HUMAN-CENTERED ALGORITHMS AND ETHICAL PRACTICES TO UNDERSTAND DEVIANT MENTAL HEALTH BEHAVIORS IN ONLINE COMMUNITIES

A Thesis Proposal
Presented to
The Academic Faculty

By

Stevie Chancellor, M.A

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Interactive Computing

Georgia Institute of Technology

August 2019

HUMAN-CENTERED ALGORITHMS AND ETHICAL PRACTICES TO UNDERSTAND DEVIANT MENTAL HEALTH BEHAVIORS IN ONLINE COMMUNITIES

Approved by:

Dr. Munmun De Choudhury, Advisor
School of Interactive Computing
Georgia Institute of Technology

Dr. Amy Bruckman
School of Interactive Computing
Georgia Institute of Technology

Dr. Eric Gilbert
School of Information
University of Michigan

Dr. Scott Counts
Urban Innovation Initiative
Microsoft Research

Dr. Wanda Pratt
School of Information
University of Washington

Date Approved: June 4, 2019

To Mom, who has had my back through thick and thin

ACKNOWLEDGEMENTS

I'm so grateful for all the encouragement I've received throughout this journey. I would first like to thank my advisor, Munmun De Choudhury, for your constant and unwavering support. Your mentorship and guidance has provided a research home that I didn't know existed, your knowledge and research practices are now a foundation for my own work, and your moral support pulled me through the best and worst times of a PhD. Thank you for taking me as your first PhD student. I hope I can be half the researcher/mentor you are.

I would also like to thank my thesis committee, Eric Gilbert, Amy Bruckman, Scott Counts, and Wanda Pratt. Eric and Amy, your support and advice throughout my whole PhD has been so helpful in shaping my career (even before I decided to come to Georgia Tech). Scott and Wanda, your wisdom and advice later in the process has been instrumental in developing my research trajectory and crafting a narrative for my research.

I've also been fortunate to have an amazing network inside and outside of Georgia Tech. The *SocWeB lab* has been fantastic for pushing my thinking and providing much needed distractions and fun - thank you to Sindhu Ernala, Koustuv Saha, Dong Whi Yoo, Vedant Das Swain, and Benjamin Sugar. Michaelanne Dye, you have been my rock through the PhD as we completed together. To others at GT who have been so supportive – Ari Schlesinger, Eshwar Chandrasekharan, Sucheta Ghosal, Shagun Jhaver, Miranda Parker, and Ian Stewart – thanks for the laughs, feedback, and support! My research assistants have also been indispensable in conducting this work - Jerry Lin, Andrea Hu, Francisco Zampieri, Tahirah Ahmad, thank you for your hard work and dedication and for learning with me as I learned to mentor you. To those outside of Georgia Tech (Harman Kaur and Natalie Simpson), thank you for being an external sounding board and stress reliever.

Finally, to my siblings - Tracey, Katie, Sarah, and Sam - thank you for always cheering me on no matter what, despite being several states away. Mom and Ed, thank you for your love, guidance, and cheerleading throughout this process. I would not be the person I am without your perseverance and encouragement throughout my life. And most of all, to Owen - we did it! Thank you for being a supportive, patient partner and husband. I can't wait to see where we fly next.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

Social media has changed how individuals cope with health challenges in complex ways. Especially for stigmatized mental health conditions like depression, online communities often offer positive health outcomes and behavior change. However, in some mental health communities, individuals promote deliberate self-injury, disordered eating habits, and suicidal ideas as acceptable choices rather than dangerous actions. These communities pose risks for members who engage in and encourage these behaviors, conflicting interactions with outsiders, and risks to social networks who moderate these groups.

My research focused on the pro-eating disorder (pro-ED) community, a clandestine group that advocates for eating disorders as lifestyle choices rather than a dangerous and potentially life-threatening mental illnesses. Computational approaches in machine learning have made strides in identifying and removing related deviant behaviors, like spam and abusive content. However, data-driven approaches alone oversimplify the complexities of mental disorders and the unique effects these communities have on people and on platforms. Mental disorders are unique to the individual, yet these communities also have larger platform impacts that resist simplistic approaches to intervention. Understanding these communities, how they interact with these multifaceted stakeholders, and doing so compassionately is key if we are interested in designing interventions to promote sustainable or healthier behaviors.

This thesis develops human-centered algorithmic approaches to understand these deviant and dangerous behaviors on social media. Using large-scale social media datasets and techniques like machine learning, computational linguistics, and statistical modeling, I analyze and understand patterns of behavior in pro-ED communities, how they interact with others on the platform, and these latent impacts. I blend methods contributions from data-driven fields like machine learning, natural language processing, and data science with collaborations and theoretical insights from psychology and sociology. Towards my human-centered research focus, I also consider the impacts that methods, ethics, and practices of conducting this work have on these communities.

Through eight empirical examinations and an analytical essay, I demonstrate that computational approaches can identify pro-ED and related behaviors on social media as well as documenting larger-scale community and platform changes and interactions with dangerous content. From the computational analyses, three themes and contribution areas emerge: developing rigorous inference of mental health status; using quantitative methods to understanding normative behavior and deviance; and better contextualizing content moderation and management of dangerous communities. Following this work and my interest in human-centeredness, I also propose a taxonomy and conduct several empirical studies on the emergent methods and ethics practices within the broader field of inferring mental health status on social media.

