and social media data (Wang and Potts, 2019; Lees et al., 2021; Sabri et al., 2021), with recent work from Washington et al. (2021) being the first to consider multi-modal alternatives. Our study fills an important void in the existing landscape of covert bias research, providing an in-depth analysis of stigma in a linguistic domain that

differs dramatically from others studied in the space.

9.6 Data

We consider two clinical datasets. In addition to covering different clinical specialties, they also feature different demographic compositions.

9.6.1 Data Sources

JHM. We retrospectively acquired a dataset of 128,343 English-language progress notes written by physicians across 5 clinical specialties within the Johns Hopkins Medicine (JHM) hospital system — Internal Medicine, Emergency Medicine, Pediatrics, OB-GYN, and Surgery. Notes were processed in accordance with our institution's privacy policy after approval by our Institutional Review Board (IRB).

MIMIC. To encourage future research, we also include in our study the publicly-accessible MIMIC-IV-Note dataset (v2.2) (Johnson et al., 2023). This recently released extension of the widely-adopted MIMIC-III dataset (Johnson et al., 2016) consists of deidentified free-text clinical notes for patients admitted to an intensive care unit (ICU) or the emergency department at Beth Israel Deaconess Medical Center in Boston, MA. We focus on the 331,794 available discharge summaries, having found

minimal evidence of stigmatizing language in the associated radiology reports.

9.6.2 Preprocessing

All clinical free text in our datasets was case-normalized and converted to an ASCII encoding prior to additional processing. The MIMIC dataset was de-identified before we obtained access to it. The JHM dataset, however, was not subject to any de-identification procedures because it is protected within a secure cloud environment and we are not distributing assets derived from it.

9.6.3 Task Taxonomy

Like Sun et al. (2022), we develop a two-stage process to detect and characterize stigmatizing language in clinical notes. Possible instances of bias are first identified using "anchor" n-grams and then classified using a machine learning classifier. Unlike the single, sentiment-like classification task considered by Sun et al. (2022), we formulate three independent classification tasks that discriminate between instances of bias based on impact. The taxonomy was developed by clinicians on our team, drawing upon previous literature (Beach et al., 2021; Sun et al., 2022). We present the task taxonomy developed for this study in Table 9.1, along with de-identified examples for each of the stigmatizing language classes.

Stigma Type	Class	Definition	Examples
Credibility & Obstinacy	Disbelief	Insinuates doubt about a patient's stated testimony.	*adamant* he doesn't smoke; *claims* to see a therapist
	Difficult	Describes patient (or patient's family) perspective as inflexible/difficult/entrenched, typically with respect to their intentions.	*insists* on being admitted; *adamantly* opposed to limiting fruit intake
	Exclude	Word/phrase is not used to characterize the patient or describe the patient's behavior; may refer to medical condition or treatment or to another person or context.	patient's friend *insisted* she go to the hospital; test *claims* submitted to insurance
Compliance	Negative	Patient not, unlikely to, or questionably following medical advice.	*adherence* to therapeutic medication is unclear; mother *declines* vaccines; struggles with medication and follow-up *compliance*
	Neutral	Not used to describe whether the patient is not following medical advice or rejecting treatment; often used to describe generically some future plan involving a hypothetical. Alternatively, see Exclude (Credibility & Obstinacy).	discussed the medication *compliance*; school *refuses* to provide adequate accommodations; feels that her parents' health has *declined*
	Positive	Patient following medical advice.	continues to be *compliant* with aspirin regimen; reports excellent *adherence*
Descriptors	Negative	Patient's demeanor or behavior is cast in a negative light; insinuates the patients is not being forthright or transparent; patient may be falsifying symptoms to get something they want.	*drug-seeking* behavior; concern for *secondary gain*; *unwilling* to meet with case manager; unfortunately a poor *historian*
	Neutral	Negation of negative descriptors; insinuates the patient was expected to have a negative demeanor or be difficult to interact with.	his mother is the primary *historian*; interactive and *cooperative*; not *combative* or *belligerent*; dad seems *angry* with patient at times
	Positive	Patient's demeanor or behavior is described in a positive light; patient is easy to interact with.	*lovely* 80 year old woman; *well-groomed* and holds good eye contact; *pleasant* and appropriate interaction with staff
	Exclude	Patient self-description or description of another individual. Alternatively, see Exclude (Credibility & Obstinacy).	does not want providers to think she's *malingering*; reports feeling *angry* before her period; lives on *pleasant* avenue downtown

Table 9.1: Taxonomy of stigmatizing language. Complete anchor sets for each task can be found in Figure 9.2. Annotators were provided a comprehensive guide with general examples and edge cases for each anchor n-gram in our taxonomy.

9.6.4 Anchor List

We take the union of n-grams curated by Beach et al. (2021) and Sun et al. (2022) as our anchor set. Both studies leveraged a time-intensive qualitative review of thousands of clinical notes by clinicians and medical ethicists to arrive at their respective lists. Future work may consider using context-aware keyword discovery (Naseri et al., 2021; Zheng et al., 2021) to expand the current anchor n-gram list.

9.6.5 Annotation

We ran our anchor list against both datasets, caching each match and up to 10 words to the left and right which make up its context. We used Python's `re` package to identify the anchors. The 10-word context size was specified *a priori* based on guidance from our clinical collaborators.[3]

We randomly sampled 5,201 and 5,043 of the identified instances from the JHM and MIMIC datasets, respectively, to be annotated. Three annotators – one clinician C1 and two research assistants R1, R2 – were responsible for labeling all data used in our study. We present agreement matrices in Figure 9.1 for the MIMIC and JHM datasets. Each instance in the JHM dataset was labeled by at least two annotators, with a subset labeled by three. A subset of instances in the MIMIC dataset were labeled by two annotators, with the remainder labeled by a single annotator. Annotators labeled the data independently and then met with the larger team to resolve disagreements and discuss ambiguous cases.

Agreement scores prior to resolution were quite high, suggesting 1) the annotation taxonomy was clear and 2) the stigmatizing language we considered was generally not ambiguous in its impact. We observed similar agreement trends for both datasets; the Descriptors task had the highest agreement, while the Credibility & Obstinacy

[3] Future work may consider evaluating the effect this choice has on annotation and modeling outcomes.

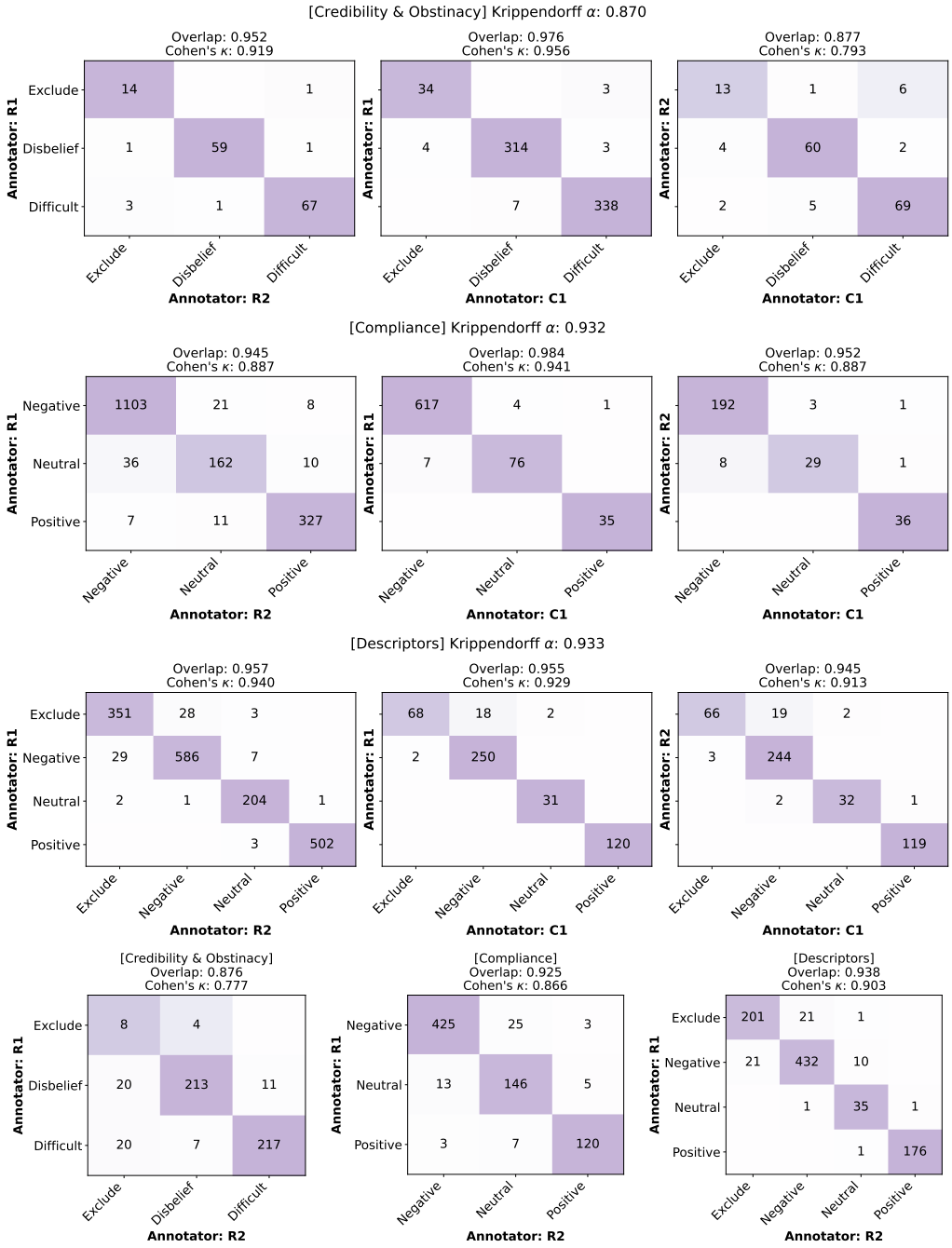


Figure 9.1: Pairwise interannotator agreement for the JHM dataset (first 3 rows) and MIMIC dataset (last row).

Task	Class	JHM	MIMIC
Credibility & Obstinacy	Difficult	413	526
	Disbelief	438	609
	Exclude	77	115
Compliance	Negative	1,578	893
	Neutral	283	439
	Positive	357	271
Descriptors	Exclude	430	496
	Negative	843	1,221
	Neutral	233	96
	Positive	549	377

Table 9.2: Resolved label distribution for each task.

task had the lowest agreement. The former consists of several highly polar anchor n-grams (e.g., pleasantly, unkempt), while the latter requires a higher degree of personal interpretation.

9.6.6 Sample Statistics

We provide the distribution of labels for each task in Table 9.2. This distribution is further broken down by anchor n-gram in Figure 9.2. Each task contains a subset of anchors with extreme class imbalance.

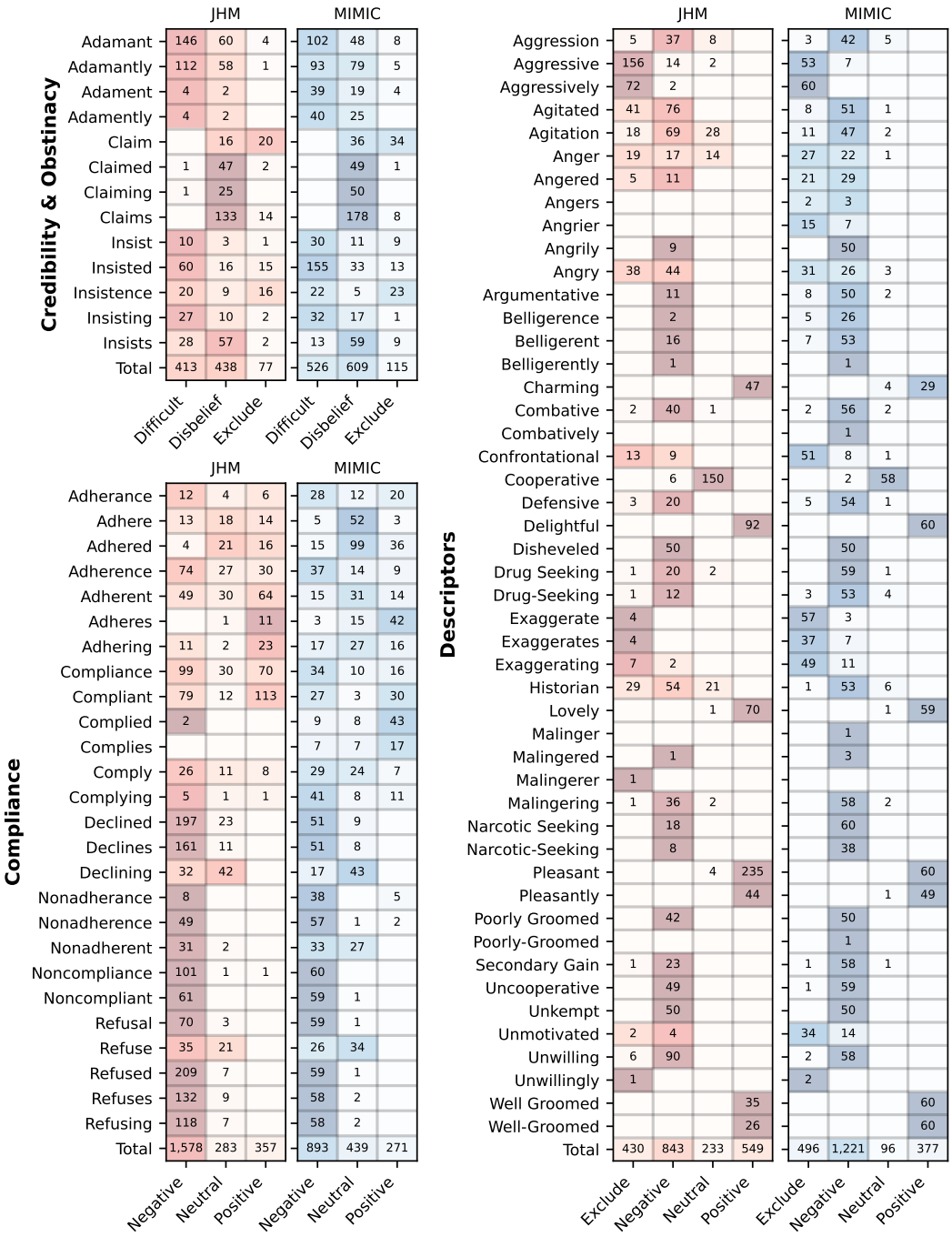


Figure 9.2: Joint anchor and label distribution for each task.

9.7 Modeling Stigmatizing Language

9.7.1 What role does context play in characterizing stigmatizing language?

Some forms of harmful language are stigmatizing in isolation, while others critically depend on context to invoke meaning (Waseem et al., 2017). Prior work has not provided insight regarding where stigmatizing language in medical records lies on this spectrum (Sun et al., 2022). We hypothesize that context around a stigmatizing instance is necessary, but insufficient, for characterizing the utterance.

9.7.1.1 Methods

We test our hypothesis by varying feature representations such that they encode different degrees of the stigmatizing anchor term and its surrounding context.

Models. We consider 3 classes of models. The first two classes allow us to understand the interaction between context and the anchor n-grams in an additive manner. The third class captures more complex dynamics between anchor n-grams and their context.

Majority. 2 naive majority classifiers – one that has knowledge of the anchor

(Majority Per Anchor) and one that does not (Majority). The Majority model outputs the most common class seen at training time. The Majority Per Anchor model outputs the following class probabilities given an input anchor n-gram w:

$$p(y \mid w) = \frac{C(w, y) + \alpha}{\sum_{y' \in \mathcal{Y}} C(w, y') + |\mathcal{Y}|\alpha}$$

where $C(w, y)$ is the number of examples with anchor w having class y in the training data, $\mathcal{Y}$ is the set of possible classes y, and α is a smoothing hyperparameter. We use $\alpha = 1$ for all of our experiments.

Logistic Regression (LR). 2 models, both leveraging TF-IDF feature representations in line with Sun et al. (2022). One model includes the anchor n-gram, while the other model excludes the anchor n-gram. We use a custom pipeline to transform the raw text into feature space. Instances are first tokenized using a clinical domain tokenizer implemented in the medspaCy library (Eyre et al., 2021). Tokens are recursively merged together to form phrases based on the bi-gram scoring function introduced by Mikolov et al. (2013b) and implemented in Gensim (Řehůřek and Sojka, 2010). We use a scoring threshold of 10, minimum vocabulary frequency of 5, and recurse twice to identify 1-4 grams. We use scikit-learn (Pedregosa et al., 2011) for TF-IDF data transformations and classifier training. For the TF-IDF representations, we use an ℓ_2 row-wise norm. As a classifier, we use multinomial logistic regression optimized using

lbfgs (Zhu et al., 1997). We balance class weights and perform a grid search over the following ℓ_2 regularization parameters: {0.01, 0.03, 0.1, 0.3, 1, 3, 5, 10}. The model which maximizes macro F1 score in each training split's associated development set is chosen for application on the test set.

BERT. 2 models – one version trained on web data (Devlin et al., 2019) and one version trained on clinical notes (Alsentzer et al., 2019). We use Hugging Face's transformers library (Wolf et al., 2019) to initialize all BERT models and fine-tune them using code written in PyTorch (Paszke et al., 2019). We train all models using a batch size of 16, a fixed learning rate of 5e-05, a dropout probability of 0.1, and class-balanced cross-entropy loss. As an optimizer, we use AdamW (Loshchilov and Hutter, 2018). We evaluate the model every 50 updates and save the model which maximizes macro F1 score on the training split's associated development data. Due to compute limitations in our HIPAA-compliant environment (i.e., limited GPU access), we do an initial exploration of the ℓ_2 regularization strength on one split of the data for each classification task. We find the regularization strength to have minimal effect on performance for decay values of {1e-5, 1e-4, and 1e-3}; we set a decay weight of 1e-5 for all remaining experiments.

Pooling. We also compare four methods of pooling BERT's final hidden layer for

input into the task classification head.

Anchor Mean: Arithmetic mean of tokens (subwords) composing the anchor n-gram.

CLS: Embedding for the classification token.

Sentence Mean: Arithmetic mean of all tokens in the instance, excluding special tokens.

BERT Pooler: Weighted pooling of all tokens; weights learned at training time.

Experimental Design. The annotated dataset is split into training, development, and test subsets at a 70/20/10 ratio. Stigmatizing language instances are assigned randomly into each subset, using their associated patient identifiers as stratification criteria to limit data leakage. The training and development subsets are further split at random to facilitate 5-fold cross-validation. Readers should keep in mind that the clinical BERT models (Alsentzer et al., 2019) were pretrained on MIMIC-III (Johnson et al., 2016), which may have a small amount of note and/or patient overlap with our MIMIC-IV discharge summary sample.

	Credibility & Obstinacy		Compliance		Descriptors	
Model	**JHM**	**MIMIC**	**JHM**	**MIMIC**	**JHM**	**MIMIC**
Majority Overall	0.21 ± 0.00	0.17 ± 0.00	0.29 ± 0.00	0.24 ± 0.00	0.16 ± 0.00	0.19 ± 0.00
Majority Per Anchor	0.67 ± 0.10	0.55 ± 0.04	0.68 ± 0.04	0.73 ± 0.01	0.82 ± 0.01	0.83 ± 0.00
LR (Context)	0.60 ± 0.05	0.58 ± 0.04	0.55 ± 0.01	0.68 ± 0.02	0.74 ± 0.03	0.60 ± 0.04
LR (Context + Anchor)	0.69 ± 0.02	0.65 ± 0.03	0.68 ± 0.04	0.80 ± 0.02	0.86 ± 0.02	0.76 ± 0.05
Bert (Web)	0.85 ± 0.04	0.76 ± 0.02	**0.86 ± 0.01**	**0.92 ± 0.02**	**0.93 ± 0.01**	**0.86 ± 0.01**
Bert (Clinical)	**0.89 ± 0.03**	**0.78 ± 0.03**	0.85 ± 0.02	**0.92 ± 0.02**	**0.93 ± 0.02**	**0.86 ± 0.01**
– CLS Token	0.89 ± 0.04	0.69 ± 0.03	0.84 ± 0.03	0.92 ± 0.01	0.90 ± 0.01	0.84 ± 0.03
– Sentence Mean	0.85 ± 0.06	0.69 ± 0.06	0.84 ± 0.03	0.92 ± 0.01	0.91 ± 0.01	0.84 ± 0.02
– BERT Pooler	0.83 ± 0.08	0.70 ± 0.07	0.84 ± 0.02	0.91 ± 0.02	0.89 ± 0.03	0.80 ± 0.03

Table 9.3: Test macro F1 score ($\mu \pm \sigma$) for each classification task. Underlining indicates a pooling method is significantly worse than anchor mean pooling (paired t-test $p < .05$). The best model(s) for each classification task are bolded.

9.7.1.2 Results

The final four rows in Table 9.3 show clinical BERT's test-set macro F1 score for each pooling method across the three classification tasks; the web version of BERT performs similarly. Although not always statistically significant, the anchored pooling method consistently outperforms the alternative pooling approaches across all tasks and datasets. Under this setting, the classification head lacks direct access to information in each anchor's context window. Classification performance can be thought of as a measure of how well the closed set of anchor n-grams are separated in semantic space. That the anchor pooling approach outperforms the alternative methods suggests characterizing stigmatizing language in medical records can be thought of as a word-sense-disambiguation task more than a sequence classification task.

The majority and logistic regression model outcomes (first four rows of Table 9.3)

lend additional support to this claim. We see that anchors used as classification criteria in isolation provide a significant improvement over the majority overall model in all cases. The context window used in isolation provides a relatively smaller increase in performance over the majority overall model. Jointly modeling the anchors and their context achieves the largest improvement over the majority overall model in 4 of 6 tasks. This outcome suggests that both subsets of text provide different, but complementary, information.

The BERT models effectively capture the interaction between anchors and their surrounding context. Fine-tuning both BERT models significantly increases macro F1 over the best non-BERT model in all settings. Interestingly, the difference in performance between the web and clinical BERT models is not consistently significant, even for the MIMIC-IV dataset that potentially has overlap with the clinical BERT pretraining data. We hypothesize that understanding social bias may be more important than understanding clinical jargon for our tasks, but leave this as an open question for future work. It may also just be the case that, as seen in Harrigian et al. (2023a), clinical language models don't offer much of an advantage in the presence of any distributional shift.

9.7.2 Is stigma conveyed in the same manner about different demographic groups?

The majority of bias-related tasks in NLP examine language which, while covert, contains some indication of the targeted demographic of identity group (e.g., racial slurs, sexist microaggressions) (Sue, 2010; Waseem et al., 2017). Here, we show that stigmatizing language in medical records uniquely *does not* target any racial group or sex.

9.7.2.1 Methods

Results from §9.7.1 verify that our BERT encoders learn semantic representations of the anchor n-grams that are informative for the downstream stigma characterization tasks. If language is used differently for different demographic groups, we expect the encoders to reflect this (Adam et al., 2022). We can test our hypothesis by attempting to infer a patient's self-reported race and sex using each anchor n-gram's BERT representation.

Experimental Design. We train new clinical BERT models for each of the three classification tasks. This time, we forego cross-validation and instead use a single training, development, and test split. We detach each task's classification head

and pass the anchor n-grams through their respective models to extract their internal mean-pooled representation.

Maintaining separation between the three classification tasks, we randomly split the subset of patients whose data was used for training the BERT models into 5 non-overlapping groups and use these groups as folds for cross-validation. Using 4 of the groups for training, an unregularized logistic regression classifier is fit to independently predict race and sex from the internal semantic representations. We evaluate separation using data from the held-out group. This process is repeated 5 times until each patient group has been used as the held-out test group.

The joint race and sex distribution of instances is provided in Table 9.4. We ignore instances in which a patient either declined to report or did not self-report their race or sex. After this exclusion, we are left with 5,129 of the original 5,201 instances for the JHM dataset, and 4,875 of the original 5,043 instances for the MIMIC dataset.

Reference Performance. Our clinical datasets represent a concatenation of notes from different specialties. Each speciality has a unique patient demographic pool and thus invites the possibility of conflating the encoding of specialty-specific knowledge with demographic-specific knowledge. For example, OB-GYN notes come specifically from female patients and our sample of JHM pediatric notes come from a population which is 95% Black. Encoding the speciality would naturally allow inference of patient

JHM		Black or African American		White or Caucasian		Other	
Task	**Class**	**Female**	**Male**	**Female**	**Male**	**Female**	**Male**
Credibility & Obstinacy	Difficult	159 (129)	94 (76)	87 (60)	40 (29)	11 (10)	17 (13)
	Disbelief	160 (133)	142 (117)	59 (46)	47 (39)	8 (7)	18 (17)
	Exclude	20 (18)	20 (17)	20 (13)	11 (9)	3 (3)	3 (2)
Compliance	Negative	714 (499)	480 (324)	187 (146)	104 (81)	22 (20)	41 (29)
	Neutral	107 (102)	87 (81)	43 (37)	31 (29)	4 (4)	4 (3)
	Positive	146 (135)	105 (93)	50 (45)	35 (31)	9 (7)	6 (5)
Descriptors	Exclude	146 (132)	132 (108)	68 (55)	58 (56)	8 (8)	11 (9)
	Negative	253 (172)	254 (189)	134 (72)	144 (89)	17 (11)	34 (18)
	Neutral	78 (69)	51 (50)	54 (52)	32 (29)	6 (5)	8 (8)
	Positive	232 (185)	117 (98)	111 (91)	59 (48)	19 (16)	9 (9)

MIMIC		Black or African American		White or Caucasian		Other	
Task	**Class**	**Female**	**Male**	**Female**	**Male**	**Female**	**Male**
Credibility & Obstinacy	Difficult	35 (32)	48 (47)	177 (167)	189 (177)	31 (29)	32 (31)
	Disbelief	64 (64)	56 (55)	209 (198)	191 (179)	31 (30)	41 (41)
	Exclude	13 (13)	8 (8)	36 (36)	43 (43)	7 (6)	7 (7)
Compliance	Negative	127 (121)	109 (93)	232 (219)	277 (258)	64 (61)	56 (54)
	Neutral	30 (30)	26 (25)	146 (140)	160 (157)	23 (23)	35 (33)
	Positive	23 (23)	23 (22)	81 (79)	93 (90)	23 (22)	21 (19)
Descriptors	Exclude	50 (49)	36 (35)	161 (157)	171 (162)	29 (29)	19 (19)
	Negative	106 (84)	126 (112)	341 (309)	514 (419)	49 (44)	49 (46)
	Neutral	4 (4)	10 (9)	38 (38)	29 (29)	5 (5)	6 (6)
	Positive	33 (33)	13 (13)	157 (152)	105 (104)	37 (35)	20 (20)

Table 9.4: Joint sex, race, and label distribution for the JHM and MIMIC datasets. The format is "# Examples (# Patients)". These distributions are insufficient for characterizing the extent to which demographic disparities are replicated within our dataset. A more thorough statistical analysis which controls for differences in anchor term usage, repeated measures, underlying conditions, and clinical specialty is necessary to make any substantive claims.

demographics. Additionally, any differences in prevalence of our anchor n-grams between demographic groups may be exploited by the linear classifier. The latter is expected given the extant literature which highlights demographic disparities in usage of stigmatizing language (Beach and Saha, 2021; Beach et al., 2021).

For these reasons, we ground the predictive performance achieved using the semantic representations against simple logistic regression baselines which model one-hot-encoded representations of the anchor n-gram, clinical speciality, and the primary stigmatizing language classification label. A qualitative review of instances in both datasets suggest there are likely additional auxiliary attributes not accounted for here (e.g., diagnoses) that would further explain the encoding of race and sex in the embeddings. For the MIMIC dataset, we consider the service which wrote the discharge summary (e.g., `SURG`, `GYN`, `PSYCH`) to be the speciality.

Demographic-Neutral Substitutions. During an initial run of the experiment, we recognized that patient sex could be easily inferred from the semantic representations due to the cues from gender-specific language. We adopt a naive approach to mitigate the presence of overt gender-informative language affecting conclusions within the demographic inference experiments. We replace gendered pronouns (e.g., he, herself), identifiers of sex (e.g., male, Mrs. Smith), and terms with non-uniform gender associations (e.g., husband, wife). The full mapping of substitutions is provided below in Table 9.5.

There are two limitations with this approach. First, we do not make substitutions for any patient names in the text. Second, we do not address any grammatical issues that arise after substitution of a gendered word (e.g., "he denies" $\rightarrow$ "they denies"). In

Original	Replacement
He, She	They
Him, Her	Them
His, Hers	Their
Himself, Herself	Themselves
Male, Female, Girl, Boy, Man, Woman	Person
Mr. XX, Ms. XX, Mrs. XX, Miss. XX	Patient
Husband, Wife	Partner

Table 9.5: Gender-informative words and their associated gender-neutral substitutions.

practice, the former implies that true amount of the sex-related information encoded in the learned embeddings may be lower than current estimates suggest. This case would only further strengthen our current conclusions. Regarding the latter, we find that any grammatical inconsistencies do not affect our ability to infer the stigma labels associated with each anchor embedding (Table 9.6).