This work brings data-driven methods to Human Computer Interaction and Social Computing, a human-centered approach to improve Data Science and Machine Learning, and improves our understanding of moderation, normative behavior, and deviance in socio-technical systems. My thesis contributes novel methods and pipelines to understand mental health and other complex behaviors on social media, and empirical knowledge and design guidelines for ”deviance” and its manifestation online. For human-centered research, this thesis is an important step towards an agenda for conducting human-centered AI and ML research. In sum, this thesis represents the beginnings of an interdisciplinary approach to problem-solving for complex, vulnerable communities on social media.

CHAPTER 1
INTRODUCTION

Research overwhelmingly shows that online communities can promote healthy behaviors and outcomes for management of disease and disorders [1]. Especially for stigmatized mental health conditions like depression, online groups provide guidance, advice, and a sense of community. In many cases, participation in these groups and communities results in positive outcomes for those suffering from mental illness [2, 3].

Although most use these communities for positive behavior change, other communities advocate for dangerous health behaviors. In some mental health communities, individuals promote deliberate self-injury, disordered eating habits, and suicidal ideation as acceptable and desirable choices rather than dangerous actions that are a consequence of mental distress. These behaviors are situated in a "supportive" environment, where support meaningfully promotes these identities.

My research focuses on one example of these groups, **pro-eating disorder (pro-ED) communities** across social media sites. In pro-ED communities, users promote the maintenance and glorification of eating disorders and actively eschew recovery [4]. They share advice on maintaining extreme calorie deficits, exercise plans, techniques to avoid disclosing their disorder to parents and friends, and "thinspirational" photos of thin, emaciated individuals [5, 6].

There are numerous reasons why pro-ED content and communities is worrisome for social networks, yet causes complex interactions between individuals, communities, and influence of behavior. Disordered eating behaviors, such as starvation, weight control, binging, and purging, are extremely dangerous for individuals themselves, as eating disorders have some of the highest mortality rates of any mental disorder [7], and such behaviors can cause lifelong health complications, even after recovery is achieved [8]. In addition to promoting self-harm by sharing this content online [9], research also shows that “living” through others’ self harm experiences can promote these behaviors after engaging with the content [10]. For those who do not suffer from an eating disorder but encounter this content, these behaviors can spread to other people and are considered “contagious” in social groups [4, 11] – outsiders adopt disordered body image and undesirable relationships and behaviors with food. However, for some, engaging with these behaviors is therapeutic [12] and can dissuade them from self-injury. Pro-ED communities can foster conversation about mental health as well, engaging in some behaviors that are considered therapeutic. Additionally, the entanglement of pro-ED content with those who suffer from but *do not promote* eating disorders makes identifying these health signals challenging.

Because of these social complexities and in response to outsider reproach [13, 14], pro-ED communities frequently adopt complex methods to “hide in plain sight” on publicly accessible social media platforms. By using techniques like changing their language, creating coded and convoluted vocabularies, and making “private” groups and forums, these communities avoid outsider infiltration and detection. Social networks sites struggle with finding and moderating such content because of the clandestine, evasive techniques these groups use to avoid detection. Coupled with an increasing volume of data, online social platforms face an increasing task of identifying content that violates policies and guidelines. Platforms often do not have enough moderators to manage the volume of reports for social networks already [15], let alone triage pro-ED content that is infrequently reported to site-wide moderators.

These pro-ED communities also exemplify the tensions in care, ethics, responsibility, and societal guidelines of how to interact and intervene in these scenarios. Pro-ED communities and these interactions raise challenging questions about care and responsibility, such as whose obligation it is to intervene (social networks? community administrators? friends and family?) and what are effective and compassionate methods of interventions. Related to this are new questions on conducting ethical research of public data on online platforms: does the sensitivity of the subject influence what research is allowed to happen? What are new standards for rigor, accuracy, and methods in finding and predicting on these communities?

With recent advances in computational approaches like machine learning, computational linguistics, and their use to detect mental health [16, 17], there is now an opportunity to find these otherwise clandestine communities, assess their well-being, and (potentially) offer the necessary help they may need. Pro-ED communities are large, in some cases with tens of millions of posts, and manual techniques for understanding through qualitative work do not scale. Second, computational approaches in machine learning have made strides in identifying related deviant behaviors, like spam and abusive content [18]. However, **data-driven approaches alone oversimplify the complexities of mental disorders and the unique effects these communities have on people and on platforms**. Mental disorders are unique to the individual, yet these communities also have larger platform impacts that resist simplistic approaches to intervention. Ethical tensions are also mounting, asking data-driven research to consider the appropriateness, privacy concerns, and well-being of participants within these communities. What is needed to solve this is a solution that blends computational approaches with human-centered insights about ethics and practices in these domains.

In my research, I use pro-ED as a case study to understand *deviant mental health behaviors* on social media. I deliberately draw on the sociological notion of deviance, or actions that violate the norms and behaviors of a particular community [19]. The study of deviance has sought to understand the motives, culture, and outcomes of socially undesirable behavior [20]. Complementing deviance is the study of normative behavior, or behavioral expectations of groups and communities [21]. In this case, pro-ED behaviors violate many norms, including overriding the self-preservation instinct by self-harming [22], active encouragement of illness [4], and social violations of body image and gender [23, 24, 4]. In addition to violating norms held by individuals, pro-ED behaviors violate the norms of platforms by breaking Terms of Service that explicitly prohibit such behaviors on sites like Facebook, Tumblr, and Instagram [25, 26]; yet, whether social networks should moderate pro-ED and related self-injury behaviors or intervene is controversial [27].