We briefly explored using rules to obfuscate racial identifiers as well (e.g, "43 y.o. Asian"). We found this procedure difficult to perform automatically (e.g., "wearing black T-shirt") and likely to be a low-yield process based on a qualitative review of the instances in both datasets. For this reason, we opted not to include any race-neutral substitutions. Nonetheless, the lack of obfuscation should be noted while interpreting our results.

9.7.2.2 Results

We present demographic inference results in Table 9.6. Across all but one experimental setting, inference performance achieved using the gender-neutral version of the embeddings is not significantly different from what is achieved by the metadata-only baselines. This trend suggests that the learned embeddings encode little to no information about a patient's race or sex that cannot be explained by underlying differences in prevalence between patient populations. Future work is necessary to understand whether there exist semantic differences along other axes (e.g., socioeconomic status, substance use, obesity) (Healy et al., 2022).

9.7.3 Is stigma conveyed in the same manner across different patient populations?

Machine learning models trained on one distribution often experience a loss in performance when evaluated on a different distribution (Blitzer et al., 2006; Harrigian et al., 2020). Understanding the causes of this loss is necessary for ensuring systems do not exacerbate existing social disparities (Bender et al., 2021). Here, we identify speciality-specific nuances in stigmatizing language and highlight limitations of anchor-focused modeling.

Credibility & Obstinacy	JHM					MIMIC				
	Anchor	**Label**	**Speciality**	**Sex**	**Race**	**Anchor**	**Label**	**Speciality**	**Sex**	**Race**
Majority Baseline	0.03 ± 0.00	0.20 ± 0.02	0.11 ± 0.01	0.37 ± 0.01	0.26 ± 0.02	0.02 ± 0.00	0.22 ± 0.01	0.06 ± 0.01	0.33 ± 0.01	0.27 ± 0.01
Anchor	–	0.51 ± 0.05	0.14 ± 0.03	0.50 ± 0.04	0.31 ± 0.05	–	0.51 ± 0.02	0.07 ± 0.01	0.52 ± 0.04	0.27 ± 0.01
Label	0.08 ± 0.01	–	0.11 ± 0.01	0.37 ± 0.01	0.27 ± 0.03	0.09 ± 0.01	–	0.06 ± 0.01	0.51 ± 0.05	0.27 ± 0.01
Speciality	0.07 ± 0.02	0.31 ± 0.02	–	0.44 ± 0.04	0.36 ± 0.04	0.05 ± 0.01	0.32 ± 0.03	–	0.55 ± 0.05	0.28 ± 0.02
Anchor × Label	–	–	0.18 ± 0.05	0.50 ± 0.03	0.31 ± 0.05	–	–	0.07 ± 0.01	0.49 ± 0.04	0.28 ± 0.02
Anchor × Speciality	–	0.52 ± 0.06	–	0.51 ± 0.04	0.38 ± 0.03	–	0.60 ± 0.07	–	0.51 ± 0.02	0.28 ± 0.02
Label × Speciality	0.11 ± 0.02	–	–	0.47 ± 0.04	0.38 ± 0.04	0.10 ± 0.02	–	–	0.54 ± 0.04	0.27 ± 0.01
Anchor × Label × Speciality	–	–	–	0.54 ± 0.01	0.35 ± 0.03	–	–	–	0.51 ± 0.02	0.29 ± 0.02
Embedding	0.76 ± 0.05	0.95 ± 0.03	0.24 ± 0.03	0.76 ± 0.02	0.34 ± 0.02	0.92 ± 0.02	0.87 ± 0.03	0.11 ± 0.01	0.75 ± 0.02	0.30 ± 0.03
Embedding (Gender Neutral)	0.77 ± 0.06	0.93 ± 0.02	0.25 ± 0.04	0.59 ± 0.02	0.34 ± 0.06	0.92 ± 0.01	0.86 ± 0.06	0.10 ± 0.01	0.49 ± 0.03	0.33 ± 0.02

Compliance	JHM					MIMIC				
	Anchor	**Label**	**Speciality**	**Sex**	**Race**	**Anchor**	**Label**	**Speciality**	**Sex**	**Race**
Majority Baseline	0.01 ± 0.00	0.28 ± 0.01	0.08 ± 0.00	0.37 ± 0.02	0.29 ± 0.01	0.01 ± 0.00	0.24 ± 0.01	0.05 ± 0.00	0.33 ± 0.01	0.26 ± 0.01
Anchor	–	0.59 ± 0.02	0.18 ± 0.04	0.42 ± 0.02	0.29 ± 0.01	–	0.66 ± 0.02	0.05 ± 0.00	0.54 ± 0.02	0.27 ± 0.02
Label	0.03 ± 0.00	–	0.14 ± 0.01	0.37 ± 0.02	0.29 ± 0.01	0.03 ± 0.01	–	0.05 ± 0.00	0.47 ± 0.03	0.26 ± 0.01
Speciality	0.03 ± 0.00	0.28 ± 0.01	–	0.53 ± 0.04	0.29 ± 0.01	0.02 ± 0.01	0.34 ± 0.03	–	0.55 ± 0.02	0.26 ± 0.01
Anchor × Label	–	–	0.27 ± 0.02	0.46 ± 0.03	0.30 ± 0.01	–	–	0.07 ± 0.01	0.52 ± 0.03	0.31 ± 0.02
Anchor × Speciality	–	0.62 ± 0.05	–	0.54 ± 0.02	0.35 ± 0.03	–	0.67 ± 0.03	–	0.56 ± 0.02	0.30 ± 0.02
Label × Speciality	0.08 ± 0.01	–	–	0.53 ± 0.05	0.32 ± 0.02	0.08 ± 0.02	–	–	0.56 ± 0.03	0.28 ± 0.01
Anchor × Label × Speciality	–	–	–	0.54 ± 0.03	0.36 ± 0.02	–	–	–	0.54 ± 0.01	0.30 ± 0.02
Embedding	0.77 ± 0.04	1.00 ± 0.00	0.38 ± 0.05	0.57 ± 0.01	0.36 ± 0.02	0.86 ± 0.04	1.00 ± 0.00	0.13 ± 0.04	0.56 ± 0.03	0.33 ± 0.02
Embedding (Gender Neutral)	0.74 ± 0.04	1.00 ± 0.00	0.39 ± 0.05	0.52 ± 0.01	0.35 ± 0.01	0.85 ± 0.03	1.00 ± 0.00	0.12 ± 0.02	0.50 ± 0.04	0.34 ± 0.02

Descriptors	JHM					MIMIC				
	Anchor	**Label**	**Speciality**	**Sex**	**Race**	**Anchor**	**Label**	**Speciality**	**Sex**	**Race**
Majority Baseline	0.01 ± 0.00	0.14 ± 0.01	0.10 ± 0.01	0.35 ± 0.02	0.26 ± 0.01	0.00 ± 0.00	0.18 ± 0.00	0.05 ± 0.00	0.34 ± 0.02	0.28 ± 0.00
Anchor	–	0.83 ± 0.03	0.22 ± 0.02	0.50 ± 0.02	0.30 ± 0.03	–	0.87 ± 0.03	0.13 ± 0.01	0.56 ± 0.03	0.28 ± 0.01
Label	0.07 ± 0.00	–	0.10 ± 0.01	0.46 ± 0.07	0.26 ± 0.01	0.03 ± 0.00	–	0.06 ± 0.01	0.58 ± 0.03	0.28 ± 0.00
Speciality	0.01 ± 0.00	0.28 ± 0.03	–	0.58 ± 0.03	0.32 ± 0.03	0.03 ± 0.00	0.27 ± 0.03	–	0.44 ± 0.03	0.28 ± 0.00
Anchor × Label	–	–	0.30 ± 0.02	0.53 ± 0.02	0.32 ± 0.04	–	–	0.13 ± 0.01	0.56 ± 0.02	0.29 ± 0.01
Anchor × Speciality	–	0.84 ± 0.03	–	0.56 ± 0.04	0.34 ± 0.02	–	0.86 ± 0.02	–	0.57 ± 0.02	0.31 ± 0.02
Label × Speciality	0.09 ± 0.01	–	–	0.58 ± 0.04	0.32 ± 0.03	0.11 ± 0.01	–	–	0.57 ± 0.04	0.28 ± 0.00
Anchor × Label × Speciality	–	–	–	0.55 ± 0.03	0.36 ± 0.02	–	–	–	0.58 ± 0.02	0.30 ± 0.02
Embedding	0.82 ± 0.06	1.00 ± 0.00	0.45 ± 0.02	0.61 ± 0.04	0.34 ± 0.03	0.91 ± 0.02	1.00 ± 0.00	0.24 ± 0.05	0.58 ± 0.02	0.33 ± 0.02
Embedding (Gender Neutral)	0.82 ± 0.07	1.00 ± 0.00	0.44 ± 0.04	0.52 ± 0.03	0.34 ± 0.02	0.90 ± 0.03	1.00 ± 0.00	0.24 ± 0.04	0.54 ± 0.03	0.31 ± 0.02

Table 9.6: Macro F1 score ($\mu \pm \sigma$) for each attribute considered in §9.7.2. Higher inference performance suggests an attribute is more strongly encoded by (or correlated with) a given feature set. Differences in the prevalence of racial groups and sexes across auxiliary attributes (e.g., speciality, labels) can be exploited when inferring race and sex from the anchor embeddings.

		Credibility & Obstinacy		Compliance		Descriptors	
	Target →	**JHM**	**MIMIC**	**JHM**	**MIMIC**	**JHM**	**MIMIC**
Source ↕	**JHM**	**0.89 ± 0.03**	0.70 ± 0.01	**0.85 ± 0.02**	0.86 ± 0.03	**0.93 ± 0.02**	0.81 ± 0.03
	MIMIC	0.81 ± 0.03	**0.78 ± 0.03**	0.82 ± 0.02	**0.92 ± 0.02**	0.89 ± 0.03	**0.86 ± 0.01**

Table 9.7: Average test macro F1 score ($\mu \pm \sigma$) when transferring between datasets. There exists a statistically significant loss in performance (paired t-test $p < .05$) within all transfer settings (columns).

9.7.3.1 Methods

We evaluate models trained using the JHM dataset in §9.7.1 on the test set of the MIMIC dataset, and vice-versa. We also conduct a qualitative error analysis to understand how stigmatizing language differs between the two datasets.

Experimental Design. We use the clinical BERT models trained during the §9.7.1 experiments to evaluate domain-transfer. That is, we take the clinical BERT models (with anchor pooling) trained within each cross-validation fold and apply them to the test set of the opposite dataset (JHM → MIMIC, MIMIC → JHM). We *do not* modify or otherwise tune the existing models to improve transfer performance, with the primary goal being to understand differences in stigmatizing language usage between datasets (not to optimize generalization). To facilitate our qualitative analysis, we cache all test-set predictions and organize them into four groups based on whether the in-domain (source = target) and out-of-domain (source ≠ target) models characterized them correctly.

9.7.3.2 Results

We observe consistent drops in performance when models are evaluated in a different domain than which they were trained (i.e., Table 9.7). This performance loss is significant in all 6 transfer settings. What causes this loss? Are there spurious artifacts to which our models overfit (Wang et al., 2022b)? Or does each dataset contain unique stigmatizing language that arises disproportionately across patient populations?

Although many transfer errors can be attributed to differences in each dataset's joint anchor-label distribution, some special cases emerge. For example, models trained on the JHM dataset incorrectly characterize instances in MIMIC which describe parties secondary to the patient (e.g., family). This situation is more common in the MIMIC dataset due to ICU patients often being incapacitated. Models trained on the JHM dataset also struggle with statements in MIMIC from Psych ICU notes, where patients frequently describe their own behavior.

One on hand, these shortcomings appear to be a consequence of covariate shift (Sugiyama et al., 2007), for which many general mitigation strategies exist (Ramponi and Plank, 2020). On the other hand, each of the errors we observe presents a unique linguistic challenge that may be better handled using targeted interventions. Few-shot word sense disambiguation techniques may improve transfer for low-volume anchor-label pairs (Kumar et al., 2019; Scarlini et al., 2020), while

augmented annotations ma reduce speaker/receiver confusion (Rashkin et al., 2016; Hovy and Yang, 2021).

9.8 Measuring Health Disparities

With a suite of models capable of characterizing stigmatizing language at our disposal, we now ask whether there exists a disparity in the rate at which patients are subjected to different forms of stigma as a function of their race and/or sex. Unlike prior work that has either examined disparities at an anchor-level or more broadly at a sentiment-level (i.e., negative vs. positive) (Himmelstein et al., 2022; Sun et al., 2022), we perform our analysis using conceptual groupings of anchors, each having a different implication regarding the patient. This is made possible by our taxonomy and use of multiple independent models that treat different forms of stigma differently.

9.8.1 Methods

Due to the JHM dataset being composed of notes from multiple distinct specialities and clinical sites, we focus our analysis on the more uniform MIMIC-IV dataset. First, we extract candidate stigmatizing language instances from the entire MIMIC-IV dataset, maintaining the matched anchor and 10-word context window just as before.

Then, we run inference on the subset of candidate instances that are relevant for the each of the three models – i.e., Credibility & Obstinacy, Compliance, and Descriptors – and assign each candidate instance a single-label based on the argmax of the predicted class probabilities. We use the models trained on the MIMIC-IV training data without cross-validation (i.e., §9.7.2).

We first aggregate amongst the predictions within each note (i.e., discharge summary). Each (anchor, class label) tuple is treated independently – e.g., a note can contain an instance of 'claim' that was predicted as an 'Exclude', as well as an instance of 'claim' that was predicted as indicating 'Disbelief.' The presence of each (anchor, class label) tuple is treated as a binary outcome. We then generate composite outcome variables based on multiple (anchor, class label) tuples that share a common stigmatizing implication. We enumerate the composite outcomes in Table 9.8.

The MIMIC-IV dataset includes race characteristics at two levels (e.g., Hispanic - Spanish, Hispanic - Mexican); we only consider the first race level (e.g., Hispanic). Patients who had multiple encounters across the dataset with different race metadata attributes were excluded from the analysis.[4] Patients with an 'Asian' race were excluded due to the racial group having a low sample-size that was not conducive to

[4] These were predominantly cases of missing data (i.e., 'Unknown' race), though some included multiple known races which were remapped to a 'Mixed' race. Individuals with a 'Mixed' race attribute were eventually excluded.

Composite Outcome	**Class**	**Anchors (Includes Variants)**
Credibility	Disbelief	Adamant, Claim, Insist
	Negative	Drug-Seeking, Narcotic-Seeking, Secondary-Gain, Malinger, Exaggerate, Historian
Credibility (w/o Historian)	Disbelief	Adamant, Claim, Insist
	Negative	Drug-Seeking, Narcotic-Seeking, Secondary-Gain, Malinger, Exaggerate
Obstinacy	Difficult	Adamant, Claim, Insist
	Negative	Aggressive, Agitated, Angry, Argumentative, Belligerent, Combative, Confrontational, Uncooperative, Cooperative, Defensive
Obstinacy (w/o Agitated)	Difficult	Adamant, Claim, Insist
	Negative	Aggressive, Angry, Argumentative, Belligerent, Combative, Confrontational, Uncooperative, Cooperative, Defensive
Negative Compliance	Negative	Adhere, Adherent, Nonadherent, Comply, Compliant, Noncompliant, Decline, Refuse, Unmotivated, Unwilling
Negative Compliance (Appropriate)	Negative	Adhere, Adherent, Nonadherent, Decline
Negative Compliance (Inappropriate)	Negative	Comply, Compliant, Noncompliant, Refuse, Unmotivated, Unwilling
Positive Compliance	Positive	Adhere, Adherent, Comply, Compliant
Negative Appearance	Negative	Disheveled, Poorly-Groomed, Unkempt, Well-Groomed
Positive Appearance	Positive	Well-Groomed
Positive Demeanor	Positive	Charming, Delightful, Lovely, Pleasant

Table 9.8: Composite stigmatizing language outcome variables. Each grouping of (anchor, label) pairs presents a unique stigmatizing implication regarding a patient.

statistical testing. Patients with a 'Mixed' race, 'Other' race, or 'Unknown' race were excluded due to potential heterogeneity within the groups. Thus, our final sample contains notes for patients who have a sex if 'Male' or 'Female', and a race of 'White', 'Black', or 'Hispanic or Latino.'

For descriptive purposes, we compute two rates of occurrence for the composite outcomes – 1) the rate per note and 2) the rate per patient. We compute the rates across the entire population, broken down by patient gender, and broken down by patient race. To compare differences across the groups, we compute odds ratios using mixed effects logistic regression, clustering by patient identifier to account for repeated measures. We fit one model for gender (`Outcome` $\sim$ `Gender + (1|Patient ID)`) and one model for race (`Outcome` $\sim$ `Race + (1|Patient ID)`). We use the 'Male' sex and 'White' race as reference levels, in line with recommendations to measure health disparities relative to historically advantaged groups (Braveman et al., 2018).

9.8.2 Results

We report the rates of occurrence for each of the composite outcome variables in Table 9.9, and the odds ratios relative to the historically advantaged groups in Table 9.10. Female patients are less likely than Male patients to be implicated as obstinate, and are also spoken about more positively (i.e., less negativity and more positivity

Rate Per Note	**Overall** (N=280,699)	**Female** (N=144,700)	**Male** (N=135,999)	**Hispanic** (N=15,823)	**Black** (N=46,474)	**White** (N=218,402)
Credibility	3.3 (3.2, 3.3)	3.1 (3.1, 3.2)	3.4 (3.3, 3.5)	3.0 (2.7, 3.3)	4.1 (3.9, 4.3)	3.1 (3.0, 3.2)
Credibility (w/o Historian)	1.5 (1.5, 1.5)	1.4 (1.3, 1.5)	1.6 (1.5, 1.7)	1.5 (1.3, 1.7)	1.8 (1.7, 1.9)	1.4 (1.4, 1.5)
Obstinacy	7.1 (7.1, 7.2)	6.3 (6.2, 6.5)	8.0 (7.9, 8.2)	5.8 (5.5, 6.2)	7.4 (7.2, 7.6)	7.2 (7.1, 7.3)
Obstinacy (w/o Agitated)	2.6 (2.6, 2.7)	2.3 (2.2, 2.4)	3.0 (2.9, 3.1)	2.3 (2.1, 2.5)	3.3 (3.1, 3.4)	2.5 (2.5, 2.6)
Negative Compliance	15.6 (15.4, 15.7)	15.5 (15.3, 15.6)	15.7 (15.5, 15.8)	15.5 (14.9, 16.0)	22.0 (21.6, 22.4)	14.2 (14.0, 14.3)
Negative Compliance (Appropriate)	6.6 (6.5, 6.7)	6.7 (6.6, 6.8)	6.6 (6.5, 6.7)	6.3 (5.9, 6.6)	9.3 (9.0, 9.6)	6.1 (6.0, 6.2)
Negative Compliance (Inappropriate)	10.9 (10.8, 11.0)	10.6 (10.5, 10.8)	11.2 (11.0, 11.4)	11.2 (10.7, 11.7)	16.2 (15.9, 16.6)	9.8 (9.6, 9.9)
Positive Compliance	3.9 (3.8, 3.9)	3.5 (3.4, 3.6)	4.2 (4.1, 4.3)	4.9 (4.6, 5.3)	5.1 (4.9, 5.3)	3.5 (3.4, 3.6)
Negative Appearance	0.9 (0.8, 0.9)	0.7 (0.6, 0.7)	1.1 (1.0, 1.1)	0.8 (0.6, 0.9)	0.7 (0.7, 0.8)	0.9 (0.9, 0.9)
Positive Appearance	0.8 (0.8, 0.9)	0.9 (0.9, 1.0)	0.8 (0.7, 0.8)	1.0 (0.9, 1.2)	1.0 (0.9, 1.1)	0.8 (0.7, 0.8)
Positive Demeanor	18.4 (18.2, 18.5)	19.0 (18.8, 19.2)	17.7 (17.5, 17.9)	17.1 (16.5, 17.7)	18.6 (18.2, 19.0)	18.4 (18.2, 18.5)

Rate Per Patient	**Overall** (N=122,070)	**Female** (N=63,409)	**Male** (N=58,661)	**Hispanic** (N=15,823)	**Black** (N=46,474)	**White** (N=21,8402)
Credibility	5.9 (5.8, 6.1)	5.8 (5.6, 6.0)	6.1 (5.9, 6.3)	5.7 (5.1, 6.2)	8.0 (7.6, 8.4)	5.6 (5.4, 5.7)
Credibility (w/o Historian)	2.7 (2.6, 2.8)	2.6 (2.5, 2.7)	2.8 (2.7, 2.9)	3.0 (2.6, 3.4)	3.6 (3.3, 3.9)	2.5 (2.4, 2.6)
Obstinacy	12.0 (11.8, 12.2)	10.6 (10.4, 10.9)	13.5 (13.3, 13.8)	10.1 (9.4, 10.8)	12.8 (12.3, 13.3)	12.0 (11.8, 12.2)
Obstinacy (w/o Agitated)	4.9 (4.8, 5.0)	4.3 (4.2, 4.5)	5.5 (5.4, 5.7)	4.5 (4.0, 5.0)	6.6 (6.2, 7.0)	4.6 (4.5, 4.8)
Negative Compliance	21.6 (21.4, 21.8)	21.7 (21.4, 22.0)	21.5 (21.2, 21.9)	20.6 (19.6, 21.6)	29.0 (28.3, 29.7)	20.4 (20.1, 20.6)
Negative Compliance (Appropriate)	11.3 (11.2, 11.5)	11.4 (11.1, 11.6)	11.3 (11.1, 11.6)	10.9 (10.1, 11.6)	16.2 (15.7, 16.8)	10.5 (10.3, 10.7)
Negative Compliance (Inappropriate)	15.3 (15.1, 15.5)	15.2 (14.9, 15.5)	15.5 (15.2, 15.8)	14.9 (14.0, 15.7)	21.7 (21.1, 22.4)	14.2 (14.0, 14.5)
Positive Compliance	7.0 (6.8, 7.1)	6.4 (6.2, 6.6)	7.5 (7.3, 7.8)	8.7 (8.1, 9.4)	9.9 (9.5, 10.4)	6.3 (6.2, 6.5)
Negative Appearance	1.6 (1.5, 1.7)	1.3 (1.2, 1.4)	2.0 (1.9, 2.1)	1.5 (1.2, 1.8)	1.7 (1.5, 1.9)	1.6 (1.5, 1.7)
Positive Appearance	1.8 (1.7, 1.8)	1.9 (1.8, 2.0)	1.6 (1.5, 1.7)	2.2 (1.8, 2.5)	2.4 (2.1, 2.6)	1.6 (1.5, 1.7)
Positive Demeanor	28.6 (28.3, 28.8)	29.2 (28.9, 29.6)	27.9 (27.5, 28.3)	26.1 (25.0, 27.1)	31.8 (31.1, 32.5)	28.2 (27.9, 28.5)

Table 9.9: Stigmatizing language prevalence estimates for the MIMIC-IV dataset. Rate per note (top) and rate per patient (bottom). 95% confidence intervals estimated using normal approximation for binomial distribution.

Outcome	Female OR	Hispanic OR	Black OR
Credibility	0.95 (0.83, 1.09)	0.97 (0.71, 1.32)	1.24 (1.04, 1.49) *
Credibility (w/o Historian)	0.93 (0.76, 1.14)	1.14 (0.74, 1.76)	1.18 (0.91, 1.53)
Obstinacy	0.74 (0.67, 0.82) *	0.79 (0.62, 1.00)	0.91 (0.79, 1.06)
Obstinacy (w/o Agitated)	0.77 (0.66, 0.90) *	0.93 (0.65, 1.33)	1.20 (0.98, 1.48)
Negative Compliance	1.01 (0.98, 1.05)	1.04 (0.97, 1.13)	1.79 (1.70, 1.87) *
Negative Compliance (Appropriate)	1.03 (0.94, 1.13)	1.05 (0.85, 1.28)	1.57 (1.39, 1.77) *
Negative Compliance (Inappropriate)	0.98 (0.90, 1.06)	1.07 (0.89, 1.28)	1.73 (1.55, 1.92) *
Positive Compliance	0.85 (0.76, 0.96) *	1.44 (1.13, 1.84) *	1.43 (1.22, 1.67) *
Negative Appearance	0.68 (0.50, 0.93) *	0.91 (0.45, 1.85)	0.85 (0.54, 1.33)
Positive Appearance	1.13 (1.13, 1.13) *	1.13 (0.59, 2.17)	1.29 (0.90, 1.86)
Positive Demeanor	1.10 (1.08, 1.13) *	0.91 (0.86, 0.96) *	1.04 (1.01, 1.08) *

Table 9.10: Odds ratios for stigmatizing language prevalence relative to historically advantaged groups (Male, White). Ratios were computed using mixed effects binomial logistic regression. 95% confidence intervals were estimated using the Wald test. Asterisks (*) indicate results that are statistically significant at a $p < .05$ level.

regarding appearance, more positivity regarding demeanor and compliance). In comparison, both negative and positive compliance is discussed more frequently for Hispanic and Black patients than for White patients. Over-indexing in both positive and negative sentiments of stigmatizing language may be due to the presence of an underlying provider expectation that is subverted by the patient, in turn-causing a tendency to overcorrect the implicit bias (Dovidio and Gaertner, 2004; Mendes and Koslov, 2013). Hispanic patients are slightly less likely than White patients to receive praise for positive demeanor, while Black patients are slightly more likely than White patients to receive praise for positive demeanor. Finally, Black patients are significantly more likely than White patients to have their credibility questioned, though this effect goes away when excluding instances containing 'historian.'

Overall, these results provide reasonable evidence that contemporary healthcare providers may possess an unconscious bias against individuals of historically marginalized racial groups. At the same time, there is evidence to suggest that Male patients receive unfavorable forms of stigmatization more frequently than Female patients. While the former result constitutes a health disparity given the effect on historically marginalized groups, the latter may only be classified as a public health concern because Males are typically considered historically advantaged.

Nonetheless, there are multiple potential caveats to these observations. First, the estimates are derived from a high-performing, but error prone, machine learning model. Second, this analysis does not control for various possible confounds – e.g., provider-patient concurrence (i.e., same race or sex), medical comorbidities (e.g., conditions in which patients historically do not adhere to treatment recommendations (Thier et al., 2008; Sharma et al., 2014; Ferdinand et al., 2017)), clinical setting, or practitioner type. Third, the MIMIC-IV dataset represents just a subset of encounters at a single urban, academic institution; both the patient and provider population may influence the degree of stigmatization we observe. Finally, this analysis only considers a patient's race and sex independently, ignoring potential effects of intersectionality (Homan et al., 2021). As this exploration is merely meant to be a starting point to showcase the opportunity to identify health disparities, we leave these open questions

for future work.

9.9 Review of Learnings

Technical contributions in the field of natural language processing do not always need to be centered around a model, algorithm, or mathematical expression. This study, which focuses primarily on a novel task, its relationship to existing literature, and its operationalization for health equity research, serves as a prime example.

We started the study by introducing a new NLP task – characterizing stigmatizing language in medical records – and then theoretically embedding it within the broader harmful language landscape. We argued and demonstrated empirically that stigmatizing language in medical records shares the most similarity with linguistic microaggressions, as they are both typically reflective of unconscious bias and communicated through implicit mechanisms that require context to facilitate interpretation and impact. However, unlike other forms of harmful language, stigmatizing language in medical records is typically concentrated in relatively benign anchor words (i.e., §9.7.1). Furthermore, stigmatizing anchor words share semantic similarity across demographic groups (i.e., §9.7.2), but are more likely to be used in different contexts across clinical settings (i.e., §9.7.3).

After introducing, validating, and inspecting the models that facilitate the

characterization of stigmatizing language in medical records, we measured the prevalence of different forms of stigmatizing language in the public, MIMIC-IV dataset. In line with previous studies (Beach et al., 2021; Himmelstein et al., 2022; Sun et al., 2022), we found evidence to suggest that different races and sexes experience stigmatization in the medical record at different rates (i.e. §9.8). Furthermore, some of these differences affect historically marginalized groups disproportionately and thus, by definition, represent a health disparity.

9.10 Challenges and Recommendations

The current reliance on domain experts to identify possible instances of bias using anchor terms is limiting given the adversarial relationship been abusive language and speakers (Nobata et al., 2016). It also does not address abstract forms of stigma (Kopera et al., 2015) or stigmatizing pragmatics (Beach and Saha, 2021).

Methods for discovering stigmatizing language in medical records are poised to be highly impactful (Field and Tsvetkov, 2020). Counterfactual analyses may be instrumental for better characterizing the nuance between stigmatizing and non-stigmatizing clinical language (Kaushik et al., 2019). Whether these nuances are uniform across patient populations (e.g., hospital systems, regions) and providers (e.g., nurses, resident physicians) remains an open question not answerable from our

datasets alone. Likewise, future work is necessary to understand whether clinical knowledge is necessary for models in this domain (Roberts, 2016).