The methods I adopt in my approach are primarily computational, blending methodological contributions from data-driven fields like machine learning, natural language processing, and data science with collaborations and theoretical insights from fields like psychology and sociology. My research approaches builds rigorous applications of data-driven methods to scale to millions of posts on social media that accurately identify and assess these deviant mental health behaviors. Using pro-ED as a case study, I also consider the impacts of this research and the ethics and practices that go into conducting it, and how we can best make decisions.

To understand these issues within pro-ED communities and deviant mental health behaviors, my work explores the following themes and asks these research questions:

- **Inferring Health Statuses**. How might we detect these pro-ED communities? Can we both identify these deviant behaviors and assess the well-being of participants? In Chapter 3, I explore the application of novel methods and approaches to understanding pro-ED communities on Instagram and Tumblr, and explore their use for social computing and online communities.

- **Normative Behavior and Deviance.** How are pro-ED communities engaging in deviant behavior? What are norms of behavior in and around these communities? I consider these questions in Chapter 4, where I explore techniques of communities to avoid detection through deviant behavior, and the implications of norms on building predictive systems.

- **Content Moderation and Management**. What are the impacts of moderation on deviant mental health behaviors, and are these moderation strategies effective? Should we moderate or manage these behaviors, and whose responsibility is it to intervene?

I discuss machine learning approaches to identifying content on social media that may be candidates for removal on Tumblr and Instagram in Chapter 5, and the social implications to moderation in this chapter.

- **Ethics and Practices in this Field.** What are the emergent ethical issues of conducting research that identifies and assesses the well-being of individuals online? How should researchers consider their relationships with the humans who are the data subjects within this work? First, I discuss a taxonomy that I developed to understand mental health status in Chapter 6. Then, in Chapter 7, I describe the results of an empirical systematic literature review on the body of research in mental health status prediction on social media, and the two studies that have resulted from this research.

I provide an overview of my research and its location within the larger document in Table 1.1.

In short, I have conducted several computational studies to understand the behaviors and latent impacts of pro-ED communities in online communities that attempts to do so in human-centered ways. These studies have used machine learning, computational linguistics, and statistical modeling to model and examine the behaviors of individuals and community dynamics. Outcomes of this human-centered approach in my work include a focus on ethics and privacy for participants, methods choices that focus on the community as the interest, and a commitment to ethics and practices within the field. Finally, I tie this research agenda together with several studies on the research, practices, and methods within the field, to help guide better outcomes for community members and practices for the fields of computational social science, social computing, data science, and HCI.

1.1 Contributions

At a high level, my work contributes insights towards the development of a broader **human-centered paradigm** for data-driven, predictive research about individuals and communities [35]. Through my work on pro-ED communities, I have begun to develop an approach to problem solving that puts the human at the center, necessitated by the community's sensitivities and emergent challenges. Pro-ED communities have been an illuminating case study for the challenges and constraints of methods, ethics, privacy, study design, and implications of building predictive techniques for understanding human behavior.

This human-centered paradigm is a comprehensive approach to solving research problems that deliberately refocus on the needs of people, communities, and society, rather than on the data or methods or scientific outcomes, and identifying the appropriate tools to solve problems. The human-centered computational approaches I adopt are more than just developing human-machine hybrid systems, collaborations with domain experts, or bringing "humans in the loop," although these feature prominently in my work. This ties into my vested interest in doing right by the people in these communities—as I have developed my research agenda and seen the implications of this work unfold in my own work and in popular press, the urgent need for this kind of approach is evident.

Specifically, I advance this human-centered paradigm through contributions to four primary areas: computational understandings and analysis of online behavior; empirical insights into bad behavior in online communities through pro-ED; a taxonomy and guidance on how to conduct more rigorous, ethical, and human-centered computational research; and insights that promote the design of interventions and improved clinical outcomes.

1. **Computational Contributions to Understanding Online Communities and Behavior.** In my work, I bring novel computational techniques to HCI and social computing to identify and understand deviant mental health behaviors on social media. Some of these techniques include deep learning, like image processing and deep neural networks, transfer learning, and computational linguistic techniques like word embeddings.

   Outside of porting methods to HCI, my work proposes novel techniques for improving human-centered methods to this work. I adopt combinations of human-in-the-loop systems for data labeling, curation, and human evaluation of computational techniques. Using the power of expert curation, manual labeling, and error analyses, I design and test pipelines to acquire challenging data, label data precisely and in alignment with clinical standards, and evaluate the output of computational systems. These techniques can be used on problems outside of deviance or mental health, and can help researchers analyze complex online communities and scale up annotation rates.

2. **Empirical Insights Into Deviant Behavior Online.** Pro-ED is an outstanding case study to understand deviant behavior on online social platforms. Pro-ED manifests in ways similar to other kinds of deviant behavior online, taking evasive techniques to avoid detection. Additionally, these communities and their behaviors challenge social networks with their content's impact on outsiders, adversarial relationships with other online communities, and the latent impacts on social networks and society. My research provides empirical insights into pro-ED communities and mental health.

   In addition, my empirical work in this space speaks to very urgent questions around content management, moderation, and promoting healthier and safe communities, an active and thriving area within CSCW, HCI, and in public discourse more broadly. My research indicates that straightforward approaches to content management and moderation are not effective when managing deviant mental health communities like pro-ED [28, 30]. Pro-ED is a case study for a special kind of deviant behavior online that in many ways echoes the concerns of other kinds of content, such as extremist, abusive, or dangerous content on social media. Yet, my research also speak to social concerns of how we as a society should address issues of dangerous behaviors. My prior work touches on issues of free speech and action around the right to self-harm, privacy, and societal responsibility to help those who are struggling. In my work, I provide support for new strategies to manage this content. Although these are tailored towards mental health communities, I see many of these strategies being effective for combating other kinds of pernicious behaviors in online space.