9.11 Limitations

In our work we faced numerous types of limitations that fall under different categories.

Data. Our relatively small dataset size limits our analysis, especially with the use of language models. Furthermore, the label distribution is skewed across the different specialties (domains), which affects model performance, robustness and generalizability. The differences in distribution might be the result of how the data was collected, which was not in light of the anchor words, or due to the domain's nature and/or the medical providers' language of that specialty. Furthermore, the time frame that the data was sampled from might manifest certain biases that are different from other time frames. Finally, our datasets are only representative of a small number of specialties from two medical institutions. Patient populations and providers may vary greatly across medical fields and additional institutions.

Task. The formulation of the labels for our task imposes limitations and challenges. Stigmatizing language is subjective and can vary between the perspective of the patient and the medical provider. As a result, we are aware that our medical experts' annotations might impose a bias. Additionally, the negative connotations

of language might be ambiguous and can change depending on a medical expert's identity, background and specialty, which creates a bias that is hard to mitigate.

Computational Resources. We only used IRB-approved servers to access the datasets and perform our experiments. Because these platforms had limited computational capacity and lacked the specifications required to build more complex neural models, we were not able to include more recent language models in our experiments that might have yielded better performance. In the future, we hope to have access to machines that support more recent and state-of-the-art models.

9.12 Ethical Considerations

Our datasets were collected from real patients, contain protected health information (PHI), and are subject to HIPAA regulations. As a result, we took the utmost care to maintain data integrity and privacy. First, we obtained IRB approval to access and process the data. Second, we obtained permission and approval for all applications and libraries used to process the data. Third, data storage and computational experimentation was done on IRB-approved platforms.

9.13 Discussion

This chapter serves as both a call to action and a warning for ML and NLP practitioners. On one hand, we demonstrated that while ML and NLP have an inherent risk of perpetuating existing health disparities, they can also be used to proactively identify disparities that were previously unknown or not able to be measured at scale. The insights regarding the (disparate) presence of stigmatizing language in the MIMIC-IV dataset may cause practitioners to think differently about how they train clinical language models. Meanwhile, the models we developed may be used to sunshine other instances of stigmatization in different patient populations, and then to potentially mitigate their incidence moving forward (e.g., via an 'autocorrect' writing assistant or provider report card).

On the other hand, we showed that practitioners cannot work towards these positive outcomes without also maintaining an awareness for statistical biases in training data that may inhibit model robustness. As seen in the dataset transfer experiments from §9.7.3, distribution shift can arise in task-specific ways that cannot always be easily predicted *a priori* or deciphered via metrics alone (e.g., generalization gap). Instead, they may require practitioners to perform a thorough qualitative analysis of the data and model behavior to reveal the underlying confound. Future exploration is necessary to first understand how heterogeneity in stigmatizing language

affects other patient populations and clinical settings, and then to implement corrective mechanisms that address it.

9.14 Looking Ahead

In the next and final chapter, we will we consolidate and summarize the contributions set forth in the prior chapters. We will also discuss future opportunities for promoting health equity using NLP.

Part V: Conclusion

Chapter 10

Conclusion

10.1 Contributions

This thesis presented a series of efforts to move the field of natural language processing (NLP) towards a state in which it can be used with confidence to promote health equity. These efforts fell into two broad categories. "Defensive" methods sought to prevent NLP and machine learning applications from (re-)introducing disparities. These focused specifically on measuring and understanding biases within data, as well as counteracting their effects to train robust models. "Proactive" methods, in comparison, sought to measure existing health disparities so that they could then be addressed through alternative mechanisms (potentially themselves using robust NLP methods). Broadly, our "defensive" contributions targeted *statistical* biases, while our "proactive" contributions targeted *social* biases.

In Chapter 4, we introduced an annotated repository of social media datasets used for modeling mental health status. Using this resource, we identified systematic issues with current dataset curation practices in the mental health modeling domain and then provided evidence-based recommendations to ameliorate them moving forward. We connected the current issues in this specific application domain to dataset quality in health more generally.

In Chapter 5, we performed a suite of empirical experiments to measure and understand selection biases in widely adopted social media datasets for mental

health modeling. While the experimental suite (which included some non-traditional techniques) can certainly be used as a guide for other practitioners as they refine dataset curation methods, the greater contribution is likely the measurement itself which affirmed the concerns presented in Chapter 4.

In Chapter 6, we continued our examination of social media datasets used for mental health, focusing more acutely on the issues that health status annotations based on self-disclosure present. We introduced a counterfactual explanation method for modeling problems involving nested text (e.g., sentences in a document, posts in a user history) that facilitates the discovery of selection biases. We used this approach to identify previously unrecognized personality-related and temporally-acute artifacts that arise in self-disclosure-based datasets.

In Chapter 7, we called into question the validity of longitudinal social media studies of mental health surrounding the COVID-19 pandemic. We adapted an existing semantic shift measurement to instead be used for robust and interpretable feature selection, showing that it improved generalization in the presence of distribution shift. We further demonstrated using this mechanism that semantic shift can lead to undesirable volatility in language-based estimates of longitudinal health status change.

In Chapter 8, we shifted away from non-clinical data to instead focus on distribution shift in clinical settings. Through a series of thorough experiments,

we demonstrated that, counter to existing intuition, out-of-distribution clinical pre-training does not offer an advantage to language models compared to out-of-distribution non-clinical pre-training. We used these insights to ground a recommendation that practitioners allocate more effort to adapting general language models to clinical data than to training clinical language models from scratch. As an additional note, the dataset and models curated as part of this study are being used in ongoing work to measure health disparities in the Ophthalmology domain (e.g., Cai et al. (2023) and Cai, Cindy X et al. (2024)).

Finally, in Chapter 9, we introduced a new NLP task – characterizing stigmatizing language in medical records – and grounded it within existing harmful language research via a set of empirical experiments. Through these experiments, we also provided a novel finding regarding the heterogeneity of stigmatizing language across clinical settings. Thereafter, we used data and model resources generated from the study to measure stigmatizing language in a widely-adopted public clinical NLP dataset and showed that this language was used disproportionately for historically marginalized racial groups.

10.2 Future Directions

As mentioned at the start of this thesis, many of the scientific and technical questions guiding my work have been asked and answered previously (Jeong et al., 2024). However, because both the health domain and society more generally are constantly evolving, no study or method is likely to ever be the end-all be-all for a particular inquiry. While I genuinely believe this thesis made progress toward promoting equity in healthcare, the goal posts have already moved and thus require continued innovation to ensure they do not fall out of reach. I foresee several clear opportunities to meet this demand.

10.2.1 Measuring and Understanding Distribution Shift

Large language models such as ChatGPT (Achiam et al., 2023) and Llama (Touvron et al., 2023) have shown a remarkable ability to not only encode large amounts of knowledge, but also to facilitate interaction with that knowledge via natural language instructions (Kojima et al., 2022; Ouyang et al., 2022). Some contemporary work has expressed concerns related to these models being overfit to the data used for training them (Gao et al., 2021; Biderman et al., 2024; Li and Flanigan, 2024). While such

behavior may be detrimental for their ability to perform downstream tasks, it may actually be advantageous in the pursuit of identifying issues such as selection bias and distribution shift.

Practitioners could potentially treat an LLM as a fuzzy database for a dataset (or set of datasets) (Petroni et al., 2019), and then use the LLM to probe the database in pursuit of sample artifacts. In fact, LLMs have already shown an ability to perform abstract queries of unstructured data (Zhu et al., 2023; Li et al., 2024; Zhao et al., 2024), and also shown an ability to reason about linguistic differences between sets of utterances (Chen et al., 2023b). They have even been used for superficial data science analyses (e.g., computing distributional statistics, generating visualizations) (Hassani and Silva, 2023; Ma et al., 2023). Integrating the aforementioned abilities together to perform more complex linguistic-focused analyses is likely already possible using methods such as multi-step promoting (Paranjape et al., 2023) and due to technical advancements that enable the use of massive context windows (Chen et al., 2023a; Wang et al., 2024). An example prompt may be:

> *Please retrieve instances from Dataset A and Dataset B that meet the following criteria:*
>
> - *Contain a disclosure of a mental health diagnosis (e.g., "I was diagnosed with depression." or "My anti-depressant prescription is running low.")*
> - *Written by a teenager that typically does not make posts about their mental health*

- *Written between January 2023 and March 2023*

How do documents in Dataset A that include a mental health diagnosis typically differ from documents in Dataset B that include a mental health diagnosis? Please provide specific examples to showcase your reasoning.

The bigger challenge would likely be automating the artifact and bias search procedure such that practitioners would only need to inspect a set of proposed artifacts rather than iteratively probe for them. Because sample biases are defined relative to different populations and tasks, there would still likely need to be some manual guidance to initialize the search. However, this would be significantly less burdensome. One possible strategy for accomplishing this may be to leverage chain-of-thought prompting to extract a model's reasoning regarding a task (Wei et al., 2022), and then to summarize this reasoning for a practitioner to review. Prior work on dataset bias discovery may prove useful for inspiring LLM-based methods (Le Bras et al., 2020; Liu et al., 2021a).

10.2.2 Promoting Robustness Under Distribution Shift

There are opportunities to extend the semantic-stability feature selection method presented in Chapter 7. Techniques for learning context-aware and metadata-aware

embeddings may be used to measure semantic stability across more than two discrete distributions (Hamilton et al., 2016; Bamler and Mandt, 2017; Arora et al., 2022). This would be primarily useful from a computational feasibility perspective, since the current approach requires the training of a separate embedding model for every distribution. Not only is the current approach costly, but also not always possible for some populations due to sample size constraints (Antoniak and Mimno, 2018). With the latter issue in particular, it may be useful to leverage hierarchical approaches that loosely share information across distributions but still allow for distribution-specific variation (Goswami et al., 2021). Nevertheless, work to extend the measurement and selection method to support multiple and/or continuously-defined distributions is likely to be more useful for dataset understanding than for actually facilitating model generalization given the transition away manual feature engineering in NLP.

It will likely be more fruitful for practitioners to focus their attention moving forward on improving language model performance for highly specialized domains. As discussed originally in Chapter 8 and further corroborated in ongoing work (Yi et al., 2024), language models trained from scratch on domain-related data are not necessarily the solution to the unique demands of the health domain. There simply is not enough relevant data available – due in part to the specificity of the domain and in part to privacy constraints – to achieve the same degree of generalization

achieved by LLMs trained from scratch on general text corpora. There exists a need to further develop methods for adapting general language models to specific applications, in particular in low-resource settings (Zhang et al., 2020; Diao et al., 2021). Test-time adaptation approaches may be even more relevant for certain health applications where practitioners cannot readily expect to have access to their target data distribution (McDermott et al., 2023; Shi et al., 2024).

10.2.3 Identifying and Mitigating Health Disparities

Robust models are certainly a dependency in machine learning frameworks that target issues related to health equity, but can ultimately only have an effect as large as the setting in which they operate. For this reason, practitioners interested in promoting health equity directly should likely focus their attention on the "proactive" dimension of the equity research domain. The most promising avenue for doing so is by collaborating closely with experts (e.g., clinicians, sociologists, healthcare administrators) to build actionable measurement devices and systems that facilitate behavioral change in direct response to their hypotheses. Practitioners who don't have the opportunity to do so, however, may still consider building off of the foundational

work related to characterizing stigmatizing language that we've presented.

First, there remains a need to apply our stigmatizing language models to larger, more diverse patient populations than what is available in MIMIC-IV, and to better understand the mechanisms by which stigmatizing language arises disproportionately and acts disproportionately across groups. Are there any confounding factors that drive our observations (e.g., comorbid conditions, unobserved social characteristics)? Does stigmatizing language have an effect on tangible health outcomes (e.g., transplant decisions, loss to follow up)?

Second, practitioners may ask whether the presence of stigmatizing language in clinical data affects language model behavior. For example, preliminary work from Liu et al. (2023) suggests that stigmatizing language negatively affects mortality risk prediction. How does it affect generative language models? Is stigmatizing language perpetuated in synthetic medical records? If so, how do we train language models that convey important, clinically-relevant information while also not simply refusing to say certain things (e.g., expressing doubt about a patient's testimony)? Can we train language models such that they implicitly translate stigmatizing versions of clinically-important information into non-stigmatizing versions?

Finally, there are various other forms of language in medical records that are of concern with respect to health disparities. On the stigmatizing language front, there

§	Publication	Code & Data (https://github.com/kharrigian/)
4	Harrigian et al. (2021)	mental-health-datasets
5	Harrigian et al. (2020)	emnlp-2020-mental-health-generalization
6	Harrigian and Dredze (2022b)	—
7	Harrigian and Dredze (2022a)	semantic-shift-websci-2022
8	Harrigian et al. (2023a)	ml4h-clinical-bert
9	Harrigian et al. (2023b)	ehr-stigma

Table 10.1: Publications highlighted by each chapter of the thesis, as well as source code and data released with them.

remains a need to characterize quotations and evidentials (Hobbs, 2003; Schuman and Romero, 2023), to measure stigma outside of the EHR (e.g., in patient messages, internal hospital communications), and to perform discovery of stigmatization in the absence of manually-defined anchor keywords. The latter includes the discovery of stigmatizing sentiments that are expressed in multiple sentences, potentially with or without adjacency to one another. Beyond stigmatizing language, practitioners may look to characterize humanizing and personalizing language in the EHR – treating it as an indicator of patient-provider relationship quality (McCarthy et al., 2013; Murray and McCrone, 2015).

10.3 Software Releases

Publicly released code and data is presented in Table 10.1.

Bibliography

Jacob Cohen (1960). "A coefficient of agreement for nominal scales". *Educational and psychological measurement* 20.1, pages 37–46 (cited on page 137).

Lawrence L Weed (1968). "Medical records that guide and teach". *New England Journal of Medicine* 278.12, pages 652–657 (cited on page 204).

Oscar Firschein, Martin A Fischler, L Stephen Coles, and Jay M Tenenbaum (1973). "Forecasting and assessing the impact of artificial intelligence on society". *IJCAI*. Volume 5. 1. Citeseer, pages 105–120 (cited on page 6).

Carlo Faravelli, Giorgio Albanesi, and Enrico Poli (1986). "Assessment of depression: a comparison of rating scales". *Journal of affective disorders* (cited on page 71).

William W Dressler (1991). *Stress and adaptation in the context of culture: Depression in a southern black community* (cited on page 169).

Margaret Whitehead (1992). "The concepts and principles of equity and health". *International journal of health services* 22.3, pages 429–445 (cited on pages 19, 20).

BIBLIOGRAPHY

Joao Pereira (1993). "What does equity in health mean?" *Journal of Social Policy* 22.1, pages 19–48 (cited on page 19).

William E Boebert, Thomas R Markham, and Robert A Olmsted (1994). *Data enclave and trusted path system*. US Patent 5,276,735 (cited on page 72).

Steven A Julious and Mark A Mullee (1994). "Confounding and Simpson's paradox". *Bmj* 309.6967, pages 1480–1481 (cited on page 164).

Ciyou Zhu, Richard H Byrd, Peihuang Lu, and Jorge Nocedal (1997). "Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization". *TOMS* (cited on page 264).

Aharon Ben-Tal and Arkadi Nemirovski (1998). "Robust convex optimization". *Mathematics of operations research* 23.4, pages 769–805 (cited on page 48).

Ricardo Baeza-Yates, Berthier Ribeiro-Neto, et al. (1999). *Modern information retrieval*. ACM press New York (cited on page 130).

Christopher J Murray, Emmanuela E Gakidou, and Julio Frenk (1999). "Health inequalities and social group differences: what should we measure?" *Bulletin of the World Health Organization* 77.7, page 537 (cited on page 20).

World Health Organization (2000). *The world health report 2000: health systems: improving performance*. World Health Organization (cited on page 19).

BIBLIOGRAPHY

Healthy People (2000). *Healthy People 2010: Understanding and Improving Health.* US Department of Health and Human Services (cited on page 20).

Hidetoshi Shimodaira (2000). "Improving predictive inference under covariate shift by weighting the log-likelihood function". *Journal of statistical planning and inference* 90.2, pages 227–244 (cited on page 83).

Sudhir Anand, Finn Diderichsen, Timothy Evans, Vladimir M Shkolnikov, Meg Wirth, et al. (2001). "Measuring disparities in health: methods and indicators". *Challenging inequities in health: from ethics to action*, pages 49–67 (cited on page 23).

Alexandra Bambas, Juan Antonio Casas, et al. (2001). "Assessing equity in health: conceptual criteria". *Equity and health: Views from the Pan American Sanitary Bureau*, pages 12–21 (cited on page 19).

Charles L Bowden (2001). "Strategies to reduce misdiagnosis of bipolar depression". *Psychiatric Services* (cited on page 147).

C Laurel Franklin and Mark Zimmerman (2001). "Posttraumatic stress disorder and major depressive disorder: Investigating the role of overlapping symptoms in diagnostic comorbidity". *The Journal of nervous and mental disease* (cited on page 146).

BIBLIOGRAPHY

Peter Glick and Susan T Fiske (2001). "An ambivalent alliance: Hostile and benevolent sexism as complementary justifications for gender inequality." *American psychologist* 56.2, page 109 (cited on page 252).

Gary King and Langche Zeng (2001). "Logistic regression in rare events data". *Political analysis* 9.2, pages 137–163 (cited on page 240).

James W Pennebaker, Martha E Francis, and Roger J Booth (2001). "Linguistic inquiry and word count: LIWC 2001". *Mahway: Lawrence Erlbaum Associates* 71.2001, page 2001 (cited on page 85).

Qing-Song Xu and Yi-Zeng Liang (2001). "Monte Carlo cross validation". *Chemometrics and Intelligent Laboratory Systems* (cited on page 129).

Nathan Berg and Donald Lien (2002). "Measuring the effect of sexual orientation on income: Evidence of discrimination?" *Contemporary economic policy* 20.4, pages 394–414 (cited on page 22).

Stephen L Buka (2002). "Disparities in health status and substance use: ethnicity and socioeconomic factors." *Public health reports* 117.Suppl 1, S118 (cited on page 3).

Sheri L Johnson and Andrzej Nowak (2002). "Dynamical patterns in bipolar depression". *Personality and Social Psychology Review* (cited on page 123).

Roderick JA Little and Donald B Rubin (2002). *Statistical analysis with missing data: Wiley series in probability and statistics* (cited on page 35).

BIBLIOGRAPHY

Alan Nelson (2002). "Unequal treatment: confronting racial and ethnic disparities in health care." *Journal of the national medical association* 94.8, page 666 (cited on page 248).

David M Blei, Andrew Y Ng, and Michael I Jordan (2003). "Latent dirichlet allocation". *Journal of machine Learning research* 3.Jan, pages 993–1022 (cited on pages 93, 104).

W Drew Gouvier, Sara Sytsma-Jordan, and Stephen Mayville (2003). "Patterns of discrimination in hiring job applicants with disabilities: The role of disability type, job complexity, and public contact." *Rehabilitation psychology* 48.3, page 175 (cited on page 76).

Pamela Hobbs (2003). "The use of evidentiality in physicians' progress notes". *Discourse Studies* 5.4, pages 451–478 (cited on page 301).

Ronald C Kessler, Patricia Berglund, Olga Demler, Robert Jin, Doreen Koretz, Kathleen R Merikangas, A John Rush, Ellen E Walters, and Philip S Wang (2003). "The epidemiology of major depressive disorder: results from the National Comorbidity Survey Replication (NCS-R)". *Jama* 289.23, pages 3095–3105 (cited on pages 73, 105).

BIBLIOGRAPHY

Juan Ramos et al. (2003). "Using tf-idf to determine word relevance in document queries". *Proceedings of the first instructional conference on machine learning*. Volume 242. Piscataway, NJ, pages 133–142 (cited on page 93).

Stephan Collishaw, Barbara Maughan, Robert Goodman, and Andrew Pickles (2004). "Time trends in adolescent mental health". *Journal of Child Psychology and psychiatry* 45.8 (cited on pages 122, 187).

John F Dovidio and Samuel L Gaertner (2004). "Aversive racism". *Advances in experimental social psychology* 36, pages 4–56 (cited on page 282).

Klaus Krippendorff (2004). "Reliability in content analysis: Some common misconceptions and recommendations". *Human communication research* 30.3, pages 411–433 (cited on page 137).

Tarja K Melartin, Heikki J Rytsala, Ulla S Leskela, Paula S Lestela-Mielonen, T Petteri Sokero, and Erkki T Isometsa (2004). "Severity and comorbidity predict episode duration and recurrence of DSM-IV major depressive disorder". *Journal of Clinical Psychiatry* 65.6, pages 810–819 (cited on page 113).

Stephanie Rude, Eva-Maria Gortner, and James Pennebaker (2004). "Language use of depressed and depression-vulnerable college students". *Cognition & Emotion* 18.8, pages 1121–1133 (cited on page 81).

BIBLIOGRAPHY

Bianca Zadrozny (2004). "Learning and evaluating classifiers under sample selection bias". *Proceedings of the twenty-first international conference on Machine learning*, page 114 (cited on page 34).

Kenneth Keppel, Elsie Pamuk, John Lynch, Olivia Carter-Pokras, Insun Kim, Vickie Mays, Jeffrey Pearcy, Victor Schoenbach, and Joel S Weissman (2005). "Methodological issues in measuring health disparities". *Vital and health statistics. Series 2, Data evaluation and methods research* 141, page 1 (cited on page 23).

R Jeanne Ruiz and Kay C Avant (2005). "Effects of maternal prenatal stress on infant outcomes: a synthesis of the literature". *Advances in Nursing Science* 28.4, pages 345–355 (cited on page 18).

Robert A Schoevers, DJH Deeg, W Van Tilburg, and ATF Beekman (2005). "Depression and generalized anxiety disorder: co-occurrence and longitudinal patterns in elderly patients". *The American Journal of Geriatric Psychiatry* (cited on pages 123, 146).

Meifen Wei, Daniel W Russell, and Robyn A Zakalik (2005). "Adult attachment, social self-efficacy, self-disclosure, loneliness, and subsequent depression for freshman college students: A longitudinal study." *Journal of counseling psychology* 52.4, page 602 (cited on page 147).

BIBLIOGRAPHY

David R Williams (2005). "Patterns and causes of disparities in health". *Policy challenges in modern health care*, pages 115–134 (cited on page 24).

Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira (2006). "Analysis of representations for domain adaptation". *Advances in neural information processing systems* 19 (cited on pages 42, 99).

John Blitzer, Ryan McDonald, and Fernando Pereira (2006). "Domain adaptation with structural correspondence learning". *Proceedings of the 2006 conference on empirical methods in natural language processing*, pages 120–128 (cited on pages 50, 273).

Paula Braveman (2006). "Health disparities and health equity: concepts and measurement". *Annu. Rev. Public Health* 27, pages 167–194 (cited on pages 3, 7, 20, 23).

Edith Chen, Andrew D Martin, and Karen A Matthews (2006). "Understanding health disparities: The role of race and socioeconomic status in children's health". *American journal of public health* 96.4, pages 702–708 (cited on page 40).

Hal Daume III and Daniel Marcu (2006). "Domain adaptation for statistical classifiers". *Journal of artificial Intelligence research* 26, pages 101–126 (cited on page 161).

National Institutes of Health et al. (2006). *Addressing disparities: the NIH program of action. What are health disparities* (cited on page 20).

BIBLIOGRAPHY

Serguei VS Pakhomov, James D Buntrock, and Christopher G Chute (2006). "Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques". *Journal of the American Medical Informatics Association* 13.5, pages 516–525 (cited on page 4).

Madison Powers and Ruth R Faden (2006). *Social justice: The moral foundations of public health and health policy*. Oxford University Press, USA (cited on page 4).

Masashi Sugiyama, Benjamin Blankertz, Matthias Krauledat, Guido Dornhege, and Klaus-Robert Müller (2006). "Importance-weighted cross-validation for covariate shift". *Joint Pattern Recognition Symposium*. Springer, pages 354–363 (cited on page 35).

Noori Akhtar-Danesh and Janet Landeen (2007). "Relation between depression and sociodemographic factors". *International journal of mental health systems* 1, pages 1–9 (cited on page 28).

John Blitzer, Mark Dredze, and Fernando Pereira (2007). "Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification". *Proceedings of the 45th annual meeting of the association of computational linguistics*, pages 440–447 (cited on page 50).

BIBLIOGRAPHY

Kathleen M Colleran, Allyson Richards, and Keri Shafer (2007). "Disparities in cardiovascular disease risk and treatment: demographic comparison". *Journal of Investigative Medicine* 55.8, pages 415–422 (cited on page 28).

Hal Daumé III (2007). "Frustratingly Easy Domain Adaptation". *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 256–263 (cited on page 50).

Alexander R Green, Dana R Carney, Daniel J Pallin, Long H Ngo, Kristal L Raymond, Lisa I Iezzoni, and Mahzarin R Banaji (2007). "Implicit bias among physicians and its prediction of thrombolysis decisions for black and white patients". *Journal of general internal medicine* 22.9, pages 1231–1238 (cited on page 248).

Jing Jiang and ChengXiang Zhai (2007). "Instance weighting for domain adaptation in NLP". ACL (cited on page 50).

Derald Wing Sue, Christina M Capodilupo, Gina C Torino, Jennifer M Bucceri, Aisha Holder, Kevin L Nadal, and Marta Esquilin (2007). "Racial microaggressions in everyday life: implications for clinical practice." *American psychologist* 62.4, page 271 (cited on pages 250, 252).

Masashi Sugiyama, Matthias Krauledat, and Klaus-Robert Müller (2007). "Covariate shift adaptation by importance weighted cross validation." *Journal of Machine Learning Research* 8.5 (cited on page 276).

BIBLIOGRAPHY

Jaco Voorham and Petra Denig (2007). “Computerized extraction of information on the quality of diabetes care from free text in electronic patient records of general practitioners”. *Journal of the American Medical Informatics Association* 14.3, pages 349–354 (cited on page 197).

Yariv Yogev and Oded Langer (2007). “Spontaneous preterm delivery and gestational diabetes: the impact of glycemic control”. *Archives of gynecology and obstetrics* 276, pages 361–365 (cited on page 18).

R Michael Bagby, Lena C Quilty, and Andrew C Ryder (2008). “Personality and depression”. *The Canadian Journal of Psychiatry* (cited on page 146).

Jing Jiang (2008). “A literature survey on domain adaptation of statistical classifiers”. *URL: http://sifaka. cs. uiuc. edu/jiang4/domainadaptation/survey* 3, pages 1–12 (cited on page 50).

Pascal Lavergne and Valentin Patilea (2008). “Breaking the curse of dimensionality in nonparametric testing”. *Journal of Econometrics* 143.1, pages 103–122 (cited on page 41).

Sara L Thier, KS Yu-Isenberg, BF Leas, CR Cantrell, S DeBussey, NI Goldfarb, and David B Nash (2008). “In chronic disease, nationwide data show poor adherence by patients to medication and by physicians to guidelines”. *Manag Care* 17.2, pages 48–52 (cited on page 283).

BIBLIOGRAPHY

Jules Angst, Alex Gamma, Wulf Rössler, Vladeta Ajdacic, and Daniel N Klein (2009). "Long-term depression versus episodic major depression: results from the prospective Zurich study of a community sample". *Journal of affective disorders* 115.1-2 (cited on pages 128, 187).

Varun Chandola, Arindam Banerjee, and Vipin Kumar (2009). "Anomaly detection: A survey". *ACM computing surveys (CSUR)* 41.3, pages 1–58 (cited on page 41).

Jesse R Cougle, Meghan E Keough, Christina J Riccardi, and Natalie Sachs-Ericsson (2009). "Anxiety disorders and suicidality in the National Comorbidity Survey-Replication". *Journal of psychiatric research* 43.9, pages 825–829 (cited on page 66).

Steven A Greenberg (2009). "How citation distortions create unfounded authority: analysis of a citation network". *Bmj* 339 (cited on page 155).

Jeffrey H Kahn and Angela M Garrison (2009). "Emotional self-disclosure and emotional avoidance: Relations with symptoms of depression and anxiety." *Journal of counseling psychology* 56.4, page 573 (cited on page 147).

Benjamin B Lahey (2009). "Public health significance of neuroticism." *American Psychologist* 64.4, page 241 (cited on page 146).

BIBLIOGRAPHY

Alex J Mitchell, Amol Vaze, and Sanjay Rao (2009). "Clinical diagnosis of depression in primary care: a meta-analysis". *The Lancet* 374.9690, pages 609–619 (cited on page 84).

Jenelle R Shanley, Deborah Shropshire, and Barbara L Bonner (2009). "To report or not report: A physician's dilemma". *AMA Journal of Ethics* 11.2, pages 141–145 (cited on page 253).

Timothy Waidmann (2009). *Estimating the cost of racial and ethnic health disparities*. Urban Institute Washington, DC (cited on page 3).

Cynthia A Claassen, Thomas Carmody, Sunita M Stewart, Robert M Bossarte, Gregory L Larkin, Wayne A Woodward, and Madhukar H Trivedi (2010). "Effect of 11 September 2001 terrorist attacks in the USA on suicide in areas surrounding the crash sites". *The British Journal of Psychiatry* 196.5, pages 359–364 (cited on page 164).

Hal Daumé III, Abhishek Kumar, and Avishek Saha (2010). "Frustratingly easy semi-supervised domain adaptation". *Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing*, pages 53–59 (cited on page 50).

Mark Dredze, Tim Oates, and Christine Piatko (2010). "We're not in kansas anymore: detecting domain changes in streams". *EMNLP* (cited on pages 162, 167).

BIBLIOGRAPHY

Ahmed M Elmisery and Huaiguo Fu (2010). "Privacy preserving distributed learning clustering of healthcare data using cryptography protocols". *2010 IEEE 34th Annual Computer Software and Applications Conference Workshops* (cited on page 72).

Alan J Gelenberg (2010). "The prevalence and impact of depression". *The Journal of clinical psychiatry* (cited on page 168).

William L Jarrold, Bart Peintner, Eric Yeh, Ruth Krasnow, Harold S Javitz, and Gary E Swan (2010). "Language analytics for assessing brain health: Cognitive impairment, depression and pre-symptomatic Alzheimer's disease". *International Conference on Brain Informatics*. Springer, pages 299–307 (cited on page 81).

Cynthia M Jones (2010). "The moral problem of health disparities". *American journal of public health* 100.S1, S47–S51 (cited on page 4).

Brendan O'Connor, Michel Krieger, and David Ahn (2010). "Tweetmotif: Exploratory search and topic summarization for twitter". *Fourth International AAAI Conference on Weblogs and Social Media* (cited on pages 92, 174).

Radim Řehůřek and Petr Sojka (2010). "Software Framework for Topic Modelling with Large Corpora". English. *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. http://is.muni.cz/publication/884893/en. Valletta, Malta: ELRA, pages 45–50 (cited on page 263).

BIBLIOGRAPHY

Derald Wing Sue (2010). *Microaggressions in everyday life: Race, gender, and sexual orientation*. John Wiley & Sons (cited on page 268).

Yla R Tausczik and James W Pennebaker (2010). "The psychological meaning of words: LIWC and computerized text analysis methods". *Journal of language and social psychology* 29.1, pages 24–54 (cited on pages 93, 106).

Giovanni Tripepi, Kitty J Jager, Friedo W Dekker, and Carmine Zoccali (2010). "Selection bias and information bias in clinical research". *Nephron Clinical Practice* 115.2, pages c94–c99 (cited on page 34).

Judith Bell and Mary M Lee (2011). "Why place and race matter: Impacting health through a focus on race and place". *Oakland, CA: PolicyLink* (cited on page 248).

Paula A Braveman, Shiriki Kumanyika, Jonathan Fielding, Thomas LaVeist, Luisa N Borrell, Ron Manderscheid, and Adewale Troutman (2011). "Health disparities and health equity: the issue is justice". *American journal of public health* 101.S1, S149–S155 (cited on page 7).

Armin Brott, Adam Dougherty, Scott T Williams, Janet H Matope, Ana Fadich, and Muguleta Taddelle (2011). "The economic burden shouldered by public and private entities as a consequence of health disparities between men and women". *American journal of men's health* 5.6, pages 528–539 (cited on page 4).

BIBLIOGRAPHY

Wendy W Chapman, Prakash M Nadkarni, Lynette Hirschman, Leonard W D'avolio, Guergana K Savova, and Ozlem Uzuner (2011). *Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions* (cited on page 154).

Kathryn H Gordon, Konrad Bresin, Joseph Dombeck, Clay Routledge, and Joseph A Wonderlich (2011). "The impact of the 2009 Red River Flood on interpersonal risk factors for suicide". *Crisis* (cited on page 164).

Klaus Krippendorff (2011). "Computing Krippendorff's alpha-reliability" (cited on page 137).

Jean-Pierre Lépine and Mike Briley (2011). "The increasing burden of depression". *Neuropsychiatric disease and treatment* (cited on page 169).

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. (2011). "Scikit-learn: Machine learning in Python". *JMLR* (cited on pages 175, 176, 263).

Alan Ritter, Sam Clark, Oren Etzioni, et al. (2011). "Named entity recognition in tweets: an experimental study". *Proceedings of the conference on empirical methods in natural language processing*. Association for Computational Linguistics, pages 1524–1534 (cited on page 113).

Konstantinos Sechidis, Grigorios Tsoumakas, and Ioannis Vlahavas (2011). "On the stratification of multi-label data". *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2011, Athens, Greece, September 5-9, 2011, Proceedings, Part III 22*. Springer, pages 145–158 (cited on page 233).

Michelle Van Ryn, Diana J Burgess, John F Dovidio, Sean M Phelan, Somnath Saha, Jennifer Malat, Joan M Griffin, Steven S Fu, and Sylvia Perry (2011). "The impact of racism on clinician cognition, behavior, and clinical decision making". *Du Bois review: social science research on race* 8.1, pages 199–218 (cited on page 248).

Jerry J Vaske (2011). "Advantages and disadvantages of internet surveys: Introduction to the special issue". *Human Dimensions of Wildlife* (cited on page 190).

Stephen Wu and Hongfang Liu (2011). "Semantic characteristics of NLP-extracted concepts in clinical notes vs. biomedical literature". *AMIA Annual Symposium Proceedings*. Volume 2011. American Medical Informatics Association, page 1550 (cited on page 202).

Faiyaz Al Zamal, Wendy Liu, and Derek Ruths (2012). "Homophily and latent attribute inference: Inferring latent attributes of twitter users from neighbors". *ICWSM* (cited on page 161).

John Blosnich and Robert Bossarte (2012). "Drivers of disparity: Differences in socially based risk factors of self-injurious and suicidal behaviors among sexual minority

college students". *Journal of American College Health* 60.2, pages 141–149 (cited on page 3).

Hau-wen Chang, Dongwon Lee, Mohammed Eltaher, and Jeongkyu Lee (2012a). "@ Phillies tweeting from Philly? Predicting Twitter user locations with spatial word usage". *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. IEEE, pages 111–118 (cited on pages 93, 104).

Sung Man Chang, Jin-Pyo Hong, and Maeng Je Cho (2012b). "Economic burden of depression in South Korea". *Social psychiatry and psychiatric epidemiology* (cited on page 169).

Lisa A Cooper, Debra L Roter, Kathryn A Carson, Mary Catherine Beach, Janice A Sabin, Anthony G Greenwald, and Thomas S Inui (2012). "The associations of clinicians' implicit attitudes about race with medical visit communication and patient ratings of interpersonal care". *American journal of public health* 102.5, pages 979–987 (cited on page 248).

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola (2012). "A kernel two-sample test". *The Journal of Machine Learning Research* 13.1, pages 723–773 (cited on page 44).

Chin-Lan Huang, Cindy K Chung, Natalie Hui, Yi-Cheng Lin, Yi-Tai Seih, Ben CP Lam, Wei-Chuan Chen, Michael H Bond, and James W Pennebaker (2012). "The

development of the Chinese linguistic inquiry and word count dictionary." *Chinese Journal of Psychology* (cited on page 85).

Zurida Ishak, Azizah Jaafar, and Azlina Ahmad (2012). "Interface design for cultural differences". *Procedia-Social and Behavioral Sciences* 65, pages 793–801 (cited on page 54).

Hongfang Liu, Stephen T Wu, Dingcheng Li, Siddhartha Jonnalagadda, Sunghwan Sohn, Kavishwar Wagholikar, Peter J Haug, Stanley M Huff, and Christopher G Chute (2012). "Towards a semantic lexicon for clinical natural language processing". *AMIA Annual Symposium Proceedings*. Volume 2012. American Medical Informatics Association, page 568 (cited on page 202).

Marco Lui and Timothy Baldwin (2012). "langid. py: An off-the-shelf language identification tool". *ACL 2012 system demonstrations* (cited on pages 126, 171, 172).

Jose G Moreno-Torres, Troy Raeder, Rocío Alaiz-Rodríguez, Nitesh V Chawla, and Francisco Herrera (2012). "A unifying view on dataset shift in classification". *Pattern recognition* 45.1, pages 521–530 (cited on page 34).

Minsu Park, Chiyoung Cha, and Meeyoung Cha (2012). "Depressive moods of users portrayed in Twitter". *Proceedings of the ACM SIGKDD Workshop on healthcare informatics (HI-KDD)*. Volume 2012, pages 1–8 (cited on pages 69, 82).

BIBLIOGRAPHY

David R Williams (2012). "Miles to go before we sleep: Racial inequities in health". *Journal of health and social behavior* 53.3, pages 279–295 (cited on page 23).

Huan Xu and Shie Mannor (2012). "Robustness and generalization". *Machine learning* 86.3, pages 391–423 (cited on page 48).

Joanne WY Yau, Sophie L Rogers, Ryo Kawasaki, Ecosse L Lamoureux, Jonathan W Kowalski, Toke Bek, Shih-Jen Chen, Jacqueline M Dekker, Astrid Fletcher, Jakob Grauslund, et al. (2012). "Global prevalence and major risk factors of diabetic retinopathy". *Diabetes care* 35.3, pages 556–564 (cited on page 198).

American Psychiatric Association APA (2013). *Diagnostic and statistical manual of mental disorders (DSM-5®)*. American Psychiatric Pub (cited on pages 61, 113, 133).

John W Ayers, Benjamin M Althouse, Jon-Patrick Allem, J Niels Rosenquist, and Daniel E Ford (2013). "Seasonality in seeking mental health information on Google". *American journal of preventive medicine* 44.5, pages 520–525 (cited on page 62).

Timothy Baldwin, Paul Cook, Marco Lui, Andrew MacKinlay, and Li Wang (2013). "How noisy social media text, how diffrnt social media sources?" *Proceedings of the sixth international joint conference on natural language processing*, pages 356–364 (cited on pages 47, 103).

BIBLIOGRAPHY

Stephanie L Budge, Jill L Adelson, and Kimberly AS Howard (2013). "Anxiety and depression in transgender individuals: the roles of transition status, loss, social support, and coping." *Journal of consulting and clinical psychology* 81.3, page 545 (cited on page 105).

Munmun De Choudhury, Scott Counts, and Eric Horvitz (2013a). "Social media as a measurement tool of depression in populations". *5th annual ACM web science conference* (cited on page 170).

Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz (2013b). "Predicting depression via social media". *Seventh international AAAI conference on weblogs and social media* (cited on pages 58, 82).

Benoit Frenay and Michel Verleysen (2013). "Classification in the presence of label noise: a survey". *IEEE transactions on neural networks and learning systems* (cited on page 128).

Jacquie R Halladay, Katrina E Donahue, Alan L Hinderliter, Doyle M Cummings, Crystal W Cene, Cassie L Miller, Beverly A Garcia, Jim Tillman, Darren DeWalt, and Heart Healthy Lenoir Research Team (2013). "The heart healthy lenoir project-an intervention to reduce disparities in hypertension control: study protocol". *BMC health services research* 13, pages 1–11 (cited on page 24).

BIBLIOGRAPHY

Keith Hawton, Carolina Casañas i Comabella, Camilla Haw, and Kate Saunders (2013). "Risk factors for suicide in individuals with depression: a systematic review". *Journal of affective disorders* 147.1-3, pages 17–28 (cited on page 86).

Danielle M McCarthy, Barbara A Buckley, Kirsten G Engel, Victoria E Forth, James G Adams, and Kenzie A Cameron (2013). "Understanding patient–provider conversations: what are we talking about?" *Academic Emergency Medicine* 20.5, pages 441–448 (cited on page 301).

Wendy Berry Mendes and Katrina Koslov (2013). "Brittle smiles: positive biases toward stigmatized and outgroup targets." *Journal of Experimental Psychology: General* 142.3, page 923 (cited on page 282).

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean (2013a). "Efficient estimation of word representations in vector space". *arXiv preprint arXiv:1301.3781* (cited on page 175).

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean (2013b). "Distributed representations of words and phrases and their compositionality". *NeurIPS* (cited on pages 174, 175, 263).

Mark Myslín, Shu-Hong Zhu, Wendy Chapman, and Mike Conway (2013). "Using twitter to examine smoking behavior and perceptions of emerging tobacco products". *JMIR* (cited on page 160).

BIBLIOGRAPHY

H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Stephanie M Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin EP Seligman, et al. (2013). "Personality, gender, and age in the language of social media: The open-vocabulary approach". *PloS one* 8.9, e73791 (cited on page 87).

Kun Zhang, Bernhard Schölkopf, Krikamol Muandet, and Zhikun Wang (2013). "Domain adaptation under target and conditional shift". *International conference on machine learning*. Pmlr, pages 819–827 (cited on page 50).

Elizabeth A Bell, Lucila Ohno-Machado, and M Adela Grando (2014). "Sharing my health data: a survey of data sharing preferences of healthy individuals". *AMIA annual symposium proceedings*. Volume 2014. American Medical Informatics Association, page 1699 (cited on page 76).

Joel A Blanco and Lynn A Barnett (2014). "The effects of depression on leisure: Varying relationships between enjoyment, sociability, participation, and desired outcomes in college students". *Leisure Sciences* 36.5, pages 458–478 (cited on page 105).

Glen Coppersmith, Mark Dredze, and Craig Harman (2014a). "Quantifying Mental Health Signals in Twitter". *CLPsych* (cited on pages 70, 86, 121, 150, 161).

BIBLIOGRAPHY

Glen Coppersmith, Craig Harman, and Mark Dredze (2014b). "Measuring Post Traumatic Stress Disorder in Twitter". *ICWSM* (cited on page 70).

Munmun De Choudhury and Sushovan De (2014). "Mental health discourse on reddit: Self-disclosure, social support, and anonymity". *Eighth international AAAI conference on weblogs and social media* (cited on pages 90, 126, 172).

João Gama, Indrė Žliobaitė, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia (2014). "A survey on concept drift adaptation". *ACM computing surveys (CSUR)* 46.4, pages 1–37 (cited on page 42).

Jonathan Gratch, Ron Artstein, Gale M. Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood, Jill Boberg, David DeVault, Stacy Marsella, David R. Traum, Albert A. Rizzo, and Louis-Philippe Morency (2014). "The Distress Analysis Interview Corpus of human and computer interviews". *LREC* (cited on page 62).

Jared Jashinsky, Scott H. Burton, Carl Lee Hanson, Joshua H. West, Christophe G. Giraud-Carrier, Michael D Barnes, and Trenton Argyle (2014). "Tracking suicide risk factors through Twitter in the US." *Crisis* 35 1, pages 51–9 (cited on page 70).

Yungchang Ku, Chaochang Chiu, Yulei Zhang, Hsinchun Chen, and Handsome Su (2014). "Text mining self-disclosing health information for public health

service". *Journal of the Association for Information Science and Technology* 65.5, pages 928–947 (cited on page 75).

Huijie Lin, Jia Jia, Quan Guo, Yuanyuan Xue, Qi Li, Jie Huang, Lianhong Cai, and Ling Feng (2014). "User-level psychological stress detection from social media using deep neural network". *22nd ACM international conference on Multimedia* (cited on page 70).

Jeffrey Pennington, Richard Socher, and Christopher D Manning (2014). "Glove: Global vectors for word representation". *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543 (cited on pages 93, 104).

Chris Poulin, Brian Shiner, Paul Thompson, Linas Vepstas, Yinong Young-Xu, Benjamin Goertzel, Bradley Watts, Laura Flashman, and Thomas McAllister (2014). "Predicting the risk of suicide by analyzing the text of clinical notes". *PloS one* 9.1, e85733 (cited on page 57).

H. Andrew Schwartz, Johannes Eichstaedt, Margaret L. Kern, Gregory Park, Maarten Sap, David Stillwell, Michal Kosinski, and Lyle Ungar (2014). "Towards Assessing Changes in Degree of Depression through Facebook". *CLPsych* (cited on page 69).

BIBLIOGRAPHY

Isis H Settles and Nicole T Buchanan (2014). "Multiple groups, multiple identities, and intersectionality". *The Oxford handbook of multicultural identity* 1, pages 160–180 (cited on page 40).

Taruna Sharma, Juhi Kalra, D Dhasmana, and Harish Basera (2014). "Poor adherence to treatment: A major challenge in diabetes". *Age (Yrs)* 31.40, page 40 (cited on page 283).

Amy E Sickel, Jason D Seacat, and Nina A Nabors (2014). "Mental health stigma update: A review of consequences". *Advances in Mental Health* 12.3, pages 202–215 (cited on page 76).

Lei Zhang, Xiaolei Huang, Tianli Liu, Ang Li, Zhenxiang Chen, and Tingshao Zhu (2014). "Using linguistic features to estimate suicide probability of Chinese microblog users". *International Conference on Human Centered Computing*. Springer, pages 549–559 (cited on page 58).

Borwin Bandelow and Sophie Michaelis (2015). "Epidemiology of anxiety disorders in the 21st century". *Dialogues in clinical neuroscience* 17.3, page 327 (cited on page 66).

Victoria Betton, Rohan Borschmann, Mary Docherty, Stephen Coleman, Mark Brown, and Claire Henderson (2015). "The role of social media in reducing stigma and discrimination". *The British Journal of Psychiatry* (cited on page 147).

BIBLIOGRAPHY

Lucy Bowes, Rebecca Carnegie, Rebecca Pearson, Becky Mars, Lucy Biddle, Barbara Maughan, Glyn Lewis, Charles Fernyhough, and Jon Heron (2015). "Risk of depression and self-harm in teenagers identifying with goth subculture: a longitudinal cohort study". *The Lancet Psychiatry* 2.9, pages 793–800 (cited on page 105).

Igor Brigadir, Derek Greene, and Pádraig Cunningham (2015). "Analyzing discourse communities with distributional semantic models". *Proceedings of the ACM Web Science Conference*, pages 1–10 (cited on pages 98, 161).

Pete Burnap and Matthew L Williams (2015). "Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making". *Policy & internet* (cited on page 160).

Luca Canzian and Mirco Musolesi (2015). "Trajectories of depression: unobtrusive monitoring of depressive states by means of smartphone mobility traces analysis". *2015 ACM international joint conference on pervasive and ubiquitous computing* (cited on page 123).

Glen Coppersmith, Mark Dredze, Craig Harman, and Kristy Hollingshead (2015a). "From ADHD to SAD: Analyzing the Language of Mental Health on Twitter through Self-Reported Diagnoses". *CLPsych* (cited on pages 66, 70).

BIBLIOGRAPHY

Glen Coppersmith, Mark Dredze, Craig Harman, and Kristy Hollingshead (2015b). "From ADHD to SAD: Analyzing the language of mental health on Twitter through self-reported diagnoses". *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 1–10 (cited on page 81).

Glen Coppersmith, Mark Dredze, Craig Harman, Kristy Hollingshead, and Margaret Mitchell (2015c). "CLPsych 2015 Shared Task: Depression and PTSD on Twitter". *CLPsych* (cited on pages 58, 64, 70, 72, 74).

Glen Coppersmith, Mark Dredze, Craig Harman, Kristy Hollingshead, and Margaret Mitchell (2015d). "CLPsych 2015 shared task: Depression and PTSD on Twitter". *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 31–39 (cited on pages 73, 82, 86, 87, 122, 125, 136, 169).

Nicholas Cummins, Stefan Scherer, Jarek Krajewski, Sebastian Schnieder, Julien Epps, and Thomas F Quatieri (2015). "A review of depression and suicide risk assessment using speech analysis". *Speech Communication* 71, pages 10–49 (cited on page 84).

Munmun De Choudhury (2015). "Anorexia on tumblr: A characterization study". *5th international conference on digital health 2015* (cited on page 70).

BIBLIOGRAPHY

Geli Fei and Bing Liu (2015). "Social media text classification under negative covariate shift". *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 2347–2356 (cited on page 43).

Christian Fuchs (2015). *Culture and economy in the age of social media*. Routledge (cited on page 57).

Guy M Goodwin (2015). "The overlap between anxiety, depression, and obsessive-compulsive disorder". *Dialogues in clinical neuroscience* (cited on page 146).

Garth Graham (2015). "Disparities in cardiovascular disease risk in the United States". *Current cardiology reviews* 11.3, pages 238–245 (cited on page 38).

Melissa W Graham, Elizabeth J Avery, and Sejin Park (2015). "The role of social media in local government crisis communications". *Public Relations Review* (cited on page 57).

Jennifer S Haas, Jeffrey A Linder, Elyse R Park, Irina Gonzalez, Nancy A Rigotti, Elissa V Klinger, Emily Z Kontos, Alan M Zaslavsky, Phyllis Brawarsky, Lucas X Marinacci, et al. (2015). "Proactive tobacco cessation outreach to smokers of low socioeconomic status: a randomized clinical trial". *JAMA internal medicine* 175.2, pages 218–226 (cited on page 24).

BIBLIOGRAPHY

William J Hall, Mimi V Chapman, Kent M Lee, Yesenia M Merino, Tainayah W Thomas, B Keith Payne, Eugenia Eng, Steven H Day, and Tamera Coyne-Beasley (2015). "Implicit racial/ethnic bias among health care professionals and its influence on health care outcomes: a systematic review". *American journal of public health* 105.12, e60–e76 (cited on page 248).

Bernadette Davantes Heckman and Ashley Joi Britton (2015). "Headache in African Americans: an overlooked disparity". *Journal of the National Medical Association* 107.2, pages 39–45 (cited on page 21).

Guido W Imbens and Donald B Rubin (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press (cited on pages 129, 150).

Maciej Kopera, Hubert Suszek, Erin Bonar, Maciej Myszka, Bartłomiej Gmaj, Mark Ilgen, and Marcin Wojnar (2015). "Evaluating explicit and implicit stigma of mental illness in mental health professionals and medical students". *Community mental health journal* 51.5, pages 628–634 (cited on page 285).

Mrinal Kumar, Mark Dredze, Glen Coppersmith, and Munmun De Choudhury (2015). "Detecting changes in suicide content manifested in social media following celebrity suicides". *Proceedings of the 26th ACM conference on Hypertext & Social Media*, pages 85–94 (cited on pages 58, 70).

BIBLIOGRAPHY

Momin M Malik, Hemank Lamba, Constantine Nakos, and Jürgen Pfeffer (2015). "Population bias in geotagged tweets". *Ninth international AAAI conference on web and social media* (cited on page 112).

Danielle Mowery, Craig Bryan, and Mike Conway (2015). "Towards Developing an Annotation Scheme for Depressive Disorder Symptoms: A Preliminary Study using Twitter Data". *CLPsych* (cited on pages 66, 70).

Billie Murray and Susan McCrone (2015). "An integrative review of promoting trust in the patient–primary care provider relationship". *Journal of Advanced Nursing* 71.1, pages 3–23 (cited on page 301).

James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn (2015). *The development and psychometric properties of LIWC2015*. Technical report (cited on page 106).

Andrew Perrin (2015). "Social media usage". *Pew research center*, pages 52–68 (cited on page 57).

Rimma Pivovarov, Adler J Perotte, Edouard Grave, John Angiolillo, Chris H Wiggins, and Noémie Elhadad (2015). "Learning probabilistic phenotypes from heterogeneous EHR data". *Journal of biomedical informatics* 58, pages 156–165 (cited on page 4).

BIBLIOGRAPHY

Daniel Preoţiuc-Pietro, Johannes Eichstaedt, Gregory Park, Maarten Sap, Laura Smith, Victoria Tobolsky, H Andrew Schwartz, and Lyle Ungar (2015). "The role of personality, age, and gender in tweeting about mental illness". *Proceedings of the 2nd workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality*, pages 21–30 (cited on pages 84, 86, 114, 123, 150).

Martin Shelton, Katherine Lo, and Bonnie Nardi (2015). "Online media forums as separate social lives: A qualitative study of disclosure within and beyond Reddit". *iConference 2015 Proceedings* (cited on page 112).

Maxwell J Smith (2015). "Health equity in public health: clarifying our commitment". *Public Health Ethics* 8.2, pages 173–184 (cited on page 19).

Shiliang Sun, Honglei Shi, and Yuanbin Wu (2015). "A survey of multi-source domain adaptation". *Information Fusion* 24, pages 84–92 (cited on page 50).

Svitlana Volkova, Yoram Bachrach, Michael Armstrong, and Vijay Sharma (2015). "Inferring latent user properties from texts published in social media". *AAAI* (cited on page 161).

Yazhe Wang, Jamie Callan, and Baihua Zheng (2015). "Should we use the sample? Analyzing datasets sampled from Twitter's stream API". *ACM Transactions on the Web (TWEB)* 9.3, pages 1–23 (cited on page 169).

BIBLIOGRAPHY

Wei-Qi Wei and Joshua C Denny (2015). “Extracting research-quality phenotypes from electronic health records to support precision medicine”. *Genome medicine* 7, pages 1–14 (cited on page 4).

Yonghui Wu, Jun Xu, Yaoyun Zhang, and Hua Xu (2015). “Clinical abbreviation disambiguation using neural word embeddings”. *Proceedings of BioNLP 15*, pages 171–176 (cited on page 202).

Patricia A Cavazos-Rehg, Melissa J Krauss, Shaina Sowles, Sarah Connolly, Carlos Rosas, Meghana Bharadwaj, and Laura J Bierut (2016). “A content analysis of depression-related tweets”. *Computers in human behavior* 54, pages 351–357 (cited on page 82).

Stevie Chancellor, Zhiyuan Lin, Erica L Goodman, Stephanie Zerwas, and Munmun De Choudhury (2016a). “Quantifying and predicting mental illness severity in online pro-eating disorder communities”. *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, pages 1171–1184 (cited on pages 65, 124).

Stevie Chancellor, Tanushree Mitra, and Munmun De Choudhury (2016b). “Recovery amid pro-anorexia: Analysis of recovery in social media”. *CHI* (cited on page 70).

BIBLIOGRAPHY

Dongho Choi, Ziad Matni, and Chirag Shah (2016). "What social media data should i use in my research?: A comparative analysis of twitter, youtube, reddit, and the new york times comments". *ASIS&T* (cited on page 65).

Mike Conway and Daniel O'Connor (2016). "Social media, big data, and mental health: current advances and ethical implications". *Current opinion in psychology* (cited on pages 152, 189).

Glen Coppersmith, Kim Ngo, Ryan Leary, and Anthony Wood (2016). "Exploratory Analysis of Social Media Prior to a Suicide Attempt". *CLPsych* (cited on page 70).

Munmun De Choudhury, Emre Kiciman, Mark Dredze, Glen Coppersmith, and Mrinal Kumar (2016). "Discovering shifts to suicidal ideation from mental health content in social media". *2016 CHI conference on human factors in computing systems* (cited on pages 67, 70, 124, 150).

Mark Dredze, Miles Osborne, and Prabhanjan Kambadur (2016). "Geolocation for twitter: Timing matters". *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1064–1069 (cited on page 98).

Anthony W Flores, Kristin Bechtel, and Christopher T Lowenkamp (2016). "False positives, false negatives, and false analyses: A rejoinder to machine bias: There's

software used across the country to predict future criminals. and it's biased against blacks". *Fed. Probation* 80, page 38 (cited on page 30).

George Gkotsis, Anika Oellrich, Tim Hubbard, Richard Dobson, Maria Liakata, Sumithra Velupillai, and Rina Dutta (2016). "The language of mental health problems in social media". *CLPsych* (cited on pages 64, 70).

William L Hamilton, Jure Leskovec, and Dan Jurafsky (2016). "Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change". *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501 (cited on pages 167, 188, 298).

Butool Hisam, Cheryl K Zogg, Muhammad A Chaudhary, Ammar Ahmed, Hammad Khan, Shalini Selvarajah, Maya J Torain, Navin R Changoor, and Adil H Haider (2016). "From understanding to action: interventions for surgical disparities". *journal of surgical research* 200.2, pages 560–578 (cited on page 30).

Kelly M Hoffman, Sophie Trawalter, Jordan R Axt, and M Norman Oliver (2016). "Racial bias in pain assessment and treatment recommendations, and false beliefs about biological differences between blacks and whites". *Proceedings of the National Academy of Sciences* 113.16, pages 4296–4301 (cited on page 248).

BIBLIOGRAPHY

John W Jackson, David R Williams, and Tyler J VanderWeele (2016). "Disparities at the intersection of marginalized groups". *Social psychiatry and psychiatric epidemiology* 51, pages 1349–1359 (cited on page 3).

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark (2016). "MIMIC-III, a freely accessible critical care database". *Scientific data* 3.1, pages 1–9 (cited on pages 255, 265).

LaGarrett J King and Prentice T Chandler (2016). "From non-racism to anti-racism in social studies teacher education: Social studies and racial pedagogical content knowledge". *Rethinking social studies teacher education in the twenty-first century*, pages 3–21 (cited on page 9).