3. **Helping Set Standards for Better Ethics and Practices.** As computational techniques become more ubiquitous in human-centered problem solving, we will need standards of conducting such research in ethical and rigorous ways. My computational work helps in modeling how to conduct human-centered research.

   My interest in ethics and practices for predicting mental health status is also important. The taxonomy of mental health status I developed brings to light many of the concerns (both theoretical and practical) of using social media to predict mental health status on social media data. The empirical contributions of the literature review, and the analyses that have resulted, are key to this agenda of raising ethical "consciousness" because they empirically demonstrate the gaps that exist in current practices. I see this literature review as influencing inter-disciplinary fields of computer science to more critically consider how to maintain rigorous and ethical standards.

4. **Informing Human-Centered Design of Social Systems.** Thinking more broadly to HCI and other fields, I also see my work influencing how designers and clinicians engage with these communities in providing meaningful support. Through the new understandings gathered from my work, I can help online communities and social networks develop better moderation systems and community guidelines that restrict the sharing of such dangerous mental health content. For designing moderation systems, my work on pro-ED communities can help both avoiding sharing dangerous behaviors and also encourage positive online communities.

   For clinicians who are concerned about such content, my work also informs their own practices for managing patients who may participate on these social networks. As people move to therapy, clinicians will need to understand dynamics of support that their patients receive offline and online. For online support, this may mean finding those encouraging of recovery or it may mean finding subversive support for dangerous pro-ED behaviors. The interplay between these communities and managing mental illnesses will influence treatment, and my work helps clinicians understand these trends and make better decisions for their patients. Our understandings of the pro-ED community can help researchers better understand the intricacies of pro-ED sentiments and their relationships to eating disorders.

Overall, I aim to contribute empirical understandings of deviant mental health behaviors to online communities, and insights that can inform better design and practice of using computational methods to understand bad behavior in online communities.

1.2 Overview of Thesis

This thesis is organized as follows. This chapter introduces the thesis and provides my contributions to the fields of social computing, data science, and HCI.

Chapter 2 is a literature review explaining my subject of interest, the pro-ED community in social networks, and my motivations for selecting them as a topic.

In the next three chapters (Chapters 3, 4, and 5), I review my computational studies on pro-ED communities in light of the thematic contributions of my work. In light of the three thematic areas of interest I identify, each chapter contains two projects most relevant to that theme, prior work related to them, and implications. Chapter 3 overviews computational methods for inferring mental health statuses of individuals online. Chapter 4 overviews norms and deviant behavior and their study in online communities. Chapter 5 discusses online content moderation and community management.

Next, in Chapter 6, I move to the 4th area of contributions, understanding the ethical and methods concerns with conducting predictive research on mental health status from social media data. I present here a taxonomy of the ethical, methods, and collaborative challenges this research poses to the interdisciplinary fields involved.

Chapter 7 contains the empirical research of a systematic literature review of this space and my findings. This contains two studies of a corpus of 55 papers on the topic. First, I discuss the findings exploring the study design, methods, and algorithmic decisions made to conduct this work. Then, I examine the representations of the human "research subject" within computational studies of online mental health status.

In Chapter 8, I reflect on this research. I discuss potential impacts to design of online communities and social systems, provide some recommendations for conducting more ethical and rigorous work in this space, outline one potential agenda for human-centered algorithmic approaches, limitations, future work. Finally, I conclude by summarizing the findings and contributions of my research.

Table 1.1: Overview of research within this thesis, organized by project.

Study	**Thematic Area**	**Summary**	**SNS**	**Location**
#thyghgapp [28]	Norms and Deviance	Understanding the lexical, behavioral, and engagement changes of the pro-ED community who use lexical variants of banned hashtags.	Instagram	Ch 3
Mental Illness Severity [29]	Computational Techniques	Developing a computational approach to assess and measure mental illness severity in the pro-ED community.	Instagram	Ch 5
“This Post Will Just Get Taken Down” [30]	Moderation and Management	Predicting whether a post will be removed or deleted	Instagram	Ch 4
Anorexia Recovery [31]	Computational Techniques	Assessing the likelihood and predictors for anorexia recovery after engaging in pro-ED content	Tumblr	Ch 5
Multimodal Removal [32]	Moderation and Management	Implementing a deep learning system using text and image data to predict if posts will be removed for violating community guidelines around self-harm	Instagram	Ch 4
Norms Matter [33]	Norms and Deviance	Understanding norms of behavior in two online weight loss communities	Reddit	Ch 3
A Taxonomy [34]		Overview of methods and ethics issues in predicting mental health with social media	n/a	Ch 6
Empirical Analysis for Methods (*in prep*)		Systematic review of methods, study design, and algorithm selection	n/a	Ch 7
Who is the Human? (*under submission*)		Examining representations of the human research subject in this work	n/a	Ch 7

CHAPTER 2
AN OVERVIEW OF PRO-EATING DISORDER COMMUNITIES

In this chapter, I provide a brief overview of the literature in eating disorders and online pro-ED communities. This is an interdisciplinary research area across sociology, clinical psychology, psychology, and HCI. Then, I discuss the motivations for studying pro-ED communities as an exemplar for deviant mental health behaviors.