Wouter M Kouw, Laurens JP Van Der Maaten, Jesse H Krijthe, and Marco Loog (2016). "Feature-level domain adaptation". *Journal of Machine Learning Research* 17.171, pages 1–32 (cited on page 51).

Giwoong Lee, Eunho Yang, and Sung Hwang (2016). "Asymmetric multi-task learning based on task relatedness and loss". *International conference on machine learning*. PMLR, pages 230–238 (cited on page 240).

Jiwei Li, Will Monroe, and Dan Jurafsky (2016). "Understanding neural networks through representation erasure". *arXiv preprint arXiv:1612.08220* (cited on page 156).

Huijie Lin, Jia Jia, Liqiang Nie, Guangyao Shen, and Tat-Seng Chua (2016). "What Does Social Media Say about Your Stress?." *IJCAI*, pages 3775–3781 (cited on pages 66, 70).

Jean M McMahon and Kimberly Barsamian Kahn (2016). "Benevolent racism? The impact of target race on ambivalent sexism". *Group Processes & Intergroup Relations* 19.2, pages 169–183 (cited on page 252).

David N. Milne, Glen Pink, Ben Hachey, and Rafael A. Calvo (2016). "CLPsych 2016 Shared Task: Triaging content in online peer-support forums". *CLPsych* (cited on page 70).

Danielle L Mowery, Y Albert Park, Craig Bryan, and Mike Conway (2016). "Towards automatically classifying depressive symptoms from Twitter data for population health". *Proceedings of the Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media (PEOPLES)*, pages 182–191 (cited on pages 70, 81, 106).

BIBLIOGRAPHY

Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang (2016). "Abusive language detection in online user content". *Proceedings of the 25th international conference on world wide web*, pages 145–153 (cited on page 285).

Galen Panger (2016). "Reassessing the Facebook experiment: critical thinking about the validity of Big Data research". *Information, Communication & Society* 19.8, pages 1108–1126 (cited on page 65).

Michael J Paul, Abeed Sarker, John S Brownstein, Azadeh Nikfarjam, Matthew Scotch, Karen L Smith, and Graciela Gonzalez (2016). "Social media mining for public health monitoring and surveillance". *Biocomputing* (cited on pages 160, 189).

Hannah Rashkin, Sameer Singh, and Yejin Choi (2016). "Connotation Frames: A Data-Driven Investigation". *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 311–321 (cited on page 277).

Kirk Roberts (2016). "Assessing the corpus size vs. similarity trade-off for word embeddings in clinical NLP". *Proceedings of the clinical natural language processing workshop (ClinicalNLP)*, pages 54–63 (cited on pages 198, 286).

Haining Zhu, Joanna Colgan, Madhu Reddy, and Eun Kyoung Choe (2016). "Sharing patient-generated data in clinical practices: an interview study". *AMIA* (cited on page 72).

BIBLIOGRAPHY

Silvio Amir, Glen Coppersmith, Paula Carvalho, Mario J Silva, and Bryon C Wallace (2017). "Quantifying Mental Health from Social Media with Neural User Embeddings". *Machine Learning for Healthcare Conference*, pages 306–321 (cited on page 150).

Alina Baciu, Yamrot Negussie, Amy Geller, James N Weinstein, National Academies of Sciences, Engineering, and Medicine, et al. (2017a). "The root causes of health inequity". *Communities in action: Pathways to health equity*. National Academies Press (US) (cited on page 24).

Alina Baciu, Yamrot Negussie, Amy Geller, James N Weinstein, National Academies of Sciences, Engineering, and Medicine, et al. (2017b). "The state of health disparities in the United States". *Communities in action: Pathways to health equity*. National Academies Press (US) (cited on page 248).

Shrey Bagroy, Ponnurangam Kumaraguru, and Munmun De Choudhury (2017). "A social media based index of mental well-being in college campuses". *Proceedings of the 2017 CHI Conference on Human factors in Computing Systems*, pages 1634–1646 (cited on pages 70, 160).

Robert Bamler and Stephan Mandt (2017). "Dynamic word embeddings". *International conference on Machine learning*. PMLR, pages 380–389 (cited on page 298).

BIBLIOGRAPHY

Nicholas Beauchamp (2017). "Predicting and interpolating state-level polls using Twitter textual data". *American Journal of Political Science* (cited on page 160).

Adrian Benton, Glen Coppersmith, and Mark Dredze (2017a). "Ethical research protocols for social media health research". *First ACL Workshop on Ethics in Natural Language Processing* (cited on pages 58, 65, 115, 125, 153, 188).

Adrian Benton, Margaret Mitchell, and Dirk Hovy (2017b). "Multitask learning for mental health conditions with limited social media data". *Conference of the European Chapter of the Association for Computational Linguistics (EACL)*. Association for Computational Linguistics (cited on pages 73, 86, 88, 90, 94, 114, 120, 169).

Munmun De Choudhury and Emre Kiciman (2017). "The Language of Social Support in Social Media and its Effect on Suicidal Ideation Risk". *ICWSM* (cited on page 70).

Munmun De Choudhury, Sanket S Sharma, Tomaz Logar, Wouter Eekhout, and René Clausen Nielsen (2017). "Gender and cross-cultural differences in social media disclosures of mental illness". *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*, pages 353–369 (cited on pages 58, 68, 73, 74, 114).

BIBLIOGRAPHY

Orianna DeMasi, Konrad Kording, and Benjamin Recht (2017). "Meaningless comparisons lead to false optimism in medical machine learning". *PloS one* 12.9, e0184604 (cited on pages 94, 127, 187).

Elia J Duh, Jennifer K Sun, and Alan W Stitt (2017). "Diabetic retinopathy: current understanding, mechanisms, and treatment strategies". *JCI insight* 2.14 (cited on page 200).

Keith C Ferdinand, Kapil Yadav, Samar A Nasser, Helene D Clayton-Jeter, John Lewin, Dennis R Cryer, and Fortunato Fred Senatore (2017). "Disparities in hypertension and cardiovascular disease in blacks: the critical role of medication adherence". *The Journal of Clinical Hypertension* 19.10, pages 1015–1024 (cited on page 283).

Yonatan Geifman and Ran El-Yaniv (2017). "Selective classification for deep neural networks". *Advances in neural information processing systems* 30 (cited on page 41).

Su Golder, Shahd Ahmed, Gill Norman, and Andrew Booth (2017). "Attitudes toward the ethics of research using social media: a systematic review". *JMIR* (cited on page 189).

Sharath Chandra Guntuku, David B Yaden, Margaret L Kern, Lyle H Ungar, and Johannes C Eichstaedt (2017). "Detecting depression and mental illness on

social media: an integrative review". *Current Opinion in Behavioral Sciences* 18, pages 43–49 (cited on pages 59, 60, 123).

Nicholas S Holtzman et al. (2017). "A meta-analysis of correlations between depression and first person singular pronoun use". *Journal of Research in Personality* (cited on page 144).

Efsun Sarioglu Kayi, Mona Diab, Luca Pauselli, Michael Compton, and Glen Coppersmith (2017). "Predictive Linguistic Features of Schizophrenia". *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (* SEM 2017)*, pages 241–250 (cited on page 81).

Pang Wei Koh and Percy Liang (2017). "Understanding black-box predictions via influence functions". *International conference on machine learning*. PMLR, pages 1885–1894 (cited on page 133).

Liyan Liu, Neal H Shorstein, Laura B Amsden, and Lisa J Herrinton (2017a). "Natural language processing to ascertain two key variables from operative reports in ophthalmology". *Pharmacoepidemiology and drug safety* 26.4, pages 378–385 (cited on page 201).

Tong Liu, Qijin Cheng, Christopher M. Homan, and Vincent M. B. Silenzio (2017b). "Learning from various labeling strategies for suicide-related messages on social media: An experimental study". *CoRR* abs/1701.08796 (cited on page 84).

BIBLIOGRAPHY

David E Losada, Fabio Crestani, and Javier Parapar (2017). "eRISK 2017: CLEF lab on early risk prediction on the internet: experimental foundations". *CLEF* (cited on page 70).

Xingliang Mao, Fangfang Li, Yu Duan, and Hao Wang (2017). "Named entity recognition of electronic medical record in ophthalmology based on crf model". *2017 International conference on computer technology, electronics and communication (ICCTEC)*. IEEE, pages 785–788 (cited on page 201).

Thomas H McCoy Jr, Deanna C Chaukos, Leslie A Snapper, Kamber L Hart, Theodore A Stern, and Roy H Perlis (2017). "Enhancing delirium case definitions in electronic health records using clinical free text". *Psychosomatics* 58.2, pages 113–120 (cited on page 197).

Jude Mikal, Samantha Hurst, and Mike Conway (2017). "Investigating patient attitudes towards the use of social media data to augment depression diagnosis and treatment: a qualitative study". *fourth workshop on computational linguistics and clinical psychology—from linguistic signal to clinical reality* (cited on page 152).

Michelle Morales, Stefan Scherer, and Rivka Levitan (2017). "A cross-modal review of indicators for depression detection systems". *Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology—From Linguistic Signal to Clinical Reality*, pages 1–12 (cited on pages 81, 95).

BIBLIOGRAPHY

Danielle Mowery, Hilary Smith, Tyler Cheney, Greg Stoddard, Glen Coppersmith, Craig Bryan, and Mike Conway (2017). "Understanding depressive symptoms and psychosocial stressors on Twitter: a corpus-based study". *Journal of medical Internet research* 19.2 (cited on page 121).

Jan Mutchler, Yang Li, and Ping Xu (2017). "Living below the line: Economic insecurity and older Americans, racial and ethnic disparities in insecurity, 2016" (cited on page 22).

World Health Organization et al. (2017). *Depression and other common mental disorders: global health estimates*. Technical report. World Health Organization (cited on page 86).

Carla Parra Escartín, Wessel Reijers, Teresa Lynn, Joss Moorkens, Andy Way, and Chao-Hong Liu (2017). "Ethical considerations in NLP shared tasks". Association for Computational Linguistics (cited on page 154).

Michael J Paul and Mark Dredze (2017). "Social monitoring for public health". *Synthesis Lectures on Information Concepts, Retrieval, and Services* (cited on page 190).

Nanyun Peng and Mark Dredze (2017). "Multi-task Domain Adaptation for Sequence Tagging". *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 91–100 (cited on page 115).

BIBLIOGRAPHY

Ana Radovic, Theresa Gmelin, Bradley D Stein, and Elizabeth Miller (2017). "Depressed adolescents' positive and negative use of social media". *Journal of adolescence* 55, pages 5–15 (cited on page 147).

Andrew G Reece, Andrew J Reagan, Katharina LM Lix, Peter Sheridan Dodds, Christopher M Danforth, and Ellen J Langer (2017). "Forecasting the onset and course of mental illness with Twitter data". *Scientific reports* (cited on page 124).

Koustuv Saha and Munmun De Choudhury (2017). "Modeling stress with social media around incidents of gun violence on college campuses". *Proceedings of the ACM on Human-Computer Interaction* 1.CSCW, pages 1–27 (cited on pages 70, 160).

Anna Schmidt and Michael Wiegand (2017). "A survey on hate speech detection using natural language processing". *Proceedings of the fifth international workshop on natural language processing for social media*, pages 1–10 (cited on page 254).

Joan Serra, Ilias Leontiadis, Dimitris Spathis, Gianluca Stringhini, Jeremy Blackburn, and Athena Vakali (2017). "Class-based prediction errors to detect hate speech with out-of-vocabulary words". *Proceedings of the First Workshop on Abusive Language Online*, pages 36–40 (cited on page 103).

Tegjyot Singh Sethi and Mehmed Kantardzic (2017). "On the reliable detection of concept drift from streaming unlabeled data". *Expert Systems with Applications* 82, pages 77–99 (cited on page 46).

Guangyao Shen, Jia Jia, Liqiang Nie, Fuli Feng, Cunjun Zhang, Tianrui Hu, Tat-Seng Chua, and Wenwu Zhu (2017). "Depression Detection via Harvesting Social Media: A Multimodal Dictionary Learning Solution." *IJCAI* (cited on page 70).

Judy Hanwen Shen and Frank Rudzicz (2017). "Detecting anxiety through reddit". *CLPsych* (cited on page 70).

Sharon D Solomon, Emily Chew, Elia J Duh, Lucia Sobrin, Jennifer K Sun, Brian L VanderBeek, Charles C Wykoff, and Thomas W Gardner (2017). "Diabetic retinopathy: a position statement by the American Diabetes Association". *Diabetes care* 40.3, pages 412–418 (cited on pages 197, 199).

Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell (2017). "Adversarial discriminative domain adaptation". *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7167–7176 (cited on pages 51, 113).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin (2017). "Attention is all you need". *Advances in neural information processing systems* 30 (cited on page 193).

Nikhita Vedula and Srinivasan Parthasarathy (2017). "Emotional and linguistic cues of depression from social media". *Proceedings of the 2017 International Conference on Digital Health*, pages 127–136 (cited on pages 82, 92).

BIBLIOGRAPHY

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber (2017). "Understanding Abuse: A Typology of Abusive Language Detection Subtasks". *Proceedings of the First Workshop on Abusive Language Online*, pages 78–84 (cited on pages 250, 252, 262, 268).

Akkapon Wongkoblap, Miguel A Vadillo, and Vasa Curcin (2017). "Researching mental health disorders in the era of social media: systematic review". *JMIR* 19.6, e228 (cited on pages 59, 60).

Andrew Yates, Arman Cohan, and Nazli Goharian (2017). "Depression and Self-Harm Risk Assessment in Online Forums". *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2968–2978 (cited on pages 64, 70, 86, 88, 91).

Xinzhi Zhang, Eliseo J Pérez-Stable, Philip E Bourne, Emmanuel Peprah, O Kenrik Duru, Nancy Breen, David Berrigan, Fred Wood, James S Jackson, David WS Wong, et al. (2017). "Big data science: opportunities and challenges to address minority health and health disparities in the 21st century". *Ethnicity & disease* 27.2, page 95 (cited on page 27).

Adewole S Adamson and Avery Smith (2018). "Machine learning and health care disparities in dermatology". *JAMA dermatology* 154.11, pages 1247–1248 (cited on page 27).

BIBLIOGRAPHY

Maria Antoniak and David Mimno (2018). "Evaluating the stability of embedding-based word similarities". *Transactions of the Association for Computational Linguistics* 6, pages 107–119 (cited on page 298).

Alina Arseniev-Koehler, Sharon Mozgai, and Stefan Scherer (2018). "What type of happiness are you looking for?-A closer look at detecting mental health from language". *CLPsych* (cited on pages 58, 71, 84, 114, 149).

John W Ayers, Theodore L Caputi, Camille Nebeker, and Mark Dredze (2018). "Don't quote me: reverse identification of research participants in social media studies". *NPJ digital medicine* (cited on page 134).

Sivaraman Balakrishnan and Larry Wasserman (2018). "Hypothesis testing for high-dimensional multinomials: A selective review" (cited on page 41).

Paula Braveman, Elaine Arkin, Tracy Orleans, Dwayne Proctor, Julia Acker, and Alonzo Plough (2018). "What is health equity?" *Behavioral science & policy* 4.1, pages 1–14 (cited on pages 19, 280).

Joy Buolamwini and Timnit Gebru (2018). "Gender shades: Intersectional accuracy disparities in commercial gender classification". *Conference on fairness, accountability and transparency*. PMLR, pages 77–91 (cited on page 30).

BIBLIOGRAPHY

Stevie Chancellor, Andrea Hu, and Munmun De Choudhury (2018). "Norms matter: contrasting social support around behavior change in online weight loss communities". *CHI* (cited on page 70).

Arman Cohan, Bart Desmet, Andrew Yates, Luca Soldaini, Sean MacAvaney, and Nazli Goharian (2018). "SMHD: a Large-Scale Resource for Exploring Online Language Usage for Multiple Mental Health Conditions". *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1485–1497 (cited on pages 68, 70, 72, 87, 89, 91, 150, 169, 170).

Nicholas Dingwall and Christopher Potts (2018). "Mittens: an Extension of GloVe for Learning Domain-Specialized Representations". *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 212–217 (cited on page 202).

Julia Dressel and Hany Farid (2018). "The accuracy, fairness, and limits of predicting recidivism". *Science advances* 4.1, eaao5580 (cited on page 30).

Sarmistha Dutta, Jennifer Ma, and Munmun De Choudhury (2018). "Measuring the Impact of Anxiety on Online Social Interactions." *ICWSM*, pages 584–587 (cited on page 70).

BIBLIOGRAPHY

Johannes C Eichstaedt, Robert J Smith, Raina M Merchant, Lyle H Ungar, Patrick Crutchley, Daniel Preoţiuc-Pietro, David A Asch, and H Andrew Schwartz (2018). "Facebook language predicts depression in medical records". *Proceedings of the National Academy of Sciences* (cited on page 58).

Michele Filannino and Özlem Uzuner (2018). "Advancing the state of the art in clinical natural language processing through shared tasks". *Yearbook of medical informatics* 27.01, pages 184–192 (cited on page 154).

Devin Gaffney and J Nathan Matias (2018). "Caveat emptor, computational social science: Large-scale missing data in a widely-published Reddit corpus". *PloS one* (cited on page 171).

Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal (2018). "Explaining explanations: An overview of interpretability of machine learning". *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*. IEEE, pages 80–89 (cited on page 52).

Anna P Goddu, Katie J O'Conor, Sophie Lanzkron, Mustapha O Saheed, Somnath Saha, Monica E Peek, Carlton Haywood, and Mary Catherine Beach (2018). "Do words matter? Stigmatizing language and the transmission of bias in the medical record". *Journal of general internal medicine* 33.5, pages 685–691 (cited on pages 249, 253).

BIBLIOGRAPHY

Parampal S Grewal, Faraz Oloumi, Uriel Rubin, and Matthew TS Tennant (2018). "Deep learning in ophthalmology: a review". *Canadian Journal of Ophthalmology* 53.4, pages 309–313 (cited on page 200).

Jeremy Howard and Sebastian Ruder (2018). "Universal Language Model Fine-tuning for Text Classification". *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339 (cited on page 198).

Xiaolei Huang and Michael Paul (2018). "Examining temporality in document classification". *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 694–699 (cited on pages 98, 162).

Molly Ireland and Micah Iserman (2018). "Within and between-person differences in language used across anxiety support and neutral reddit communities". *CLPsych* (cited on pages 70, 84, 90).

Dan Iter, Jong Yoon, and Dan Jurafsky (2018). "Automatic detection of incoherent speech for diagnosing schizophrenia". *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 136–146 (cited on pages 62, 81).

Julia Ive, George Gkotsis, Rina Dutta, Robert Stewart, and Sumithra Velupillai (2018). "Hierarchical neural model with attention mechanisms for the classification of social media text related to mental health". *Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*. Association for Computational Linguistics. DOI: 10.18653/v1/W18-0607. URL: https://aclanthology.org/W18-0607 (cited on page 147).

Andrey Kutuzov, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal (2018). "Diachronic word embeddings and semantic shifts: a survey". *COLING* (cited on page 167).

Hongmin Li, Doina Caragea, Cornelia Caragea, and Nic Herndon (2018a). "Disaster response aided by tweet classification with a domain adaptation approach". *Journal of Contingencies and Crisis Management* 26.1, pages 16–27 (cited on page 115).

Yaoyiran Li, Rada Mihalcea, and Steven R Wilson (2018b). "Text-based detection and understanding of changes in mental health". *International Conference on Social Informatics* (cited on page 70).

Tom Lippincott and Annabelle Carrell (2018). "Observational Comparison of Geo-tagged and Randomly-drawn Tweets". *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, pages 50–55 (cited on page 84).

Zachary Lipton, Yu-Xiang Wang, and Alexander Smola (2018). "Detecting and correcting for label shift with black box predictors". *International conference on machine learning*. PMLR, pages 3122–3130 (cited on page 44).

Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan (2018). "Conditional adversarial domain adaptation". *Advances in neural information processing systems* 31 (cited on page 51).

David E Losada, Fabio Crestani, and Javier Parapar (2018). "Overview of eRisk: early risk prediction on the internet". *CLEF* (cited on page 70).

Ilya Loshchilov and Frank Hutter (2018). "Decoupled Weight Decay Regularization". *International Conference on Learning Representations* (cited on pages 237, 264).

Kate Loveys, Jonathan Torrez, Alex Fine, Glen Moriarty, and Glen Coppersmith (2018). "Cross-cultural differences in language markers of depression online". *Proceedings of the fifth workshop on computational linguistics and clinical psychology: from keyboard to clinic*, pages 78–87 (cited on pages 62, 68, 74, 148).

Sean MacAvaney, Bart Desmet, Arman Cohan, Luca Soldaini, Andrew Yates, Ayah Zirikly, and Nazli Goharian (2018). "RSDD-Time: Temporal Annotation of Self-Reported Mental Health Diagnoses". *Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic* (cited on pages 62, 122, 149).

BIBLIOGRAPHY

Alicia L Nobles, Caitlin N Dreisbach, Jessica Keim-Malpass, and Laura E Barnes (2018a). "" Is This an STD? Please Help!": Online Information Seeking for Sexually Transmitted Diseases on Reddit". *ICWSM* (cited on page 160).

Alicia L Nobles, Jeffrey J Glenn, Kamran Kowsari, Bethany A Teachman, and Laura E Barnes (2018b). "Identification of imminent suicide risk among young adults using text messages". *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–11 (cited on page 124).

Ahmed Husseini Orabi, Prasadith Buddhitha, Mahmoud Husseini Orabi, and Diana Inkpen (2018). "Deep learning for depression detection of twitter users". *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 88–97 (cited on page 82).

Inna Pirina and Çağrı Çöltekin (2018). "Identifying depression on reddit: The effect of training data". *SMM4H* (cited on pages 70, 71, 150).

Adam Poliak, Jason Naradoesky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme (2018). "Hypothesis Only Baselines in Natural Language Inference". *Seventh Joint Conference on Lexical and Computational Semantics* (cited on page 151).

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. (2018). "Improving language understanding by generative pre-training" (cited on page 194).

BIBLIOGRAPHY

Eva-Maria Rathner, Julia Djamali, Yannik Terhorst, Björn Schuller, Nicholas Cummins, Gudrun Salamon, Christina Hunger-Schoppe, and Harald Baumeister (2018). "How did you like 2017? detection of language markers of depression and narcissism in personal narratives". *Future* (cited on page 82).

Brenna N Renn, Abhishek Pratap, David C Atkins, Sean D Mooney, and Patricia A Areán (2018). "Smartphone-based passive assessment of mobility in depression: Challenges and opportunities". *Mental health and physical activity* 14, pages 136–139 (cited on page 62).

Farig Sadeque, Dongfang Xu, and Steven Bethard (2018). "Measuring the latency of depression detection in social media". *Eleventh ACM International Conference on Web Search and Data Mining* (cited on page 122).

Sebastian Schelter, Felix Biessmann, Tim Januschowski, David Salinas, Stephan Seufert, and Gyuri Szarvas (2018). "On challenges in machine learning model management" (cited on page 52).

Ivan Sekulic, Matej Gjurković, and Jan Šnajder (2018). "Not Just Depressed: Bipolar Disorder Prediction on Reddit". *9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis* (cited on page 70).

Tiancheng Shen, Jia Jia, Guangyao Shen, Fuli Feng, Xiangnan He, Huanbo Luan, Jie Tang, Thanassis Tiropanis, Tat Seng Chua, and Wendy Hall (2018). "Cross-domain

depression detection via harvesting social media". International Joint Conferences on Artificial Intelligence (cited on pages 85, 150).

Han-Chin Shing, Suraj Nair, Ayah Zirikly, Meir Friedenberg, Hal Daumé III, and Philip Resnik (2018). "Expert, crowdsourced, and machine assessment of suicide risk via online postings". *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 25–36 (cited on pages 70, 84).

Aaron Smith and Monica Anderson (2018). "Social media use in 2018". *Pew* (cited on page 68).

Hoyun Song, Jinseon You, and Jin-Woo Chung Jong C Park (2018). "Feature Attention Network: Interpretable Depression Detection from Social Media". *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation* (cited on page 82).

Marisa Torres-Ruiz, Kaitlynn Robinson-Ector, Dionna Attinson, Jamie Trotter, Ayodola Anise, and Steven Clauser (2018). "A portfolio analysis of culturally tailored trials to address health and healthcare disparities". *International journal of environmental research and public health* 15.9, page 1859 (cited on page 24).

Adam Tsakalidis, Maria Liakata, Theo Damoulas, and Alexandra I Cristea (2018). "Can we assess mental health through social media and smart devices? Addressing

bias in methodology and evaluation". *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, pages 407–423 (cited on pages 101, 127, 187).

Gaël Varoquaux (2018). "Cross-validation failure: small sample sizes lead to large error bars". *Neuroimage* 180 (cited on page 129).

Tiffany C Veinot, Hannah Mitchell, and Jessica S Ancker (2018). "Good intentions are not enough: how informatics interventions can worsen inequality". *Journal of the American Medical Informatics Association* 25.8, pages 1080–1088 (cited on page 27).

Genta Indra Winata, Onno Pepijn Kampman, and Pascale Fung (2018). "Attention-based lstm for psychological stress detection from spoken language using distant supervision". *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pages 6204–6208 (cited on page 81).

JT Wolohan, Misato Hiraga, Atreyee Mukherjee, Zeeshan Ali Sayyed, and Matthew Millard (2018a). "Detecting Linguistic Traces of Depression in Topic-Restricted Text: Attending to Self-Stigmatized Depression with NLP". *LCCM Workshop* (cited on pages 68, 70).

JT Wolohan, Misato Hiraga, Atreyee Mukherjee, Zeeshan Ali Sayyed, and Matthew Millard (2018b). "Detecting linguistic traces of depression in topic-restricted

text: attending to self-stigmatized depression with NLP". *Proceedings of the First International Workshop on Language Cognition and Computational Models*, pages 11–21 (cited on pages 82, 84, 87, 89, 91, 143, 150, 169, 170).

Han Zhao, Shanghang Zhang, Guanhang Wu, José MF Moura, Joao P Costeira, and Geoffrey J Gordon (2018a). "Adversarial multiple source domain adaptation". *Advances in neural information processing systems* 31 (cited on page 51).

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang (2018b). "Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods". *NAACL HLT* (cited on page 190).

Randall Akee, Maggie R Jones, and Sonya R Porter (2019). "Race matters: Income shares, income inequality, and income mobility for all US races". *Demography* 56.3, pages 999–1021 (cited on page 22).

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott (2019). "Publicly Available Clinical BERT Embeddings". *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78 (cited on pages 198, 202, 217, 218, 237, 264, 265).

Silvio Amir, Mark Dredze, and John W. Ayers (2019). "Mental Health Surveillance over Social Media with Digital Cohorts". *CLPsych* (cited on pages 58, 72, 84, 148, 153, 160).

BIBLIOGRAPHY

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz (2019). "Invariant risk minimization". *arXiv preprint arXiv:1907.02893* (cited on page 52).

Shikha Bordia and Samuel R Bowman (2019). "Identifying and reducing gender bias in word-level language models". *2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT 2019-Student Research Workshop, SRW 2019*. Association for Computational Linguistics (ACL), pages 7–15 (cited on page 28).

Luke Breitfeller, Emily Ahn, David Jurgens, and Yulia Tsvetkov (2019). "Finding microaggressions in the wild: A case for locating elusive phenomena in social media posts". *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 1664–1674 (cited on pages 247, 254).

Marc-Etienne Brunet, Colleen Alkalay-Houlihan, Ashton Anderson, and Richard Zemel (2019). "Understanding the origins of bias in word embeddings". *ICML* (cited on page 189).

Sandra Bucci, Matthias Schwannauer, and Natalie Berry (2019). "The digital revolution and its impact on mental health care". *Psychology and Psychotherapy: Theory, Research and Practice* (cited on page 57).

BIBLIOGRAPHY

Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma (2019). "Learning imbalanced datasets with label-distribution-aware margin loss". *Advances in neural information processing systems* 32 (cited on page 44).

Stevie Chancellor, Michael L Birnbaum, Eric D Caine, Vincent MB Silenzio, and Munmun De Choudhury (2019). "A taxonomy of ethical tensions in inferring mental health states from social media". *conference on fairness, accountability, and transparency* (cited on page 152).

Di Chen and Carla P Gomes (2019). "Bias Reduction via End-to-End Shift Learning: Application to Citizen Science". *Proceedings of the AAAI Conference on Artificial Intelligence*. Volume 33, pages 493–500 (cited on page 103).

Elena Davcheva, Martin Adam, and Alexander Benlian (2019). "User Dynamics in Mental Health Forums – A Sentiment Analysis Perspective". *Wirtschaftsinformatik* (cited on page 62).

Orianna Demasi, Marti A. Hearst, and Benjamin Recht (2019). "Towards augmenting crisis counselor training by improving message retrieval". *CLPsych* (cited on page 62).

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". *Proceedings of the 2019 Conference of the North American Chapter of the*

Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Edited by Jill Burstein, Christy Doran, and Thamar Solorio. Minneapolis, Minnesota: Association for Computational Linguistics, pages 4171–4186. DOI: 10.18653/v1/N19-1423. URL: https://aclanthology.org/N19-1423 (cited on pages 179, 193, 217, 218, 221, 237, 264).

Ashutosh Dhar Dwivedi, Gautam Srivastava, Shalini Dhar, and Rajani Singh (2019). "A decentralized privacy-preserving healthcare blockchain for IoT". *Sensors* (cited on page 72).

Hady Elsahar and Matthias Gallé (2019). "To annotate or not? predicting performance drop under domain shift". *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2163–2173 (cited on pages 33, 45).

Sindhu Kiranmai Ernala, Michael L Birnbaum, Kristin A Candan, Asra F Rizvi, William A Sterling, John M Kane, and Munmun De Choudhury (2019). "Methodological gaps in predicting mental health states from social media: Triangulating diagnostic signals". *CHI* (cited on pages 58, 72, 85, 145, 148, 170).

Elizabeth Ford, Keegan Curlewis, Akkapon Wongkoblap, and Vasa Curcin (2019). "Public opinions on using social media content to identify users with depression

and target mental health care advertising: mixed methods survey". *JMIR mental health* (cited on page 152).

Prasadith Kirinde Gamaarachchige and Diana Inkpen (2019). "Multi-Task, Multi-Channel, Multi-Input Learning for Mental Illness Detection using Social Media Text". *Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019)*, pages 54–64 (cited on page 86).

Hila Gonen and Yoav Goldberg (2019). "Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them". *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 609–614 (cited on page 28).

Ömer Gözüaçık, Alican Büyükçakır, Hamed Bonab, and Fazli Can (2019). "Unsupervised concept drift detection with a discriminative classifier". *Proceedings of the 28th ACM international conference on information and knowledge management*, pages 2365–2368 (cited on page 42).

Eben Holderness, Philip Cawkwell, Kirsten Bolton, James Pustejovsky, and Mei-Hua Hall (2019). "Distinguishing Clinical Sentiment: The Importance of Domain Adaptation in Psychiatric Patient Health Records". *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 117–123 (cited on page 62).

BIBLIOGRAPHY

Binxuan Huang and Kathleen M Carley (2019). “A hierarchical location prediction neural network for twitter user geolocation” (cited on page 73).

Xiaolei Huang and Michael Paul (2019). “Neural Temporality Adaptation for Document Classification: Diachronic Word Embeddings and Domain Adaptation Models”. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4113–4123 (cited on pages 115, 129, 161, 188).

Sarthak Jain and Byron C Wallace (2019). “Attention is not Explanation”. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3543–3556 (cited on pages 46, 156).

Sandra Just, Erik Haegert, Nora Kořánová, Anna-Lena Bröcker, Ivan Nenchev, Jakob Funcke, Christiane Montag, and Manfred Stede (2019). “Coherence models in schizophrenia”. *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 126–136 (cited on page 81).

Divyansh Kaushik, Eduard Hovy, and Zachary Lipton (2019). “Learning The Difference That Makes A Difference With Counterfactually-Augmented Data”. *International Conference on Learning Representations* (cited on pages 49, 285).

BIBLIOGRAPHY

Faiza Khan Khattak, Serena Jeblee, Chloé Pou-Prom, Mohamed Abdalla, Christopher Meaney, and Frank Rudzicz (2019). "A survey of word embeddings for clinical text". *Journal of Biomedical Informatics* 100, page 100057 (cited on page 202).

Akhil Alfons Kodiyan (2019). "An overview of ethical issues in using AI systems in hiring with a case study of Amazon's AI based hiring tool". *Researchgate Preprint*, pages 1–19 (cited on page 30).

Theresa A Koleck, Caitlin Dreisbach, Philip E Bourne, and Suzanne Bakken (2019). "Natural language processing of symptoms documented in free-text narratives of electronic health records: a systematic review". *Journal of the American Medical Informatics Association* 26.4, pages 364–379 (cited on page 197).

Wouter M Kouw and Marco Loog (2019). "A review of domain adaptation without target labels". *IEEE transactions on pattern analysis and machine intelligence* 43.3, pages 766–785 (cited on page 50).

Sawan Kumar, Sharmistha Jat, Karan Saxena, and Partha Talukdar (2019). "Zero-shot word sense disambiguation using sense definition embeddings". *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5670–5681 (cited on page 276).

Nelson F Liu, Matt Gardner, Yonatan Belinkov, Matthew E Peters, and Noah A Smith (2019). "Linguistic Knowledge and Transferability of Contextual Representations".

Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 1073–1094 (cited on page 217).

Yaoli Mao, Dakuo Wang, Michael Muller, Kush R Varshney, Ioana Baldini, Casey Dugan, and Aleksandra Mojsilović (2019). "How data scientistswork together with domain experts in scientific collaborations: To find the right answer or to ask the right question?" *Proceedings of the ACM on Human-Computer Interaction* 3.GROUP, pages 1–23 (cited on page 155).

Gerry McCartney, Franck Popham, Robert McMaster, and Andrew Cumbers (2019). "Defining health and health inequalities". *Public health* 172, pages 22–30 (cited on page 3).

Nimisha G Nair, Pallavi Satpathy, Jabez Christopher, et al. (2019). "Covariate shift: A review and analysis on classifiers". *2019 Global Conference for Advancement in Technology (GCAT)*. IEEE, pages 1–6 (cited on pages 42, 44).

John A Naslund and Kelly A Aschbrenner (2019). "Risks to privacy with use of social media: understanding the views of social media users with serious mental illness". *Psychiatric services* (cited on page 152).

Judy H Ng, Lauren M Ward, Madeleine Shea, Liz Hart, Paul Guerino, and Sarah Hudson Scholle (2019). "Explaining the relationship between minority group

status and health disparities: a review of selected concepts". *Health Equity* 3.1, pages 47–60 (cited on page 3).

Azadeh Nikfarjam, Julia D Ransohoff, Alison Callahan, Erik Jones, Brian Loew, Bernice Y Kwong, Kavita Y Sarin, Nigam H Shah, et al. (2019). "Early detection of adverse drug reactions in social health networks: a natural language processing pipeline for signal detection". *JMIR public health and surveillance* 5.2, e11264 (cited on page 4).

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. (2019). "Pytorch: An imperative style, high-performance deep learning library". *Advances in neural information processing systems* 32 (cited on page 264).

Ian Pearce, Rafael Simó, Monica Lövestam-Adrian, David T Wong, and Marc Evans (2019). "Association between diabetic eye disease and other complications of diabetes: implications for care. A systematic review". *Diabetes, obesity and metabolism* 21.3, pages 467–478 (cited on page 198).

Sachin R Pendse, Kate Niederhoffer, and Amit Sharma (2019). "Cross-cultural differences in the use of online mental health support forums". *Proceedings of the ACM on Human-Computer Interaction* 3.CSCW, pages 1–29 (cited on page 82).

BIBLIOGRAPHY

A Perrin and M Anderson (2019). *Share of US adults using social media, including Facebook, is mostly unchanged since 2018. Pew Research Center* (cited on page 65).

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller (2019). "Language Models as Knowledge Bases?" *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473 (cited on page 296).

Stephan Rabanser, Stephan Günnemann, and Zachary Lipton (2019). "Failing loudly: An empirical study of methods for detecting dataset shift". *Advances in Neural Information Processing Systems* 32 (cited on page 43).

Sarah Rush, Sara Britt, and John Marcotte (2019). "ICPSR Virtual Data Enclave as a Collaboratory for Team Science" (cited on page 72).

Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang (2019). "Distributionally Robust Neural Networks". *International Conference on Learning Representations* (cited on pages 28, 49, 52).

Koustuv Saha, Sang Chan Kim, Manikanta D Reddy, Albert J Carter, Eva Sharma, Oliver L Haimson, and Munmun De Choudhury (2019). "The language of LGBTQ+ minority stress experiences on social media". *Proceedings of the ACM on human-computer interaction* 3.CSCW, pages 1–22 (cited on page 82).

Jose Ramon Saura, Ana Reyes-Menendez, and Pedro Palos-Sanchez (2019). “Are black Friday deals worth it? Mining Twitter users’ sentiment and behavior response”. *JOItmC* (cited on page 160).

Lance Garrett Shaver, Ahmed Khawer, Yanqing Yi, Kris Aubrey-Bassler, Holly Etchegary, Barbara Roebothan, Shabnam Asghari, and Peizhong Peter Wang (2019). “Using Facebook advertising to recruit representative samples: feasibility assessment of a cross-sectional survey”. *JMIR* (cited on page 190).

Petar Stojanov, Ahmed Hassan Awadallah, Paul Bennett, and Saghar Hosseini (2019). “On Domain Transfer When Predicting Intent in Text” (cited on page 103).

Feng Sun, Hanrui Wu, Zhihang Luo, Wenwen Gu, Yuguang Yan, and Qing Du (2019a). “Informative feature selection for domain adaptation”. *IEEE Access* 7, pages 142551–142563 (cited on pages 51, 159).

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang (2019b). “Mitigating Gender Bias in Natural Language Processing: Literature Review”. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640 (cited on page 28).

Allison M Tackman, David A Sbarra, Angela L Carey, M Brent Donnellan, Andrea B Horn, Nicholas S Holtzman, To’Meisha S Edwards, James W Pennebaker, and

Matthias R Mehl (2019). “Depression, negative emotionality, and self-referential language: A multi-lab, multi-measure, and multi-language-task research synthesis.” *Journal of personality and social psychology* 116.5, page 817 (cited on page 146).

Yi Chern Tan and L Elisa Celis (2019). “Assessing social and intersectional biases in contextualized word representations”. *Advances in neural information processing systems* 32 (cited on page 40).

Petteri Teikari, Raymond P Najjar, Leopold Schmetterer, and Dan Milea (2019). “Embedded deep learning in ophthalmology: making ophthalmic imaging smarter”. *Therapeutic advances in ophthalmology* 11, page 2515841419827172 (cited on page 200).

Daniel Shu Wei Ting, Louis R Pasquale, Lily Peng, John Peter Campbell, Aaron Y Lee, Rajiv Raman, Gavin Siew Wei Tan, Leopold Schmetterer, Pearse A Keane, and Tien Yin Wong (2019). “Artificial intelligence and deep learning in ophthalmology”. *British Journal of Ophthalmology* 103.2, pages 167–175 (cited on page 200).

Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry (2019). “Robustness May Be at Odds with Accuracy”. *International Conference on Learning Representations* (cited on page 193).

BIBLIOGRAPHY

Elsbeth Turcan and Kathleen McKeown (2019a). “Dreaddit: A Reddit Dataset for Stress Analysis in Social Media”. *EMNLP-IJCNLP 2019*, page 97 (cited on page 106).

Elsbeth Turcan and Kathy McKeown (2019b). “Dreaddit: A Reddit Dataset for Stress Analysis in Social Media”. *LOUHI* (cited on page 70).

Zijian Wang and Christopher Potts (2019). “TalkDown: A Corpus for Condescension Detection in Context”. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3711–3719 (cited on page 254).

Sarah Wiegreffe and Yuval Pinter (2019). “Attention is not not Explanation”. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 11–20 (cited on page 156).

David R Williams, Jourdyn A Lawrence, and Brigette A Davis (2019). “Racism and health: evidence and needed research”. *Annual review of public health* 40, pages 105–125 (cited on page 248).

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al.

(2019). “Huggingface’s transformers: State-of-the-art natural language processing”. *arXiv preprint arXiv:1910.03771* (cited on page 264).

Akkapon Wongkoblap, Miguel A Vadillo, and Vasa Curcin (2019). “Predicting social network users with depression from simulated temporal data”. *IEEE EUROCON 2019-18th International Conference on Smart Technologies*. IEEE (cited on page 124).

Diyi Yang, Zheng Yao, Joseph Seering, and Robert Kraut (2019a). “The channel matters: Self-disclosure, reciprocity and social support in online cancer support groups”. *Proceedings of the 2019 chi conference on human factors in computing systems*, pages 1–15 (cited on page 75).

Qian-Hui Yang, Yan Zhang, Xiao-Min Zhang, and Xiao-Rong Li (2019b). “Prevalence of diabetic retinopathy, proliferative diabetic retinopathy and non-proliferative diabetic retinopathy in Asian T2DM patients: a systematic review and meta-analysis”. *International journal of ophthalmology* 12.2, page 302 (cited on page 198).

Chi Yuan, Patrick B Ryan, Casey Ta, Yixuan Guo, Ziran Li, Jill Hardin, Rupa Makadia, Peng Jin, Ning Shang, Tian Kang, et al. (2019). “Criteria2Query: a natural language interface to clinical databases for cohort definition”. *Journal of the American Medical Informatics Association* 26.4, pages 294–305 (cited on page 4).

BIBLIOGRAPHY

Frederick J Zimmerman and Nathaniel W Anderson (2019). "Trends in health equity in the United States by race/ethnicity, sex, and income, 1993-2017". *JAMA network open* 2.6, e196386–e196386 (cited on pages 4, 24).

Ayah Zirikly, Philip Resnik, Ozlem Uzuner, and Kristy Hollingshead (2019). "CLPsych 2019 shared task: Predicting the degree of suicide risk in Reddit posts". *sixth workshop on computational linguistics and clinical psychology* (cited on pages 70, 75).

Allison E Aiello, Audrey Renson, and Paul Zivich (2020). "Social media-and internet-based disease surveillance for public health". *Annual review of public health* 41, page 101 (cited on page 160).

Omar Ali, Nancy Scheidt, Alexander Gegov, Ella Haig, Mo Adda, and Benjamin Aziz (2020). "Automated detection of racial microaggressions using machine learning". *2020 IEEE symposium series on computational intelligence (SSCI)*. IEEE, pages 2477–2484 (cited on page 247).

Martin M Antony and David H Barlow (2020). *Handbook of assessment and treatment planning for psychological disorders*. Guilford Publications (cited on page 66).

John W Ayers, Eric C Leas, Derek C Johnson, Adam Poliak, Benjamin M Althouse, Mark Dredze, and Alicia L Nobles (2020). "Internet searches for acute anxiety

during the early stages of the COVID-19 pandemic". *JAMA Internal Medicine* (cited on pages 163, 164).

Michele Banko, Brendon MacKeen, and Laurie Ray (2020). "A unified taxonomy of harmful content". *Proceedings of the fourth workshop on online abuse and harms*, pages 125–137 (cited on page 247).

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn (2020). "The pushshift reddit dataset". *ICWSM* (cited on pages 169, 171).

Rohit Bhattacharya, Razieh Nabi, Ilya Shpitser, and James M Robins (2020). "Identification in missing data models represented by directed acyclic graphs". *Uncertainty in Artificial Intelligence*. PMLR, pages 1149–1158 (cited on page 35).

Laura Biester, Katie Matton, Janarthanan Rajendran, Emily Mower Provost, and Rada Mihalcea (2020). "Quantifying the Effects of COVID-19 on Mental Health Support Forums". *NLP for COVID-19* (cited on pages 164, 166).

Azzedine Boukerche, Lining Zheng, and Omar Alfandi (2020). "Outlier detection: Methods, models, and classification". *ACM Computing Surveys (CSUR)* 53.3, pages 1–37 (cited on page 41).

Peter Bühlmann (2020). "Invariance, causality and robustness". *Statistical Science* 35.3, pages 404–426 (cited on page 50).

BIBLIOGRAPHY

Kelsi Carolan, Ernest Gonzales, Kathy Lee, and Robert A Harootyan (2020). "Institutional and individual factors affecting health and employment for low-income women with chronic health conditions". *The Journals of Gerontology: Series B* 75.5, pages 1062–1071 (cited on page 76).

Jessica P Cerdeña, Marie V Plaisime, and Jennifer Tsai (2020). "From race-based to race-conscious medicine: how anti-racist uprisings call us to act". *The Lancet* 396.10257, pages 1125–1128 (cited on page 9).

Stevie Chancellor and Munmun De Choudhury (2020). "Methods in predictive techniques for mental health status on social media: a critical review". *NPJ digital medicine* (cited on pages 59, 60, 69, 71, 86, 95, 121, 123, 153, 170).

Irene Y Chen, Shalmali Joshi, and Marzyeh Ghassemi (2020). "Treating health disparities with artificial intelligence". *Nature medicine* 26.1, pages 16–17 (cited on page 27).

Yo Joong Choe, Jiyeon Ham, and Kyubyong Park (2020). "An empirical study of invariant risk minimization". *arXiv preprint arXiv:2004.05007* (cited on page 52).

Daejin Choi, Steven A Sumner, Kristin M Holland, John Draper, Sean Murphy, Daniel A Bowen, Marissa Zwald, Jing Wang, Royal Law, Jordan Taylor, et al. (2020). "Development of a machine learning model using multiple, heterogeneous

data sources to estimate weekly US suicide fatalities". *JAMA network open* (cited on page 127).

Mark É Czeisler, Rashon I Lane, Emiko Petrosky, Joshua F Wiley, Aleta Christensen, Rashid Njai, Matthew D Weaver, Rebecca Robbins, Elise R Facer-Childs, Laura K Barger, et al. (2020). "Mental health, substance use, and suicidal ideation during the COVID-19 pandemic—United States, June 24–30, 2020". *Morbidity and Mortality Weekly Report* 69.32, page 1049 (cited on page 164).

Ketki V Deshpande, Shimei Pan, and James R Foulds (2020). "Mitigating demographic Bias in AI-based resume filtering". *Adjunct publication of the 28th ACM conference on user modeling, adaptation and personalization*, pages 268–275 (cited on page 28).

Anjalie Field and Yulia Tsvetkov (2020). "Unsupervised Discovery of Implicit Gender Bias". *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 596–608 (cited on page 285).

Alex Fine, Patrick Crutchley, Jenny Blase, Joshua Carroll, and Glen Coppersmith (2020). "Assessing population-level symptoms of anxiety, depression, and suicide risk in real time using NLP applied to social media data". *Fourth Workshop on Natural Language Processing and Computational Social Science* (cited on pages 127, 191).

BIBLIOGRAPHY

Alexandre Flage (2020). “Discrimination against gays and lesbians in hiring decisions: a meta-analysis”. *International Journal of Manpower* 41.6, pages 671–691 (cited on page 22).

Christina J Flaxel, Ron A Adelman, Steven T Bailey, Amani Fawzi, Jennifer I Lim, G Atma Vemulakonda, and Gui-shuang Ying (2020). “Diabetic retinopathy preferred practice pattern®”. *Ophthalmology* 127.1, P66–P145 (cited on pages 197, 200).

Sandro Galea, Raina M Merchant, and Nicole Lurie (2020). “The mental health consequences of COVID-19 and physical distancing: the need for prevention and early intervention”. *JAMA internal medicine* (cited on pages 149, 162, 163).

Saurabh Garg, Yifan Wu, Sivaraman Balakrishnan, and Zachary Lipton (2020). “A unified view of label shift estimation”. *Advances in Neural Information Processing Systems* 33, pages 3290–3300 (cited on pages 35, 44).

Chuanxing Geng, Sheng-jun Huang, and Songcan Chen (2020). “Recent advances in open set recognition: A survey”. *IEEE transactions on pattern analysis and machine intelligence* 43.10, pages 3614–3631 (cited on page 41).

Abigail Goben and Robert J Sandusky (2020). “Open data repositories: current risks and opportunities”. *College & Research Libraries News* 81.2, page 62 (cited on page 154).

Danijela Godinic, Bojan Obrenovic, Akmal Khudaykulov, et al. (2020). “Effects of economic uncertainty on mental health in the COVID-19 pandemic context: social identity disturbance, job uncertainty and psychological well-being model”. *Int. J. Innov. Econ. Dev* (cited on page 163).

Hila Gonen, Ganesh Jawahar, Djamé Seddah, and Yoav Goldberg (2020). “Simple, interpretable and stable method for detecting words with usage change across corpora”. *ACL* (cited on pages 162, 167, 172, 175, 178, 186, 193).

Daniel Grießhaber, Johannes Maucher, and Ngoc Thang Vu (2020). “Fine-tuning BERT for Low-Resource Natural Language Understanding via Active Learning”. *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1158–1171 (cited on page 220).

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith (2020). “Don’t Stop Pretraining: Adapt Language Models to Domains and Tasks”. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360 (cited on pages 51, 198, 220).

Xiaochuang Han, Byron C Wallace, and Yulia Tsvetkov (2020). “Explaining Black Box Predictions and Unveiling Data Artifacts through Influence Functions”.

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 5553–5563 (cited on pages 133, 156).

Keith Harrigian, Carlos Aguirre, and Mark Dredze (2020). "Do Models of Mental Health Based on Social Media Data Generalize?" *"Findings of ACL: EMNLP"* (cited on pages 12, 80, 121, 123, 273, 301).

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei (2020). "Scaling laws for neural language models". *arXiv preprint arXiv:2001.08361* (cited on page 76).

Duleeka Knipe, Hannah Evans, Amanda Marchant, David Gunnell, and Ann John (2020). "Mapping population mental health concerns related to COVID-19 and the consequences of physical distancing: a Google trends analysis". *Wellcome Open Research* (cited on page 164).

Jing Xuan Koh and Tau Ming Liew (2020). "How loneliness is talked about in social media during COVID-19 pandemic: text mining of 4,492 Twitter feeds". *Journal of psychiatric research* (cited on page 163).

Alexandra Lautarescu, Michael C Craig, and Vivette Glover (2020). "Prenatal stress: Effects on fetal and child brain development". *International review of neurobiology* 150, pages 17–40 (cited on page 18).

BIBLIOGRAPHY

Danielle M Law, Jennifer D Shapka, and Rebecca J Collie (2020). "Who might flourish and who might languish? Adolescent social and mental health profiles and their online experiences and behaviors". *Human Behavior and Emerging Technologies* 2.1, pages 82–92 (cited on page 147).

Ronan Le Bras, Swabha Swayamdipta, Chandra Bhagavatula, Rowan Zellers, Matthew Peters, Ashish Sabharwal, and Yejin Choi (2020). "Adversarial filters of dataset biases". *International Conference on Machine Learning*. PMLR (cited on pages 49, 150, 297).

Amanda A Lee, Aimee S James, and Jean M Hunleth (2020a). "Waiting for care: Chronic illness and health system uncertainties in the United States". *Social Science & Medicine* 264, page 113296 (cited on page 3).

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang (2020b). "BioBERT: a pre-trained biomedical language representation model for biomedical text mining". *Bioinformatics* 36.4, pages 1234–1240 (cited on pages 198, 202, 228).

Patrick Lewis, Myle Ott, Jingfei Du, and Veselin Stoyanov (2020). "Pretrained language models for biomedical and clinical tasks: understanding and extending the state-of-the-art". *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 146–157 (cited on pages 202, 218).

BIBLIOGRAPHY

Chen Lin, Steven Bethard, Dmitriy Dligach, Farig Sadeque, Guergana Savova, and Timothy A Miller (2020). "Does BERT need domain adaptation for clinical negation detection?" *Journal of the American Medical Informatics Association* 27.4, pages 584–591 (cited on page 226).

Mufan Luo and Jeffrey T Hancock (2020). "Self-disclosure and social media: motivations, mechanisms and psychological well-being". *Current Opinion in Psychology* 31, pages 110–115 (cited on page 147).

Humphrey Mensah, Lu Xiao, and Sucheta Soundarajan (2020). "Characterizing the Evolution of Communities on Reddit". *International Conference on Social Media and Society*, pages 58–64 (cited on page 52).

Alicia L Nobles, Eric C Leas, Mark Dredze, and John W Ayers (2020). "Examining peer-to-peer and patient-provider interactions on a social media community facilitating ask the doctor services". *Proceedings of the International AAAI Conference on Web and Social Media*. Volume 14, pages 464–475 (cited on page 29).

Beau Norgeot, Giorgio Quer, Brett K Beaulieu-Jones, Ali Torkamani, Raquel Dias, Milena Gianfrancesco, Rima Arnaout, Isaac S Kohane, Suchi Saria, Eric Topol, et al. (2020). "Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist". *Nature medicine* (cited on page 71).

BIBLIOGRAPHY

Francisco Perales and Alice Campbell (2020). “Health disparities between sexual minority and different-sex-attracted adolescents: Quantifying the intervening role of social support and school belonging”. *LGBT health* 7.3, pages 146–154 (cited on page 40).

Cecilia Plaza (2020). “Miss Diagnosis: Gendered injustice in medical malpractice law”. *Colum. J. Gender & L.* 39, page 91 (cited on page 3).

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu (2020). “Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer”. *Journal of Machine Learning Research* 21, pages 1–67 (cited on page 194).

Alan Ramponi and Barbara Plank (2020). “Neural Unsupervised Domain Adaptation in NLP—A Survey”. *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6838–6855 (cited on pages 50, 159, 276).

Daniel Ranti, Katie Hanss, Shan Zhao, Varun Arvind, Joseph Titano, Anthony Costa, and Eric Oermann (2020). “The utility of general domain transfer learning for medical language tasks”. *arXiv preprint arXiv:2002.06670* (cited on page 226).

Guozheng Rao, Yue Zhang, Li Zhang, Qing Cong, and Zhiyong Feng (2020). “MGL-CNN: A hierarchical posts representations model for identifying depressed individuals in online forums”. *IEEE Access* (cited on page 124).

BIBLIOGRAPHY

Alexander Rietzler, Sebastian Stabinger, Paul Opitz, and Stefan Engl (2020). "Adapt or Get Left Behind: Domain Adaptation through BERT Language Model Finetuning for Aspect-Target Sentiment Classification". *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4933–4941 (cited on page 51).

Koustuv Saha, John Torous, Eric D Caine, Munmun De Choudhury, et al. (2020). "Psychosocial effects of the COVID-19 pandemic: large-scale quasi-experimental study on social media". *Journal of medical internet research* 22.11, e22600 (cited on page 170).

Bianca Scarlini, Tommaso Pasini, and Roberto Navigli (2020). "Sensembert: Context-enhanced sense embeddings for multilingual word sense disambiguation". *Proceedings of the AAAI conference on artificial intelligence*. Volume 34, pages 8758–8765 (cited on page 276).

Deven Santosh Shah, H Andrew Schwartz, and Dirk Hovy (2020). "Predictive Biases in Natural Language Processing Models: A Conceptual Framework and Overview". *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5248–5264 (cited on page 73).

Shahid Munir Shah and Rizwan Ahmed Khan (2020). "Secondary use of electronic health record: Opportunities and challenges". *IEEE access* 8, pages 136947–136965 (cited on page 76).

Nimit Sohoni, Jared Dunnmon, Geoffrey Angus, Albert Gu, and Christopher Ré (2020). "No subclass left behind: Fine-grained robustness in coarse-grained classification problems". *Advances in Neural Information Processing Systems* 33, pages 19339–19352 (cited on page 49).

Irena Spasic, Goran Nenadic, et al. (2020). "Clinical text data in machine learning: systematic review". *JMIR medical informatics* 8.3, e17984 (cited on pages 197, 228).

Stefan Stijelja and Brian L Mishara (2020). "COVID-19 and Psychological Distress—Changes in Internet Searches for Mental Health Issues in New York During the Pandemic". *JAMA internal medicine* (cited on pages 164, 166).

Adarsh Subbaswamy and Suchi Saria (2020). "From development to deployment: dataset shift, causality, and shift-stable models in health AI". *Biostatistics* 21.2, pages 345–352 (cited on page 83).

Tom Tabak and Matthew Purver (2020). "Temporal Mental Health Dynamics on Social Media". *NLP for COVID-19* (cited on page 191).

BIBLIOGRAPHY

Alon Talmor, Yanai Elazar, Yoav Goldberg, and Jonathan Berant (2020). "oLMpics-on what language model pre-training captures". *Transactions of the Association for Computational Linguistics* 8, pages 743–758 (cited on page 198).

Damien Teney, Ehsan Abbasnedjad, and Anton van den Hengel (2020). "Learning what makes a difference from counterfactual examples and gradient supervision". *European Conference on Computer Vision*. Springer, pages 580–599 (cited on page 49).

John Torous, Keris Jän Myrick, Natali Rauseo-Ricupero, and Joseph Firth (2020). "Digital mental health and COVID-19: using technology today to accelerate the curve on access and quality tomorrow". *JMIR mental health* (cited on page 163).

Lifu Tu, Garima Lalwani, Spandana Gella, and He He (2020). "An empirical study on robustness to spurious correlations using pre-trained language models". *Transactions of the Association for Computational Linguistics* 8, pages 621–633 (cited on page 48).

Zhao Wang and Aron Culotta (2020). "Identifying Spurious Correlations for Robust Text Classification". *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3431–3440 (cited on page 49).

Emily Brown Weida, Pam Phojanakong, Falguni Patel, and Mariana Chilton (2020). "Financial health as a measurable social determinant of health". *PloS one* 15.5, e0233359 (cited on page 22).