2.1 Eating Disorders

Eating disorders are a genre of psycho-social and behavioral disorders characterized by both obsessions with weight and body image as well as abnormal behaviors and preoccupations with eating and exercise [8]. According to the Diagnostic and Statistical Manual of Mental Disorders (DSM-V), symptoms include food restriction, bingeing, purging, avoiding certain foods, obsession about weight and body image, and other extreme emotional responses to eating, exercise, and body image [8]. Anorexia nervosa and bulimia nervosa are the two well-known eating disorders specified in the DSM-V, but the most commonly diagnosed eating disorders are OSFEDs – other specified feeding or eating disorders (formerly known as EDNOS, or eating disorder not otherwise specified) [8]. Despite these differences and awareness in diagnosis, all eating disorder diagnoses are serious and pose major threats to health [8, 36].

In the US alone, it is estimated that 20 million men and 10 million men suffer from an eating disorder at some point in their life, yet only a small fraction of these expected cases are ever diagnosed and receive clinical care [36]. Eating disorders also have very high comorbidities (80%) with other mental health disorders, like depression, anxiety, and obsessive-compulsive disorder [37]. Even after clinical treatment, eating disorders pose long-term risks to health, such as heart problems later in life, bone density and early-onset osteoporosis, and complications with the esophagus and throat, leading to increased risk for cancers [38].

2.2 Pro-Eating Disorder Communities

Pro-eating disorder, or pro-ED communities, are communities that normalize eating disorders as alternative lifestyle choices. Users share restrictive dieting plans, techniques to conceal their symptoms or behaviors, and exchange "thinspiration" to maintain their disordered behaviors [4]. Pro-ED sentiments are particularly dangerous given that eating disorders have the highest mortality rate of any mental illness [36] and pro-ED users actively encourage others to maintain their condition and to avoid receiving treatment. There are variants of pro-ED behavior on social networking sites, focused on the specific disorder of the person (pro-anorexia or pro-bulimia). In some forums, these communities conflict with each other [13], but in most modern social networks, these groups blur together as the pro-ED community.

Eating disorders existed before digital communication; however, the emergence of pro-ED communities appears to be tightly tied to the rise of online social interactions. Once considered a fringe disorder, those with eating disorders who were physically disconnected moved online and now use the Internet to connect with other sufferers. It is not clear why pro-ED communities have become so prolific – some researchers imagine that these communities provide members with a strong sense of identity in spite of social rejection of their eating disorder behaviors [39]. Early studies of these communities were on pro-ana or pro-ED websites [40, 4], standalone forums and bulletin boards [41, 13], and blogs [5].

Pro-ED communities and behaviors are of interest across multiple fields, including sociology and clinical psychology. In sociology, research extensively examined how pro-anorexia and pro-ED communities adopt and manage their perceptions of self [42], their conceptions of their eating disorder [23, 24], as well as concepts of their own bodies [6, 43], and worthiness [44]. For health and clinical researcher, concern exists about how these behaviors promote eating disorder behaviors and avoidance of treatment [45, 46]. In recent research, Gerrard explored the standards for content moderation through a thematic analysis of multiple online communities [47].

Recently, computational researchers in HCI, computational social science, and similar fields have begun to study the interactions and behavior of these sites on social media websites. Early research scoped the problem of pro-ana on the microblogging site Tumblr [48], and further research conducted a more nuanced descriptive study of the anorexia community, both for pro-ED and also for recovery sentiments [49]. Across multiple platforms, researchers have qualitatively examined the presentation of these disorders [6]. Yom-Tov *et al.* identify the "warring tribes" of pro-recovery and pro-ED sentiments on Flickr pro-ana photo groups [14]. Finally, Syed-Abdul *et al.* consider pro-anorexia as misinformation when they analyze the content of YouTube videos [50].

2.3 Pro-ED Content and Social Networks

Social networks have been challenged by the presence of pro-ED communities. In many cases, pro-ED content and communities are banned or restricted because they promote self-injurious behaviors. Pro-ED content promotes disordered eating behaviors, like starvation or binging/purging that research argues is self-injury [51]. In some cases in these communities, individuals promote more explicit self-injurious behaviors like suicidal ideation and cutting [29]. Many social networks ban the promotion of violence to self or others, therefore pro-ED communities are frequently banned.

In the context of eating disorders, while there is no obvious moderation on eating disorder-related content on Twitter, YouTube, or Reddit, other platforms more rigorously ban tags and terms around it. Tumblr bans content that "content that urges or encourages others to...cut or injure themselves; [or to] embrace anorexia, bulimia, or other eating disorders" [25]. Tumblr additionally issues content PSAs on tags known to be associated with eating disorders, warning the user that graphic content may occur behind the tag. Similar to Tumblr, Instagram also bans content that "glorifies self-injury," saying that, "encouraging or urging people to embrace self-injury is counter to this environment of support, and [Instagram will] remove it or disable accounts if it's reported to [Instagram]." [52]. In addition, Instagram outwardly bans several pro-ED tags and provides content advisories on others [28].

2.4 Why Pro-ED for This Thesis?

Why are pro-ED communities a relevant case study for understanding computational methods to identify deviant mental health content? I provide several motivations, framed around three motivations I present in the Introduction – norms and deviance, computational techniques, and moderation.

Norms/Deviance. Pro-ED behaviors exhibit many behaviors that are considered deviant. Pro-ED is a well-established psychological phenomenon, documented across sociological and psychological literature. As for deviance and norms, it violates basic instincts against self-harm and illness [8], contrasts with societal expectations of health and body [43], and conflicts with and occasionally antagonizes outside communities [4, 14]. Even if these communities did no harm to others, the potential for contagion of these behaviors to others is alarming [4, 11].

Additionally, pro-ED communities challenge our normative understandings of health communities promoting universally positive goals [1, 3, 53]. Pro-ED communities deliberately promote subversive behaviors for health communities, challenging our norms of health content in online spaces.