Naftali Weinberger and Seamus Bradley (2020). "Making sense of non-factual disagreement in science". *Studies in History and Philosophy of Science Part A* 83, pages 36–43 (cited on page 155).

Garrett Wilson and Diane J Cook (2020). "A survey of unsupervised deep domain adaptation". *ACM Transactions on Intelligent Systems and Technology (TIST)* 11.5, pages 1–46 (cited on page 50).

JT Wolohan (2020). "Estimating the effect of COVID-19 on mental health: Linguistic indicators of depression during a global pandemic". *1st Workshop on NLP for COVID-19 at ACL 2020* (cited on pages 153, 163, 166, 170).

Zach Wood-Doughty, Paiheng Xu, Xiao Liu, and Mark Dredze (2020). "Using Noisy Self-Reports to Predict Twitter User Demographics" (cited on page 73).

Sen Wu, Hongyang R Zhang, and Christopher Ré (2020). "Understanding and improving information transfer in multi-task learning". *arXiv preprint arXiv:2005.00944* (cited on page 240).

Hao Yao, Jian-Hua Chen, and Yi-Feng Xu (2020). "Patients with mental health disorders in the COVID-19 epidemic." (cited on page 163).

BIBLIOGRAPHY

Rong Zhang, Revanth Gangi Reddy, Md Arafat Sultan, Vittorio Castelli, Anthony Ferritto, Radu Florian, Efsun Sarioglu Kayi, Salim Roukos, Avirup Sil, and Todd Ward (2020). "Multi-Stage Pre-training for Low-Resource Domain Adaptation". *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5461–5468 (cited on page 299).

Anastazia Zunic, Padraig Corcoran, and Irena Spasic (2020). "Sentiment analysis in health and well-being: systematic review". *JMIR medical informatics* 8.1, e16023 (cited on page 48).

Carlos Aguirre, Keith Harrigian, and Mark Dredze (2021). "Gender and Racial Fairness in Depression Research using Social Media". *Proceedings of the 16th conference of the European Chapter of the Association for Computational Linguistics (EACL)* (cited on pages 13, 73, 121, 145, 148, 189).

FB Ahmad and JA Cisewski (2021). "Quarterly provisional estimates for selected indicators of mortality, 2018-Quarter 4, 2020". *National Center for Health Statistics. National Vital Statistics System, Vital Statistics Rapid Release Program* (cited on page 164).

Rukshan Alexander, Nik Thompson, Tanya McGill, and David Murray (2021). "The influence of user culture on website usability". *International journal of human-computer studies* 154, page 102688 (cited on page 54).

BIBLIOGRAPHY

John W Ayers, Adam Poliak, Derek C Johnson, Eric C Leas, Mark Dredze, Theodore Caputi, and Alicia L Nobles (2021). "Suicide-related internet searches during the early stages of the COVID-19 pandemic in the US". *JAMA network open* (cited on pages 163, 164).

Mary Catherine Beach and Somnath Saha (2021). "Quoting patients in clinical notes: First, do no harm". *Annals of internal medicine* 174.10, pages 1454–1455 (cited on pages 270, 285).

Mary Catherine Beach, Somnath Saha, Jenny Park, Janiece Taylor, Paul Drew, Eve Plank, Lisa A Cooper, and Brant Chee (2021). "Testimonial injustice: linguistic bias in the medical records of Black patients and women". *Journal of general internal medicine* 36.6, pages 1708–1714 (cited on pages 248, 251, 253, 256, 257, 270, 285).

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell (2021). "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" *FAccT* (cited on pages 189, 273).

Laura Biester, Katie Matton, Janarthanan Rajendran, Emily Mower Provost, and Rada Mihalcea (2021). "Understanding the impact of COVID-19 on online mental health forums". *ACM Transactions on Management Information Systems (TMIS)* (cited on page 153).

BIBLIOGRAPHY

Charlotte Blease, Jan Walker, Catherine M DesRoches, and Tom Delbanco (2021). *New US law mandates access to clinical notes: implications for patients and clinicians* (cited on page 249).

Erin Bondy, David AA Baranger, Jared Balbona, Kendall Sputo, Sarah E Paul, Thomas F Oltmanns, and Ryan Bogdan (2021). "Neuroticism and reward-related ventral striatum activity: Probing vulnerability to stress-related depression." *Journal of Abnormal Psychology* 130.3, page 223 (cited on page 146).

Michael Johnathan Charles Bray, Nicholas Omid Daneshvari, Indu Radhakrishnan, Janel Cubbage, Michael Eagle, Pamela Southall, and Paul Sasha Nestadt (2021). "Racial differences in statewide suicide mortality trends in Maryland during the coronavirus disease 2019 (COVID-19) pandemic". *JAMA psychiatry* (cited on page 162).

Cindy X Cai, Suzanne M Michalak, Sandra S Stinnett, Kelly W Muir, Sharon Fekrat, and Durga S Borkar (2021). "Effect of ICD-9 to ICD-10 transition on accuracy of codes for stage of diabetic retinopathy and related complications: results from the CODER study". *Ophthalmology Retina* 5.4, pages 374–380 (cited on pages 197, 200).

Steffen Castle, Robert Schwarzenberg, and Mohsen Pourvali (2021). "Detecting covariate drift with explanations". *CCF International Conference on Natural*

Language Processing and Chinese Computing. Springer, pages 317–322 (cited on page 44).

Qingyu Chen, Robert Leaman, Alexis Allot, Ling Luo, Chih-Hsuan Wei, Shankai Yan, and Zhiyong Lu (2021). "Artificial intelligence in action: addressing the COVID-19 pandemic with natural language processing". *Annual review of biomedical data science* 4, pages 313–339 (cited on page 4).

Lucio M Dery, Paul Michel, Ameet Talwalkar, and Graham Neubig (2021). "Should We Be Pre-training? An Argument for End-task Aware Training as an Alternative". *International Conference on Learning Representations* (cited on page 220).

Shizhe Diao, Ruijia Xu, Hongjin Su, Yilei Jiang, Yan Song, and Tong Zhang (2021). "Taming pre-trained language models with n-gram representations for low-resource domain adaptation". *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3336–3349 (cited on pages 51, 299).

Michelle Duvall, Frederick North, William Leasure, and Jennifer Pecina (2021). "Patient portal message characteristics and reported thoughts of self-harm and suicide: A retrospective cohort study". *Journal of telemedicine and telecare* 27.8, pages 501–508 (cited on page 37).

Hannah Eyre, Alec B Chapman, Kelly S Peterson, Jianlin Shi, Patrick R Alba, Makoto M Jones, Tamara L Box, Scott L DuVall, and Olga V Patterson (2021). "Launching into clinical space with medspaCy: a new clinical text processing toolkit in Python". *AMIA Annual Symposium Proceedings*. Volume 2021. American Medical Informatics Association, page 438 (cited on page 263).

Abolfazl Farahani, Sahar Voghoei, Khaled Rasheed, and Hamid R Arabnia (2021). "A brief review of domain adaptation". *Advances in data science and information engineering: proceedings from ICDATA 2020 and IKE 2020*, pages 877–894 (cited on page 50).

Yangye Fu, Ming Zhang, Xing Xu, Zuo Cao, Chao Ma, Yanli Ji, Kai Zuo, and Huimin Lu (2021). "Partial feature selection and alignment for multi-source domain adaptation". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16654–16663 (cited on pages 51, 159).

Michael J Gale, Brittni A Scruggs, and Christina J Flaxel (2021). "Diabetic eye disease: a review of screening and management recommendations". *Clinical & Experimental Ophthalmology* 49.2, pages 128–145 (cited on pages 198, 200).

Tianyu Gao, Adam Fisch, and Danqi Chen (2021). "Making Pre-trained Language Models Better Few-shot Learners". *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International*

Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 3816–3830 (cited on pages 198, 295).

Matt Gardner, William Merrill, Jesse Dodge, Matthew E Peters, Alexis Ross, Sameer Singh, and Noah A Smith (2021). "Competency Problems: On Finding and Removing Artifacts in Language Data". *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1801–1813 (cited on page 49).

Saurabh Garg, Sivaraman Balakrishnan, Zachary Chase Lipton, Behnam Neyshabur, and Hanie Sedghi (2021). "Leveraging Unlabeled Data to Predict Out-of-Distribution Performance". *NeurIPS 2021 Workshop on Distribution Shifts: Connecting Methods and Applications* (cited on page 33).

Yingqiang Ge, Shuchang Liu, Zelong Li, Shuyuan Xu, Shijie Geng, Yunqi Li, Juntao Tan, Fei Sun, and Yongfeng Zhang (2021). "Counterfactual evaluation for explainable AI". *arXiv preprint arXiv:2109.01962* (cited on pages 134, 151).

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford (2021). "Datasheets for datasets". *Communications of the ACM* (cited on page 151).

BIBLIOGRAPHY

Salvatore Giorgi, Veronica Lynn, Keshav Gupta, Farhan Ahmed, Sandra Matz, Lyle Ungar, and H Andrew Schwartz (2021). "Correcting Sociodemographic Selection Biases for Population Prediction from Social Media" (cited on page 190).

Carmen Gonzalez, Jody Early, Vanessa Gordon-Dseagu, Teresa Mata, and Carolina Nieto (2021). "Promoting culturally tailored mHealth: A scoping review of mobile health interventions in Latinx communities". *Journal of immigrant and minority health* 23.5, pages 1065–1077 (cited on page 24).

Koustava Goswami, Sourav Dutta, Haytham Assem, Theodorus Fransen, and John Philip McCrae (2021). "Cross-lingual sentence embedding using multi-task learning". *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9099–9113 (cited on page 298).

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon (2021). "Domain-specific language model pretraining for biomedical natural language processing". *ACM Transactions on Computing for Healthcare (HEALTH)* 3.1, pages 1–23 (cited on page 198).

Alice Guan, Marilyn Thomas, Eric Vittinghoff, Lisa Bowleg, Christina Mangurian, and Paul Wesson (2021). "An investigation of quantitative methods for assessing

intersectionality in health research: A systematic review". *SSM-population health* 16, page 100977 (cited on page 23).

Ruocheng Guo, Pengchuan Zhang, Hao Liu, and Emre Kiciman (2021). "Out-of-distribution prediction with invariant risk minimization: The limitation and an effective fix". *arXiv preprint arXiv:2101.07732* (cited on page 52).

Christian Haase, Saba Anwar, Seid Muhie Yimam, Alexander Friedrich, and Chris Biemann (2021). "SCoT: Sense Clustering over Time: a tool for the analysis of lexical change". *EACL* (cited on page 179).

Xiaochuang Han and Yulia Tsvetkov (2021). "Influence Tuning: Demoting Spurious Correlations via Instance Attribution and Instance-Driven Updates". *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4398–4409 (cited on page 49).

Keith Harrigian, Carlos Aguirre, and Mark Dredze (2021). "On the State of Social Media Data for Mental Health Research". *Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access* (cited on pages 12, 57, 121, 153, 301).

Dawn Heisey-Grove, Cheryl Rathert, Laura E McClelland, Kevin Jackson, and Jonathan P DeShazo (2021). "Patient and clinician characteristics associated with

secure message content: Retrospective cohort study". *Journal of medical Internet research* 23.8, e26650 (cited on page 37).

Patricia Homan, Tyson H Brown, and Brittany King (2021). "Structural intersectionality as a new direction for health disparities research". *Journal of health and social behavior* 62.3, pages 350–370 (cited on pages 23, 283).

Dirk Hovy and Diyi Yang (2021). "The importance of modeling social factors of language: Theory and practice". *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 588–602 (cited on pages 53, 277).

Rui Huang, Andrew Geng, and Yixuan Li (2021). "On the importance of gradients for detecting distributional shifts in the wild". *Advances in Neural Information Processing Systems* 34, pages 677–689 (cited on page 45).

Said A Ibrahim and Peter J Pronovost (2021). "Diagnostic errors, health disparities, and artificial intelligence: a combination for health or harm?" *JAMA Health Forum*. Volume 2. 9. American Medical Association, e212430–e212430 (cited on page 27).

Cris Martin P Jacoba, Leo Anthony Celi, and Paolo S Silva (2021). "Biomarkers for progression in diabetic retinopathy: expanding personalized medicine through integration of AI with electronic health records". *Seminars in ophthalmology*. Taylor & Francis, pages 250–257 (cited on page 200).

Kokil Jaidka, Sharath Chandra Guntuku, Jane H Lee, Zhengyi Luo, Anneke Buffone, and Lyle H Ungar (2021). "The rural–urban stress divide: Obtaining geographical insights through Twitter". *Computers in Human Behavior* 114, page 106544 (cited on page 82).

Shaoxiong Ji, Matti Hölttä, and Pekka Marttinen (2021). "Does the magic of BERT apply to medical code assignment? A quantitative study". *Computers in biology and medicine* 139, page 104998 (cited on page 226).

Fereshte Khani and Percy Liang (2021). "Removing spurious features can hurt accuracy and affect groups disproportionately". *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 196–205 (cited on page 53).

Benjamin Kompa, Jasper Snoek, and Andrew L Beam (2021). "Second opinion needed: communicating uncertainty in medical machine learning". *NPJ Digital Medicine* 4.1, page 4 (cited on page 41).

Andrew Lee, Jonathan K Kummerfeld, Larry An, and Rada Mihalcea (2021). "Micromodels for Efficient, Explainable, and Reusable Systems: A Case Study on Mental Health". *Findings of the Association for Computational Linguistics: EMNLP 2021* (cited on page 121).

Alyssa Lees, Daniel Borkan, Ian Kivlichan, Jorge Nario, and Tesh Goyal (2021). "Capturing Covertly Toxic Speech via Crowdsourcing". *Proceedings of the First*

Workshop on Bridging Human–Computer Interaction and Natural Language Processing, pages 14–20 (cited on page 254).

Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, et al. (2021). "Datasets: A Community Library for Natural Language Processing". *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 175–184 (cited on page 154).

Paul Pu Liang, Chiyu Wu, Louis-Philippe Morency, and Ruslan Salakhutdinov (2021). "Towards understanding and mitigating social biases in language models". *International Conference on Machine Learning*. PMLR, pages 6565–6576 (cited on page 249).

Evan Z Liu, Behzad Haghgoo, Annie S Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn (2021a). "Just train twice: Improving group robustness without training group information". *International Conference on Machine Learning*. PMLR, pages 6781–6792 (cited on pages 49, 297).

Jiashuo Liu, Zheyan Shen, Yue He, Xingxuan Zhang, Renzhe Xu, Han Yu, and Peng Cui (2021b). "Towards out-of-distribution generalization: A survey". *arXiv preprint arXiv:2108.13624* (cited on page 49).

BIBLIOGRAPHY

Sean Macavaney, Anjali Mittu, Glen Coppersmith, Jeff Leintz, and Philip Resnik (2021). "Community-level research on suicidality prediction in a secure environment: Overview of the CLPsych 2021 shared task". *Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access* (cited on page 121).

Marian F MacDorman, Marie Thoma, Eugene Declcerq, and Elizabeth A Howell (2021). "Racial and ethnic disparities in maternal mortality in the United States using enhanced vital records, 2016–2017". *American journal of public health* 111.9, pages 1673–1681 (cited on page 3).

Vivek Madan, Ashish Khetan, and Zohar Karnin (2021). "Tadpole: Task adapted pre-training via anomaly detection". *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5732–5746 (cited on page 51).

Lela R McKnight-Eily, Catherine A Okoro, Tara W Strine, Jorge Verlenden, NaTasha D Hollis, Rashid Njai, Elizabeth W Mitchell, Amy Board, Richard Puddy, and Craig Thomas (2021). "Racial and ethnic disparities in the prevalence of stress and worry, mental health conditions, and increased substance use among adults during the COVID-19 pandemic—United States, April and May 2020". *Morbidity and Mortality Weekly Report* 70.5, page 162 (cited on page 164).

Paul Michel, Sebastian Ruder, and Dani Yogatama (2021). "Balancing average and worst-case accuracy in multitask learning". *arXiv preprint arXiv:2110.05838* (cited on page 193).

Syed Ahnaf Morshed, Sifat Shahriar Khan, Raihanul Bari Tanvir, and Shafkath Nur (2021). "Impact of COVID-19 pandemic on ride-hailing services based on large-scale Twitter data analysis". *Journal of Urban Management* (cited on page 191).

Aakanksha Naik, Jill Fain Lehman, and Carolyn Rose (2021). "Adapting Event Extractors to Medical Data: Bridging the Covariate Shift". *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2963–2975 (cited on page 202).

Shahrzad Naseri, Jeffrey Dalton, Andrew Yates, and James Allan (2021). "Ceqe: Contextualized embeddings for query expansion". *Advances in Information Retrieval: 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28–April 1, 2021, Proceedings, Part I 43*. Springer, pages 467–482 (cited on page 257).

Bill Noble, Asad Sayeed, Raquel Fernández, and Staffan Larsson (2021). "Semantic shift in social networks". *Proceedings of* SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics*, pages 26–37 (cited on page 178).

Omolola I Ogunyemi, Meghal Gandhi, Martin Lee, Senait Teklehaimanot, Lauren Patty Daskivich, David Hindman, Kevin Lopez, and Ricky K Taira (2021). "Detecting diabetic retinopathy through machine learning on electronic health record data from an urban, safety net healthcare system". *JAMIA open* 4.3, ooab066 (cited on page 200).

Guansong Pang, Chunhua Shen, Longbing Cao, and Anton Van Den Hengel (2021). "Deep learning for anomaly detection: A review". *ACM computing surveys (CSUR)* 54.2, pages 1–38 (cited on page 41).

Jenny Park, Somnath Saha, Brant Chee, Janiece Taylor, and Mary Catherine Beach (2021a). "Physician use of stigmatizing language in patient medical records". *JAMA Network Open* 4.7, e2117052–e2117052 (cited on page 249).

Soya Park, April Yi Wang, Ban Kawas, Q Vera Liao, David Piorkowski, and Marina Danilevsky (2021b). "Facilitating knowledge sharing from domain experts to data scientists for building nlp models". *26th International Conference on Intelligent User Interfaces*, pages 585–596 (cited on page 155).

Braja G Patra, Mohit M Sharma, Veer Vekaria, Prakash Adekkanattu, Olga V Patterson, Benjamin Glicksberg, Lauren A Lepow, Euijung Ryu, Joanna M Biernacka, Al'ona Furmanchuk, et al. (2021). "Extracting social determinants of health from electronic health records using natural language processing: a

systematic review". *Journal of the American Medical Informatics Association* 28.12, pages 2716–2727 (cited on page 29).

Julia Raney, Ria Pal, Tiffany Lee, Samuel Ricardo Saenz, Devika Bhushan, Peter Leahy, Carrie Johnson, Cynthia Kapphahn, Michael A Gisondi, and Kim Hoang (2021). "Words matter: an antibias workshop for health care professionals to reduce stigmatizing language". *MedEdPORTAL* 17, page 11115 (cited on page 252).

Hussain Ahmed Raza, Parikshit Sen, Omaima Anis Bhatti, and Latika Gupta (2021). "Sex hormones, autoimmunity and gender disparity in COVID-19". *Rheumatology international* 41.8, pages 1375–1386 (cited on page 39).

Eliane Röösli, Brian Rice, and Tina Hernandez-Boussard (2021). "Bias at warp speed: how AI may contribute to the disparities gap in the time of COVID-19". *Journal of the American Medical Informatics Association* 28.1, pages 190–192 (cited on page 27).

Nazanin Sabri, Valerio Basile, Tommaso Caselli, et al. (2021). "Leveraging Bias in Pre-Trained Word Embeddings for Unsupervised Microaggression Detection". *CLiC-it* (cited on page 254).

Koustuv Saha, John Torous, Emre Kiciman, and Munmun De Choudhury (2021). "Understanding Side Effects of Antidepressants: Large-scale Longitudinal Study on Social Media Data". *JMIR mental health* (cited on pages 124, 128, 148).

BIBLIOGRAPHY

Beatriz Garcia Santa Cruz, Matías Nicolás Bossa, Jan Sölter, and Andreas Dominik Husch (2021). "Public covid-19 x-ray datasets and their impact on model bias–a systematic review of a significant problem". *Medical image analysis* 74, page 102225 (cited on page 31).

Mihaela van der Schaar, Ahmed M Alaa, Andres Floto, Alexander Gimson, Stefan Scholtes, Angela Wood, Eoin McKinney, Daniel Jarrett, Pietro Lio, and Ari Ercole (2021). "How artificial intelligence and machine learning can help healthcare systems respond to COVID-19". *Machine Learning* 110.1, pages 1–14 (cited on page 160).

Chandan K Sen (2021). "Human wound and its burden: updated 2020 compendium of estimates". *Advances in wound care* 10.5, pages 281–292 (cited on page 21).

Eli Sherman, Keith Harrigian, Carlos Aguirre, and Mark Dredze (2021). "Towards Understanding the Role of Gender in Deploying Social Media-Based Mental Health Surveillance Models". *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*, pages 217–223 (cited on page 13).

Giorgio Sirugo, Sarah A Tishkoff, Scott M Williams, et al. (2021). "The quagmire of race, genetic ancestry, and health disparities". *The Journal of clinical investigation* 131.11 (cited on page 39).

Annika Skoogh, Marie Louise Hall-Lord, Carina Bååth, and Ann-Kristin Sandin Bojö (2021). "Adverse events in women giving birth in a labor ward: a retrospective record review study". *BMC Health Services Research* 21, pages 1–8 (cited on page 16).

Jaimie D Steinmetz, Rupert RA Bourne, Paul Svitil Briant, Seth R Flaxman, Hugh RB Taylor, Jost B Jonas, Amir Aberhe Abdoli, Woldu Aberhe Abrha, Ahmed Abualhasan, Eman Girum Abu-Gharbieh, et al. (2021). "Causes of blindness and vision impairment in 2020 and trends over 30 years, and prevalence of avoidable blindness in relation to VISION 2020: the Right to Sight: an analysis for the Global Burden of Disease Study". *The Lancet Global Health* 9.2, e144–e160 (cited on pages 197, 200).

Madhumita Sushil, Simon Suster, and Walter Daelemans (2021). "Are we there yet? Exploring clinical domain knowledge of BERT models". *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 41–53 (cited on page 48).

Nicole M Thomasian, Carsten Eickhoff, and Eli Y Adashi (2021). "Advancing health equity with artificial intelligence". *Journal of public health policy* 42.4, pages 602–611 (cited on page 6).

Junjiao Tian, Yen-Chang Hsu, Yilin Shen, Hongxia Jin, and Zsolt Kira (2021). "Exploring covariate and concept shift for out-of-distribution detection". *NeurIPS*

2021 Workshop on Distribution Shifts: Connecting Methods and Applications (cited on page 43).

Ana Sabina Uban, Berta Chulvi, and Paolo Rosso (2021). "Understanding Patterns of Anorexia Manifestations in Social Media Data with Deep Learning". *Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access* (cited on pages 124, 151).

Anna Valdez (2021). "Words matter: Labelling, bias and stigma in nursing". *Journal of Advanced Nursing* 77.11, e33–e35 (cited on page 251).

Matej Gjurkovic Mladen Karan Iva Vukojevic and Mihaela Bošnjak Jan Šnajder (2021). "PANDORA Talks: Personality and Demographics on Reddit". *SocialNLP 2021* (cited on page 123).

Gloria J Washington, Gishawn Mance, Saurav K Aryal, and Mikel Kengni (2021). "ABL-MICRO: Opportunities for Affective AI Built Using a Multimodal Microaggression Dataset." *AffCon@ AAAI*, pages 23–29 (cited on page 254).

Colin Wei, Sang Michael Xie, and Tengyu Ma (2021). "Why do pretrained language models help in downstream tasks? an analysis of head and prompt tuning". *Advances in Neural Information Processing Systems* 34, pages 16158–16170 (cited on page 198).

Katherine J Wolsiefer, Matthias Mehl, Gordon B Moskowitz, Colleen K Cagno, Colin A Zestcott, Alma Tejeda-Padron, and Jeff Stone (2021). "Investigating the relationship between resident physician implicit bias and language use during a clinical encounter with hispanic patients". *Health Communication*, pages 1–9 (cited on page 249).

Charles C Wykoff, Rahul N Khurana, Quan Dong Nguyen, Scott P Kelly, Flora Lum, Rebecca Hall, Ibrahim M Abbass, Anna M Abolian, Ivaylo Stoilov, Tu My To, et al. (2021). "Risk of blindness among patients with diabetes and newly diagnosed diabetic retinopathy". *Diabetes care* 44.3, pages 748–756 (cited on page 197).

Yadollah Yaghoobzadeh, Soroush Mehri, Remi Tachet des Combes, Timothy J Hazen, and Alessandro Sordoni (2021). "Increasing Robustness to Spurious Correlations using Forgettable Examples". *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3319–3332 (cited on page 49).

Yuguang Yan, Hanrui Wu, Yuzhong Ye, Chaoyang Bi, Min Lu, Dapeng Liu, Qingyao Wu, and Michael K Ng (2021). "Transferable feature selection for unsupervised domain adaptation". *IEEE Transactions on Knowledge and Data Engineering* 34.11, pages 5536–5551 (cited on page 51).

Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu (2021a). "Generalized out-of-distribution detection: A survey". *arXiv preprint arXiv:2110.11334* (cited on page 41).

Lily Wei Yun Yang, Wei Yan Ng, Li Lian Foo, Yong Liu, Ming Yan, Xiaofeng Lei, Xiaoman Zhang, and Daniel Shu Wei Ting (2021b). "Deep learning-based natural language processing in ophthalmology: applications, challenges and future directions". *Current opinion in ophthalmology* 32.5, pages 397–405 (cited on page 201).

Wenjie Yin and Arkaitz Zubiaga (2021). "Towards generalisable hate speech detection: a review on obstacles and solutions". *PeerJ Computer Science* 7, e598 (cited on page 254).

Valentina A Zavala, Paige M Bracci, John M Carethers, Luis Carvajal-Carmona, Nicole B Coggins, Marcia R Cruz-Correa, Melissa Davis, Adam J de Smith, Julie Dutil, Jane C Figueiredo, et al. (2021). "Cancer health disparities in racial/ethnic minorities in the United States". *British journal of cancer* 124.2, pages 315–332 (cited on page 248).

Jon Zelner, Julien Riou, Ruth Etzioni, and Andrew Gelman (2021). "Accounting for uncertainty during a pandemic". *Patterns* (cited on page 163).

Yonggang Zhang, Mingming Gong, Tongliang Liu, Gang Niu, Xinmei Tian, Bo Han, Bernhard Schölkopf, and Kun Zhang (2021). "Adversarial Robustness Through the Lens of Causality". *International Conference on Learning Representations* (cited on page 50).

Zhi Zheng, Kai Hui, Ben He, Xianpei Han, Le Sun, and Andrew Yates (2021). "Contextualized query expansion via unsupervised chunk selection for text retrieval". *Information Processing & Management* 58.5, page 102672 (cited on page 257).

Hammaad Adam, Ming Ying Yang, Kenrick Cato, Ioana Baldini Soares, Charles Senteio, Jiaming Zeng, Moninder Singh, and Marzyeh Ghassemi (2022). "Write It Like You See It: Detectable Differences in Clinical Notes By Race Lead To Differential Model Recommendations". *AAAI/ACM Conference on AI, Ethics, and Society* (cited on page 268).

Oshin Agarwal and Ani Nenkova (2022). "Temporal Effects on Pre-trained Models for Language Processing Tasks". *Transactions of the Association for Computational Linguistics* 10, pages 904–921 (cited on page 122).

Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag (2022). "Large language models are few-shot clinical information extractors".

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 1998–2022 (cited on pages 203, 228, 230).

Simran Arora, Sen Wu, Enci Liu, and Christopher Ré (2022). "Metadata shaping: A simple approach for knowledge-enhanced language models". *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1733–1745 (cited on page 298).

Leo Anthony Celi, Jacqueline Cellini, Marie-Laure Charpignon, Edward Christopher Dee, Franck Dernoncourt, Rene Eber, William Greig Mitchell, Lama Moukheiber, Julian Schirmer, Julia Situ, et al. (2022). "Sources of bias in artificial intelligence that perpetuate healthcare disparities—A global review". *PLOS Digital Health* 1.3, e0000022 (cited on page 27).

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. (2022). "Scaling instruction-finetuned language models". *arXiv preprint arXiv:2210.11416* (cited on page 230).

Bhuwan Dhingra, Jeremy R Cole, Julian Martin Eisenschlos, Dan Gillick, Jacob Eisenstein, and William Cohen (2022). "Time-Aware Language Models as Temporal Knowledge Bases". *Transactions of the Association for Computational Linguistics* 10, pages 257–273 (cited on pages 166, 179).