Computational Techniques. Pro-ED behaviors and communities also present unique methodological challenges for identification, detection, and inference. Finding the content can be hard because of the obfuscation these communities take to avoid platform intervention [28]. Furthermore, not all content about eating disorders is pro-eating disorder, and disambiguating this content in social platforms is a challenging task. For the content that is posted by those who are pro-ED, not all of their content actively encourages the most dangerous behaviors and can represent a wide variety of emotional states.

Finally, pro-ED content is heavily multimodal (when the social network allows for it) – as text, images, videos, and audio. These images show explicit negative emotions and graphic content of thinness, motivation for starvation, and even pictures/GIFs of self-injury. This makes it more difficult to detect than a “signature” that only appears in one modality like text or behavioral cues.

Moderation and Management. These communities are clandestine on social networks, who actively avoid detection or interference. They use ever-evolving hashtag variants, community names, and other evasive techniques to avoid detection [28]. Additionally, asking moderators to manage these communities and posts manually is challenging because of staffing and labor issues [15].

Even if it were easy to identify them, the content itself is difficult to deal with. Unlike other content which may violate the rules of a site without being alarming, pro-ED content is graphic and depicts very emotionally-draining content. This content is multimodal, and in particular images and videos of graphic self-harm can be emotionally exhausting and traumatizing to deal with without adequate training [54].

CHAPTER 3
COMPUTATIONAL METHODS TO INFER MENTAL HEALTH STATUS

To assist in bringing pro-ED communities timely, meaningful, and useful assistance requires identifying and assessing their well-being and behavior. Recent advances in computational power, techniques, and analysis have made it possible to understand these behaviors in nuanced ways that can also be done across millions of posts and users.

In this chapter, I will discuss how to infer health status from social media data using computational methods, specifically the well-being and health of pro-ED communities. I will start by providing an overview of the methods used to infer mental health status from social media data. Then, I will discuss two of my prior projects that focused on inferring the health state in online pro-ED communities

The first project, "Quantifying and Predicting Mental Illness Severity in Online Pro-Eating Disorder Communities," discusses an extensive methodology to assess, predict, and quantify the severity of mental illnesses in pro-ED communities on Instagram. I offer a novel statistical method that combines topic modeling and novice/clinician annotations to infer MIS in a user's content. Alarmingly, I found that proportion of users whose content expresses high MIS have been on the rise since 2012 (13%/year increase). I show that past MIS in a users content over seven months can predict future risk with $\tilde{8}1\%$ accuracy. This model can also forecast MIS levels up to eight months in the future with performance better than baseline.

The second project, "Recovery Amid Pro-Anorexia," uses survival analysis to evaluate the influences on recovery on Tumblr pro-recovery communities after pro-ED participation and situates these results within the cognitive behavioral theory of anorexia. This model identifies content and participation measures that predict the likelihood of recovery. From the dataset of over 68M posts and 10K users that self-identify with anorexia, I found that recovery on Tumblr is protracted – only half of the population is estimated to exhibit signs of recovery after four years.

3.1 Background and Prior Work

In health research, computational analyses of online social media text to understand mental wellness behaviors online is an active area of research. Originating from broad predictions of well being and life satisfaction [55, 56], research on mental health has covered many different topics from public health [57] to analysis of specific conditions like depression [16, 58], post-traumatic stress disorder [17], and post-partum depression [59].

In this section, I will focus primarily on the emergent computational methods from mental health research and social computing, both deviant and socially acceptable.

3.1.1 An Overview of the State of the Art

The origins of predictive work come from either population-level analyses or studies of generalized and subjective well-being and affect assessment. Borrowing from advances in natural language processing [60] and psychology [61] to represent text as cues of well-being, these studies described mood shifts around political events [62], geographic differences in expressed well-being [55], and the seasonality and temporality of mood variation [63]. In addition to studying generalized well-being, researchers also assessed population happiness both on Twitter [64] and Facebook [65]. Besides establishing that psychological and health states can be inferred from this data, these findings show that people use social media to discuss their personal mood and activities honestly and candidly instead of their idealized versions [66]. Complementary to this research were studies in public health measurement with online data, termed "infodemiology." [67] This famously includes the use of human-generated data to predict influenza outbreaks through search engines [68]. Researchers also used social media data to track the spread of disease [69] and to analyze other ailments on population-scale user bases from Twitter [57].

Soon after these studies were the first predictive works on the mental health states of individuals, beginning with depression. In 2013, De Choudhury *et al.* used clinically validated depression measures to find Twitter users who tested for major depressive disorder [16]. They developed a model that could predict if someone had depression with 70% accuracy. Around the same time, Park *et al.* developed a mixed methods approach to understand how Facebook use corresponded to clinical scales for depression [70]. In 2014, Coppersmith *et al.* used self-reported disclosures of depression diagnosis on Twitter ("I was diagnosed with depression on...") to classify individuals suffering with depression, contrasting their language with those who do not self-report such diagnoses [58]. De Choudhury *et al.* also sought to identify new mothers who might be suffering from postpartum depression using Facebook and Twitter data [59, 71]. After these works, researchers began to replicate, extend, generalize, and improve on these findings [72, 73, 74] and in different cultural contexts and social media sites, beyond just English-speaking Twitter [75, 76].

From these seminal works on depression, new studies have investigated new psychiatric disorders, new social network platforms, and new modalities. Research has examined other disorders, such as post-traumatic stress disorder [17], anxiety [77], schizophrenia [78, 79], eating disorders [29, 31, 80], and suicidal ideation [81, 2, 82]. Work also now explores the symptomatology of mental disorders, such as the severity of mental illness [29, 73] and stress connected to mental health [83]. Datasets too have expanded to social networks other than Twitter and Facebook, like Sina Weibo [84], Instagram [85, 29], Tumblr [31, 86], and Reddit [87, 88]. Modalities other than text are now analyzed for their signals of mental health status. Automated image analysis can identify self-harm photos on Flickr [89], signs of depression through Instagram images [85], and mental health disclosures on Reddit [90]. Finally, new data sources have begun to supplement social media data, like active and passive sensing technologies [87].

3.1.2 What Traces and Methods are Available?

There are a variety of sources available in social media to study for inferring mental health. Text is the most common trace researchers analyze, as it is widely accessible across almost all social networks. Analysis methods for this data are extremely robust, especially when drawing from natural language processing and machine learning techniques [91, 92]. Drawn from text or behavior, researchers have also inferred emotions, such as the dimensions of affect/valence [93, 87].

Through exploration of interactions, another trace that researchers can use is behaviors on different social media platforms. Platforms allow for likes, retweets and reposting, sharing of content, and other behavioral cues of engagement [63]. Social interaction and interpersonal relationships are another focus of research. One good example of these interactions is in research for breast cancer communities and membership retention rates [94, 91].

For multimodal/non-text data, recent advances in machine learning and computer vision have made image analysis more scalable. Manikonda and De Choudhury have researched photos around disclosure [90], Reece and Danforth on markers of depression in images [85], and my prior work uses deep methods to understand pro-ED photos [32]. Video and audio are other modalities, appearing on sites like Vine, Instagram, and Facebook. However, these two traces are currently underexplored in social media and health literature.

Once the traces have been gathered, there are many methods and statistical techniques available to analyze these traces of mental health. Machine learning is frequently used in the field to make guesses about health states and mental health conditions [16]. The emphasis in these papers tends to be on developing intuitive or interpretable feature sets or models to understand how the algorithm can contribute to our understandings of mental health. Although rising in popularity in other fields, deep neural networks have only begun to start to analyze and identify health behaviors online [91, 32]. I argue that this is partially due to the interpretability issue I mentioned before.

In addition to machine learning, researchers also model trends of the data using different disciplinary strategies. Primarily pulling from NLP, studies use natural language processing and computational linguistic techniques to model word and topic relationships [95, 58]. These methods explore latent relationships between words and sentences in social media posts to understand symptomatology, the spread of disease, or signals of depression.

Various other statistical models have been used as well. Time series and related temporal analyses use time to predict items such as changes in mood and emotional stability over time [96, 97]. Work has also pulled from the medical health literature in survival analysis, a statistical technique to model the time to an event occurring. Work by Yang *et al.* , for instance, uses these models to understand support structures in breast cancer support communities [91] and interpreting time to death [98].

3.1.3 Ground Truth Verification Strategies

To infer mental health status on social media data, researchers must start with some kind of "ground truth" assessment, or information about the actual state of the person that the researchers intend to measure. In medical settings, doctors and clinicians have many tools to evaluate whether someone suffers from a given condition or symptom. This includes evaluating symptomatology, talking with the patient, observing their behaviors and reactions, and administering tests to evaluate for given conditions. In the online setting, however, researchers lack many of these options, as there is not necessarily direct engagement with a participant. To develop robust and valid methods of knowing what needs to be measure, social media researchers have developed several techniques to get these grounded assessments of states in social media data.

The first approach is *self-reported data* online, where the researchers identify key phrases that declare the state of interest. This was popularized by Coppersmith *et al.* in their work on detecting depression in Twitter data [58]. They identified users who disclosed that they suffered from Twitter using specific statements of disclosure, such as "I suffer from depression" or "I was diagnosed with depression."

Another validation measure is using *expert validations* identify individuals in the dataset that suffer from the state in question. Practicing doctors and clinicians are asked to annotate or label data. Alternatively, other experts in the research question are called in to validate and label data, like moderators who deal with sensitive online content. This approach was taken in previous work on schizophrenia [78].

Finally, researchers can also administer *tests and batteries* to users, then using the assessments from those batteries to determine the state of their mental health. These tests and instruments typically ask the patient a set of questions to assess the degree of severity of a mental health challenge. For example, the PHQ-9 assists in screening for depression.

Many projects use combinations of these techniques to further bolster their results. For instance, Birnbaum *et al.* use self-reported status of schizophrenia and expert ratings to guess when users may be experiencing psychosis [78]. In De Choudhury *et al.* 's work on depression, they recruit participants from Mechanical Turk, administer an instrument, then gather their social media data [16].

3.2 Quantifying Mental Illness Severity

The Diagnostic and Statistical Manual of Mental Disorders (DSM) [8] identifies specific behaviors and cognitions that promote self-injurious behavior, extreme weight control, and suicidal ideation. Such behaviors may be associated with specific mental illnesses like eating disorders. Based on the clinical psychology literature, I mapped manifestations of such behaviors in social media content to markers of *mental illness severity (MIS).*

Vast epidemiological and psychiatric evidence of MIS exists in clinical research [99, 100, 101, 102]; however, examinations of MIS on social media sites are limited. Most studies in CSCW or HCI have focused on identifying markers of a mental illness like depression [16, 58], not on MIS more broadly in any given mental illness-prone community.

The increasing pervasiveness of mental illness-related content on social media (such as Instagram, Tumblr, Twitter, and Reddit) now provides an opportunity for rigorous quantitative studies of markers of MIS on these platforms. The rich content on these sites may be used both to objectively quantify, measure, and characterize levels of MIS broadly and also to examine the distribution of such content over previously inaccessible timescales and population sizes.