BIBLIOGRAPHY

Laura Dwyer-Lindgren, Parkes Kendrick, Yekaterina O Kelly, Dillon O Sylte, Chris Schmidt, Brigette F Blacker, Farah Daoud, Amal A Abdi, Mathew Baumann, Farah Mouhanna, et al. (2022). "Life expectancy by county, race, and ethnicity in the USA, 2000–19: a systematic analysis of health disparities". *The Lancet* 400.10345, pages 25–38 (cited on page 3).

Tanmay Garg, Sarah Masud, Tharun Suresh, and Tanmoy Chakraborty (2022). "Handling Bias in Toxic Speech Detection: A Survey". *arXiv preprint arXiv:2202.00126* (cited on page 249).

Roger Garriga, Javier Mas, Semhar Abraha, Jon Nolan, Oliver Harrison, George Tadros, and Aleksandar Matic (2022). "Machine learning model to predict mental health crises from electronic health records". *Nature medicine* 28.6, pages 1240–1248 (cited on page 57).

Tejas Gokhale, Joshua Feinglass, and Yezhou Yang (2022). "Covariate Shift Detection via Domain Interpolation Sensitivity". *First Workshop on Interpolation Regularizers and Beyond at NeurIPS 2022* (cited on page 43).

Bernal Jiménez Gutiérrez, Nikolas McNeal, Clayton Washington, You Chen, Lang Li, Huan Sun, and Yu Su (2022). "Thinking about GPT-3 In-Context Learning for Biomedical IE? Think Again". *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4497–4512 (cited on page 228).

BIBLIOGRAPHY

Olena Hankivsky (2022). “INTERSECTIONALITY 101” (cited on page 40).

Keith Harrigian and Mark Dredze (2022a). “The Problem of Semantic Shift in Longitudinal Monitoring of Social Media”. *Proceeedings of the 14th ACM Web Science Conference* (cited on pages 13, 153, 159, 301).

Keith Harrigian and Mark Dredze (2022b). “Then and Now: Quantifying the Longitudinal Validity of Self-Disclosed Depression Diagnoses”. *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology*, pages 59–75 (cited on pages 12, 120, 301).

Jennifer Huang Harris, Nomi C Levy-Carrick, and Ashwini Nadkarni (2022). “OpenNotes: transparency versus stigma in patient care”. *The Lancet Psychiatry* 9.6, pages 426–428 (cited on page 249).

Megan Healy, Alison Richard, and Khameer Kidia (2022). “How to Reduce Stigma and Bias in Clinical Communication: a Narrative Review”. *Journal of General Internal Medicine*, pages 1–8 (cited on page 273).

Gracie Himmelstein, David Bates, and Li Zhou (2022). “Examination of stigmatizing language in the electronic health record”. *JAMA network open* 5.1, e2144967–e2144967 (cited on pages 253, 277, 285).

BIBLIOGRAPHY

Hadi Hojjati, Thi Kieu Khanh Ho, and Narges Armanfard (2022). "Self-supervised anomaly detection: A survey and outlook". *arXiv preprint arXiv:2205.05173* (cited on page 42).

Ismail Jatoi, Hyuna Sung, and Ahmedin Jemal (2022). "The emergence of the racial disparity in US breast-cancer mortality". *New England Journal of Medicine* 386.25, pages 2349–2352 (cited on pages 4, 24).

Nitish Joshi and He He (2022). "An Investigation of the (In) effectiveness of Counterfactually Augmented Data". *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3668–3681 (cited on page 49).

Shona Kalkman, Johannes van Delden, Amitava Banerjee, Benoit Tyl, Menno Mostert, and Ghislaine van Thiel (2022). "Patients' and public views and attitudes towards the sharing of health data for research: a narrative review of the empirical evidence". *Journal of medical ethics* 48.1, pages 3–13 (cited on page 76).

Anurag Katakkar, Clay H Yoo, Weiqin Wang, Zachary C Lipton, and Divyansh Kaushik (2022). "Practical Benefits of Feature Feedback Under Distribution Shift". *Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 346–355 (cited on page 159).

BIBLIOGRAPHY

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa (2022). "Large language models are zero-shot reasoners". *Advances in neural information processing systems* 35, pages 22199–22213 (cited on page 295).

Anastasios Lamproudis, Aron Henriksson, and Hercules Dalianis (2022). "Evaluating pretraining strategies for clinical bert models". *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 410–416 (cited on pages 202, 218).

Tingting Liu, Salvatore Giorgi, Xiangyu Tao, Douglas Bellew, Brenda Curtis, and Lyle Ungar (2022). "Cross-Platform Difference in Facebook and Text Messages Language Use: Illustrated by Depression Diagnosis". *ArXiv Preprint ArXiv* 220201802 (cited on page 103).

Daniel Loureiro, Francesco Barbieri, Leonardo Neves, Luis Espinosa Anke, and Jose Camacho-Collados (2022). "TimeLMs: Diachronic Language Models from Twitter". *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 251–260 (cited on pages 161, 179).

Qiuhao Lu, Dejing Dou, and Thien Nguyen (2022). "ClinicalT5: A generative language model for clinical text". *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5436–5443 (cited on page 228).

BIBLIOGRAPHY

Andreas Madsen, Siva Reddy, and Sarath Chandar (2022). “Post-hoc interpretability for neural nlp: A survey”. *ACM Computing Surveys* 55.8, pages 1–42 (cited on page 159).

Siddharth Nath, Abdullah Marie, Simon Ellershaw, Edward Korot, and Pearse A Keane (2022). “New meaning for NLP: the trials and tribulations of natural language processing with GPT-3 in ophthalmology”. *British Journal of Ophthalmology* 106.7, pages 889–892 (cited on page 201).

Giovanna Nicora, Miguel Rios, Ameen Abu-Hanna, and Riccardo Bellazzi (2022). “Evaluating pointwise reliability of machine learning prediction”. *Journal of Biomedical Informatics* 127, page 103996 (cited on page 42).

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. (2022). “Training language models to follow instructions with human feedback”. *Advances in neural information processing systems* 35, pages 27730–27744 (cited on page 295).

Aniket Pramanick, Tilman Beck, Kevin Stowe, and Iryna Gurevych (2022). “The challenges of temporal alignment on Twitter during crises”. *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2658–2672 (cited on page 103).

Julian Reiss (2022). "Why do experts disagree?" *Technocracy and the Epistemology of Human Behavior*. Routledge, pages 218–241 (cited on page 155).

Alexander Robey, Luiz Chamon, George J Pappas, and Hamed Hassani (2022). "Probabilistically robust learning: Balancing average and worst-case performance". *International Conference on Machine Learning*. PMLR, pages 18667–18686 (cited on page 193).

Andrea Schioppa, Polina Zablotskaia, David Vilar, and Artem Sokolov (2022). "Scaling up influence functions". *Proceedings of the AAAI Conference on Artificial Intelligence*. Volume 36. 8, pages 8179–8186 (cited on page 156).

Jessica Schrouff, Natalie Harris, Sanmi Koyejo, Ibrahim M Alabdulmohsin, Eva Schnider, Krista Opsahl-Ong, Alexander Brown, Subhrajit Roy, Diana Mincu, Christina Chen, et al. (2022). "Diagnosing failures of fairness transfer across distribution shift in real-world medical settings". *Advances in Neural Information Processing Systems* 35, pages 19304–19318 (cited on page 29).

Paras Sheth, Raha Moraffah, K Selçuk Candan, Adrienne Raglin, and Huan Liu (2022). "Domain Generalization–A Causal Perspective". *arXiv preprint arXiv:2209.15177* (cited on page 50).

Michael Sun, Tomasz Oliwa, Monica E Peek, and Elizabeth L Tung (2022). "Negative Patient Descriptors: Documenting Racial Bias In The Electronic Health Record:

Study examines racial bias in the patient descriptors used in the electronic health record." *Health Affairs* 41.2, pages 203–211 (cited on pages 249, 253, 254, 256, 257, 262, 263, 277, 285).

Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic (2022). "Galactica: A large language model for science". *arXiv preprint arXiv:2211.09085* (cited on page 202).

A Venigalla, J Frankle, and M Carbin (2022). "Biomedlm: a domain-specific large language model for biomedical text". *MosaicML. Accessed: Dec* 23.3, page 2 (cited on page 202).

Bohua Wan, Brian Caffo, and S Swaroop Vedula (2022). "A unified framework on generalizability of clinical prediction models". *Frontiers in Artificial Intelligence* 5, page 872720 (cited on pages 34, 35).

Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen, Wenjun Zeng, and S Yu Philip (2022a). "Generalizing to unseen domains: A survey on domain generalization". *IEEE transactions on knowledge and data engineering* 35.8, pages 8052–8072 (cited on page 49).

Tianlu Wang, Rohit Sridhar, Diyi Yang, and Xuezhi Wang (2022b). "Identifying and Mitigating Spurious Correlations for Improving Robustness in NLP Models".

Findings of the Association for Computational Linguistics: NAACL 2022, pages 1719–1729 (cited on pages 48, 49, 276).

Xuezhi Wang, Haohan Wang, and Diyi Yang (2022c). "Measure and Improve Robustness in NLP Models: A Survey". *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4569–4586 (cited on page 49).

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. (2022). "Chain-of-thought prompting elicits reasoning in large language models". *Advances in neural information processing systems* 35, pages 24824–24837 (cited on page 297).

Karen Werder, Alexa Curtis, Stephanie Reynolds, and Jason Satterfield (2022). "Addressing bias and stigma in the language we use with persons with opioid use disorder: A narrative review". *Journal of the American Psychiatric Nurses Association* 28.1, pages 9–22 (cited on page 249).

Xi Yang, Nima PourNejatian, Hoo Chang Shin, Kaleb E Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Mona G Flores, Ying Zhang, Tanja Magoc, et al. (2022). "GatorTron: A Large Clinical Language Model to Unlock Patient Information from Unstructured Electronic Health Records". *arXiv preprint arXiv:2203.03540* (cited on pages 198, 218, 228).

Zehao Yu, Xi Yang, Gianna L Sweeting, Yinghan Ma, Skylar E Stolte, Ruogu Fang, and Yonghui Wu (2022). "Identify diabetic retinopathy-related clinical concepts and their attributes using transformer-based natural language processing methods". *BMC Medical Informatics and Decision Making* 22.3, pages 1–9 (cited on pages 201, 204).

Haoran Zhang, Natalie Dullerud, Karsten Roth, Lauren Oakden-Rayner, Stephen Pfohl, and Marzyeh Ghassemi (2022). "Improving the fairness of chest x-ray classifiers". *Conference on health, inference, and learning*. PMLR, pages 204–233 (cited on page 28).

Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy (2022). "Domain generalization: A survey". *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45.4, pages 4396–4415 (cited on page 49).

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. (2023). "Gpt-4 technical report". *arXiv preprint arXiv:2303.08774* (cited on pages 5, 295).

Jide Alaga and Jonas Schuett (2023). "Coordinated pausing: An evaluation-based coordination scheme for frontier AI developers". *arXiv preprint arXiv:2310.00374* (cited on page 230).

BIBLIOGRAPHY

Yele Aluko, Susan Garfield, Perri Kasen, and Belinda Minta (2023). *Why America's health equity investment has yielded a marginal return*. https://www.ey.com/en_us/insights/health/america-s-health-equity-investment-marginal-return. [Online; accessed 09-May-2024] (cited on pages 4, 24).

Fares Antaki, Samir Touma, Daniel Milad, Jonathan El-Khoury, and Renaud Duval (2023). "Evaluating the performance of chatgpt in ophthalmology: An analysis of its successes and shortcomings". *medRxiv*, pages 2023–01 (cited on page 201).

Daniel Arias-Garzón, Reinel Tabares-Soto, Joshua Bernal-Salcedo, and Gonzalo A Ruz (2023). "Biases associated with database structure for COVID-19 detection in X-ray images". *Scientific reports* 13.1, page 3477 (cited on page 31).

Misha Armstrong, Natalie C Benda, Kenneth Seier, Christopher Rogers, Jessica S Ancker, Peter D Stetson, Yifan Peng, and Lisa C Diamond (2023). "Improving cancer care communication: identifying sociodemographic differences in patient portal secure messages not authored by the patient". *Applied Clinical Informatics* 14.02, pages 296–299 (cited on page 37).

Sedat Arslan (2023). "Exploring the potential of chat GPT in personalized obesity treatment". *Annals of biomedical engineering* 51.9, pages 1887–1888 (cited on page 32).

BIBLIOGRAPHY

Cindy X Cai, Diep Tran, Tina Tang, Wilson Liou, Keith Harrigian, Emily Scott, Paul Nagy, Hadi Kharrazi, Deidra C Crews, and Scott L Zeger (2023). "Health Disparities in Lapses in Diabetic Retinopathy Care". *Ophthalmology Science* 3.3, page 100295 (cited on pages 13, 294).

Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian (2023a). "Extending context window of large language models via positional interpolation". *arXiv preprint arXiv:2306.15595* (cited on page 296).

Yi Chen, Rui Wang, Haiyun Jiang, Shuming Shi, and Ruifeng Xu (2023b). "Exploring the use of large language models for reference-free text quality evaluation: A preliminary empirical study". *arXiv preprint arXiv:2304.00723* (cited on page 296).

Zhisheng Chen (2023). "Ethics and discrimination in artificial intelligence-enabled recruitment practices". *Humanities and Social Sciences Communications* 10.1, pages 1–12 (cited on page 30).

Brittany I Davidson, Darja Wischerath, Daniel Racek, Douglas A Parry, Emily Godwin, Joanne Hinds, Dirk van der Linden, Jonathan F Roscoe, Laura Ayravainen, and Alicia G Cork (2023). "Platform-controlled social media APIs threaten open science". *Nature Human Behaviour* 7.12, pages 2054–2057 (cited on page 81).

John Duchi, Tatsunori Hashimoto, and Hongseok Namkoong (2023). “Distributionally robust losses for latent covariate mixtures”. *Operations Research* 71.2, pages 649–664 (cited on page 49).

Hanna Dumont and Douglas D Ready (2023). “On the promise of personalized learning for educational equity”. *Npj science of learning* 8.1, page 26 (cited on page 32).

Future of Life Institute (2023). *Open Letter: Pause Giant AI Experiments*. https://futureoflife.org/open-letter/pause-giant-ai-experiments. Accessed: [insert access date here] (cited on page 230).

Bernal Jiménez Gutiérrez, Huan Sun, and Yu Su (2023). “Biomedical Language Models are Robust to Sub-optimal Tokenization”. *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 350–362 (cited on page 203).

Keith Harrigian, Tina Tang, Anthony Gonzales, Cindy X Cai, and Mark Dredze (2023a). “An Eye on Clinical BERT: Investigating Language Model Generalization for Diabetic Eye Disease Phenotyping”. *Machine Learning for Health (ML4H): Findings* (cited on pages 13, 196, 267, 301).

Keith Harrigian, Ayah Zirikly, Brant Chee, Alya Ahmad, Anne Links, Somnath Saha, Mary Catherine Beach, and Mark Dredze (2023b). “Characterization of Stigmatizing Language in Medical Records”. *Proceedings of the 61st Annual*

Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 312–329 (cited on pages 13, 203, 226, 301).

Hossein Hassani and Emmanuel Sirmal Silva (2023). "The role of ChatGPT in data science: how ai-assisted conversational interfaces are revolutionizing the field". *Big data and cognitive computing* 7.2, page 62 (cited on page 296).

Mahta HassanPour Zonoozi and Vahid Seydi (2023). "A survey on adversarial domain adaptation". *Neural Processing Letters* 55.3, pages 2429–2469 (cited on page 51).

Carrie Henning-Smith, Gabriella Meltzer, Lindsay C Kobayashi, and Jessica M Finlay (2023). "Rural/urban differences in mental health and social well-being among older US adults in the early months of the COVID-19 pandemic". *Aging & Mental Health* 27.3, pages 505–511 (cited on page 82).

Dieuwke Hupkes, Mario Giulianelli, Verna Dankers, Mikel Artetxe, Yanai Elazar, Tiago Pimentel, Christos Christodoulopoulos, Karim Lasri, Naomi Saphra, Arabella Sinclair, et al. (2023). "A taxonomy and review of generalization research in NLP". *Nature Machine Intelligence* 5.10, pages 1161–1174 (cited on page 159).

Md Saroar Jahan and Mourad Oussalah (2023). "A systematic review of hate speech automatic detection using natural language processing". *Neurocomput.* 546.C. ISSN: 0925-2312. DOI: 10.1016/j.neucom.2023.126232. URL: https://doi.org/10.1016/j.neucom.2023.126232 (cited on pages 249, 254).

BIBLIOGRAPHY

Alistair Johnson, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark (2023). *MIMIC-IV-Note: Deidentified free-text clinical notes* (cited on page 255).

Hadas Kotek, Rikker Dockum, and David Sun (2023). "Gender bias and stereotypes in large language models". *Proceedings of The ACM Collective Intelligence Conference*, pages 12–24 (cited on page 30).

Ida Kupcova, Lubos Danisovic, Martin Klein, and Stefan Harsanyi (2023). "Effects of the COVID-19 pandemic on mental health, anxiety, and depression". *BMC psychology* 11.1, page 108 (cited on page 165).

Eric Lehman, Evan Hernandez, Diwakar Mahajan, Jonas Wulff, Micah J Smith, Zachary Ziegler, Daniel Nadler, Peter Szolovits, Alistair Johnson, and Emily Alsentzer (2023). "Do We Still Need Clinical Language Models?" *Conference on Health, Inference, and Learning*. PMLR, pages 578–597 (cited on pages 196, 202, 226, 228).

Yizhi Liu, Weiguang Wang, Ritu Agarwal, et al. (2023). "People Talking and AI Listening: How Stigmatizing Language in EHR Notes Affect AI Performance". *arXiv preprint arXiv:2305.10201* (cited on page 300).

Nils Lukas, Ahmed Salem, Robert Sim, Shruti Tople, Lukas Wutschitz, and Santiago Zanella-Beguelin (2023). "Analyzing Leakage of Personally Identifiable Information

in Language Models". *2023 IEEE Symposium on Security and Privacy (SP)*. IEEE Computer Society, pages 346–363 (cited on page 229).

Pingchuan Ma, Rui Ding, Shuai Wang, Shi Han, and Dongmei Zhang (2023). "InsightPilot: An LLM-Empowered Automated Data Exploration System". *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 346–352 (cited on page 296).

Noah Thomas McDermott, Junfeng Yang, and Chengzhi Mao (2023). "Robustifying Language Models with Test-Time Adaptation". *ICLR 2023 Workshop on Pitfalls of limited data and computation for Trustworthy ML* (cited on page 299).

Elyse C Mead, Carol A Wang, Jason Phung, Joanna YX Fu, Scott M Williams, Mario Merialdi, Bo Jacobsson, Stephen Lye, Ramkumar Menon, and Craig E Pennell (2023). "The Role of Genetics in Preterm Birth". *Reproductive Sciences* 30.12, pages 3410–3427 (cited on page 16).

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz (2023). "Capabilities of gpt-4 on medical challenge problems". *arXiv preprint arXiv:2303.13375* (cited on page 230).

Bhargavi Paranjape, Scott Lundberg, Sameer Singh, Hannaneh Hajishirzi, Luke Zettlemoyer, and Marco Tulio Ribeiro (2023). "Art: Automatic multi-step

reasoning and tool-use for large language models". *arXiv preprint arXiv:2303.09014* (cited on page 296).

Juergen Pfeffer, Angelina Mooseder, Jana Lasser, Luca Hammer, Oliver Stritzel, and David Garcia (2023). "This Sample seems to be good enough! Assessing Coverage and Temporal Reliability of Twitter's Academic API". *Proceedings of the International AAAI Conference on Web and Social Media*. Volume 17, pages 720–729 (cited on page 81).

Hasin Rehana, Nur Bengisu Çam, Mert Basmaci, Yongqun He, Arzucan Özgür, and Junguk Hur (2023). "Evaluation of GPT and BERT-based models on identifying protein-protein interactions in biomedical text". *arXiv preprint arXiv:2303.17728* (cited on page 228).

Olivia Schuman and Haven Gabrielle Romero (2023). "Using Patient Quotations in Chart Notes: A Clinical Ethics Perspective". *The Journal of Clinical Ethics* 34.4, pages 352–355 (cited on page 301).

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. (2023a). "Large language models encode clinical knowledge". *Nature* 620.7972, pages 172–180 (cited on pages 203, 230).

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, et al. (2023b). "Towards expert-level medical question answering with large language models". *arXiv preprint arXiv:2305.09617* (cited on page 53).

Andrés L Suárez-Cetrulo, David Quintana, and Alejandro Cervantes (2023). "A survey on machine learning for recurring concept drifting data streams". *Expert Systems with Applications* 213, page 118934 (cited on page 41).

Ritu Thamman, Celina M Yong, Andrew H Tran, Kardie Tobb, and Eric J Brandt (2023). *Role of Artificial Intelligence in Cardiovascular Health Disparities: The Risk of Greasing the Slippery Slope* (cited on page 27).

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting (2023). "Large language models in medicine". *Nature Medicine*, pages 1–11 (cited on page 203).

Robert Tinn, Hao Cheng, Yu Gu, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon (2023). "Fine-tuning large neural language models for biomedical natural language processing". *Patterns* 4.4 (cited on page 220).

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal

Azhar, et al. (2023). “Llama: Open and efficient foundation language models”. *arXiv preprint arXiv:2302.13971* (cited on pages 5, 228, 230, 295).

Michiko Ueda, Kohei Watanabe, and Hajime Sueki (2023). “Emotional distress during COVID-19 by mental health conditions and economic vulnerability: retrospective analysis of survey-linked Twitter data with a semisupervised machine learning algorithm”. *Journal of Medical Internet Research* 25, e44965 (cited on page 165).

Soniya Vijayakumar (2023). “Interpretability in Activation Space Analysis of Transformers: A Focused Survey”. *arXiv preprint arXiv:2302.09304* (cited on page 47).

Kasumi Widner, Sunny Virmani, Jonathan Krause, Jay Nayar, Richa Tiwari, Elin Rønby Pedersen, Divleen Jeji, Naama Hammel, Yossi Matias, Greg S Corrado, et al. (2023). “Lessons learned from translating AI from development to deployment in healthcare”. *Nature Medicine* 29.6, pages 1304–1306 (cited on page 76).

Aaron A Wiegand, Taharat Sheikh, Fateha Zannath, Noah M Trudeau, Vadim Dukhanin, and Kathryn M McDonald (2023). ““It’s probably an STI because you’re gay”: a qualitative study of diagnostic error experiences in sexual and gender minority individuals”. *BMJ Quality & Safety* (cited on page 3).

Michael Wornow, Yizhe Xu, Rahul Thapa, Birju Patel, Ethan Steinberg, Scott Fleming, Michael A Pfeffer, Jason Fries, and Nigam H Shah (2023). “The shaky

foundations of large language models and foundation models for electronic health records". *npj Digital Medicine* 6.1, page 135 (cited on pages 203, 228).

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann (2023). "Bloomberggpt: A large language model for finance". *arXiv preprint arXiv:2303.17564* (cited on pages 53, 202).

Yutao Zhu, Huaying Yuan, Shuting Wang, Jiongnan Liu, Wenhan Liu, Chenlong Deng, Zhicheng Dou, and Ji-Rong Wen (2023). "Large language models for information retrieval: A survey". *arXiv preprint arXiv:2308.07107* (cited on page 296).

John W Ayers, Zechariah Zhu, Keith Harrigian, Gwenyth P Wightman, Mark Dredze, Steffanie A Strathdee, and Davey M Smith (2024). "Managing HIV during the COVID-19 pandemic: a study of help-seeking behaviors on a social media forum". *AIDS and Behavior* 28.4, pages 1166–1172 (cited on pages 13, 160).

Stella Biderman, USVSN PRASHANTH, Lintang Sutawika, Hailey Schoelkopf, Quentin Anthony, Shivanshu Purohit, and Edward Raff (2024). "Emergent and predictable memorization in large language models". *Advances in Neural Information Processing Systems* 36 (cited on page 295).

Cai, Cindy X, Keith Harrigian, Diep Tran, Tina Tang, Anthony Gonzales, Paul Nagy, Hadi Kharrazi, and Mark Dredze (2024). "Improving the identification of diabetic

retinopathy and related conditions using natural language processing methods". *Under Review* (cited on pages 14, 294).

Cory J Cascalheira, Santosh Chapagain, Ryan E Flinn, Dannie Klooster, Danica Laprade, Yuxuan Zhao, Emily M Lund, Alejandra Gonzalez, Kelsey Corro, Rikki Wheatley, et al. (2024). "The lgbtq+ minority stress on social media (missom) dataset: A labeled dataset for natural language processing and machine learning". *Proceedings of the International AAAI Conference on Web and Social Media*. Volume 18, pages 1888–1899 (cited on page 82).

Hyewon Jeong, Sarah Jabbour, Yuzhe Yang, Rahul Thapta, Hussein Mozannar, William Jongwon Han, Nikita Mehandru, Michael Wornow, Vladislav Lialin, Xin Liu, et al. (2024). "Recent Advances, Applications, and Open Challenges in Machine Learning for Health: Reflections from Research Roundtables at ML4H 2023 Symposium". *arXiv preprint arXiv:2403.01628* (cited on pages 14, 295).

Yanis Labrak, Mickaël Rouvier, and Richard Dufour (2024). "A Zero-shot and Few-shot Study of Instruction-Finetuned Large Language Models Applied to Clinical and Biomedical Tasks". *Fourteenth Language Resources and Evaluation Conference (LREC-COLING 2024)* (cited on page 228).

BIBLIOGRAPHY

Changmao Li and Jeffrey Flanigan (2024). “Task contamination: Language models may not be few-shot anymore”. *Proceedings of the AAAI Conference on Artificial Intelligence*. Volume 38. 16, pages 18471–18480 (cited on page 295).

Jingyang Li and Guoqiang Li (2024). “The Triangular Trade-off between Robustness, Accuracy and Fairness in Deep Neural Networks: A Survey”. *ACM Computing Surveys* (cited on page 53).

Jinyang Li, Binyuan Hui, Ge Qu, Jiaxi Yang, Binhua Li, Bowen Li, Bailin Wang, Bowen Qin, Ruiying Geng, Nan Huo, et al. (2024). “Can llm already serve as a database interface? a big bench for large-scale database grounded text-to-sqls”. *Advances in Neural Information Processing Systems* 36 (cited on page 296).

Zoltan P Majdik, S Scott Graham, Jade C Shiva Edward, Sabrina N Rodriguez, Martha S Karnes, Jared T Jensen, Joshua B Barbour, and Justin F Rousseau (2024). “Sample Size Considerations for Fine-Tuning Large Language Models for Named Entity Recognition Tasks: Methodological Study”. *JMIR AI* 3, e52095 (cited on page 76).

Michelle M Mello and Sherri Rose (2024). “Denial—Artificial Intelligence Tools and Health Insurance Coverage Decisions”. *JAMA Health Forum*. Volume 5. 3. American Medical Association, e240622–e240622 (cited on page 31).

Wenqi Shi, Ran Xu, Yuchen Zhuang, Yue Yu, Hang Wu, Carl Yang, and May D Wang (2024). "MedAdapter: Efficient Test-Time Adaptation of Large Language Models towards Medical Reasoning". *arXiv preprint arXiv:2405.03000* (cited on page 299).

Xindi Wang, Mahsa Salmani, Parsa Omidi, Xiangyu Ren, Mehdi Rezagholizadeh, and Armaghan Eshaghi (2024). "Beyond the Limits: A Survey of Techniques to Extend the Context Length in Large Language Models". *arXiv preprint arXiv:2402.02244* (cited on page 296).

Kristin Wolf and Julian Schmitz (2024). "Scoping review: longitudinal effects of the COVID-19 pandemic on child and adolescent mental health". *European child & adolescent psychiatry* 33.5, pages 1257–1312 (cited on page 165).

Changkun Ye, Russell Tsuchida, Lars Petersson, and Nick Barnes (2024). "Label Shift Estimation for Class-Imbalance Problem: A Bayesian Approach". *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1073–1082 (cited on page 44).

Yahan Yi, Keith Harrigian, Ayah Zirikly, and Mark Dredze (2024). "Are Clinical T5 Models Better for Clinical Text?" *Under Review* (cited on pages 14, 53, 298).

Wayne Xin Zhao, Jing Liu, Ruiyang Ren, and Ji-Rong Wen (2024). "Dense text retrieval based on pretrained language models: A survey". *ACM Transactions on Information Systems* 42.4, pages 1–60 (cited on page 296).

Vita

Keith Harrigian has a B.S. in Mathematics from Northeastern University and an M.S.E. in Computer Science from Johns Hopkins University. He has completed internships in data science and machine learning at Legendary Entertainment, True Fit, and Netflix. Prior to joining Johns Hopkins University, Keith was a Senior Quantitative Analyst at Warner Media where he leveraged machine learning and natural language processing to mine social media data for applications to film and television marketing. Keith's research focuses on the development of robust machine learning and natural language processing models in the health domain. He is particularly interested in documenting and counteracting the effects of distribution shift and selection bias.