In this paper, I study, estimate, and forecast MIS in users who share pro-ED content on Instagram. The pro-ED community glorifies eating disorders as alternative lifestyle choices rather than as psycho-social disorders [103]. Cognitions present in the pro-ED community include suicidal and thin ideal ideation, and behavioral activities include direct self injury [101] and extreme weight control behaviors (which some have termed indirect self-injury [104, 105]). The pro-ED community is an ideal community to study because up to 70% of individuals with eating disorders report a history of some form of deliberate self-injury [51]. Because of this, the community presents the largest variety of markers of MIS not present in communities focused on other mental disorders (*e.g.*, depression or obsessive-compulsive disorder) or specific activities (*e.g.*, cutting or suicidal ideation)

I make the following contributions:

- I present a novel, scalable, and robust method to quantify and characterize MIS in pro-ED Instagram content. This method employs topic modeling, specifically Latent Dirichlet Allocation on the tag content of pro-ED posts. I present a clinically-grounded framework to obtain novice and expert annotations on low, medium, and high MIS of the extracted topics.

- I develop an algorithm to combine automatically extracted topics and their severity scores into inferences of MIS levels manifested in the content of users over time.

- I present a supervised learning (regularized multinomial logistic regression) model to predict to what extent a user who shares pro-ED content on Instagram would share content with markers of low, medium, or high MIS in the future, based on MIS in their historical content.

This study uses a large dataset from Instagram spanning more than 26 million posts from 100,000 public users, specifically focused on those users who have used pro-ED tags between 2010 and 2015. My rating methodology that combines topic modeling with human annotations can measure MIS in posts with high accuracy, precision, and recall when compared against independent ratings from novices and experts. Second, MIS inferred from a user's posts over a seven month period is able to predict levels of MIS during a future month with over 81% accuracy. Our results show that despite Instagram-enforced efforts to curb dissemination of such vulnerable content, the proportion of pro-ED Instagram users sharing content with high levels of MIS has been on the rise (13% increase / year)

3.2.1 Data

I gathered a dataset of *public* posts related to eating disorders on Instagram using the official Instagram API. Note that Instagram does not have formalized community structures, like forums or private groups. Instead, communities form around more amorphous, public tags. In the case of the pro-ED community on Instagram, users cluster around tags relating to eating disorders (*e.g.*, "anorexia", "proana").

The data collection did not prioritize collecting content shared by tags directly associated with self-injury and suicide; those tags would bias the content and nature of the results. Moreover, searching specific self-injury or suicide ideation related tags would generate only a partial sample because the set of all possible MIS tags on Instagram is unknown. I used a combination of snowball sampling and human curation to gather 6.5 million posts relating to pro-ED. I use this as the initial sampling set to identify a user cohort to study MIS. This snowball sampling approach will be covered more extensively in Chapter 4 when discussing another project, but I provide a summarized version here.

First, I identified a set of nine "seed tags" that have been found to be common pro-

ED organizing tags and structures across social media platforms [49]. This includes "ed," "eating disorders," "ana," "anorexia," and "bulimia," among others. I then introduced a manual inspection phase where two researchers searched for posts on each of these nine tags to ensure there was sufficient pro-ED content. With these tags, I conducted an initial month-long crawl using Instagram's official API, identifying 222 total tags that had at least a 1% occurrence rate in the dataset.

Next we expanded the initial seed set by collating a list of all tags that co-occurred with the seed tags in the initial 434K posts. From this list, I manually checked and removed tags that did not map directly to eating disorders. This reduced the filtered co-occurrence tag list from 222 tags to 72 known to be related to eating disorders (see sample tags in Table 3.1). Then, I conducted a second longer crawl of pro-ED content focusing on these 72 tags.

Table 3.1: Example tags used for crawling pro-ED posts and users in the study.

skinny	thin	thinspo	bonespo
eatingdisorder	probulimia	anorexia	thighgap
proanorexia	mia	bulimia	promia
thinspiration	secretsociety	ana	proana
anorexianervosa			

This produced over about 8 million posts dated between January 2011 and November 2014. I removed any posts that were cross-posted to any recovery tag as well as any that had three tags ("mia," "ana," and "ed") that did not also contain another tag from the list of 72. The dataset at this phase had 6.5 million posts relating to pro-ED.

Gathering Pro-ED Users and Their Post Timelines. In the second step, to construct the candidate set of pro-ED users and their posts, I obtained a random sample of 100,000 users from the authors of the 6.5 million posts collected above. Again, I used the Instagram API to obtain the post timelines of each of the 100,000 users (all public posts of the users).

The final dataset contains over 26 million posts from 100,000 users, with post shared between October 2010 and March 2015.

3.2.2 Methods

Inferring Mental Illness Severity (MIS). A significant challenge in quantitative studies of any health risk or behavior is the availability of labeled content (ground truth) on *which users* are susceptible. In the absence of self-reported information on the mental health status of individuals, tags serve as a good indicator of whether a user's content expresses markers of high MIS. However, capturing the set of all tags related to MIS is difficult. Additionally, assessing in isolation a tag's MIS level may be difficult — *e.g.*, , "cutting" might be attributable to high MIS, however the tag "pain" is ambiguous. Furthermore, discrete human judgments on MIS may not be applicable to a user's individual posts, since a user may use their Instagram profile to share not only content with MIS but also on a variety of other topics.

To overcome these challenges in MIS inference, I adopted a hybrid approach where I leveraged both automated natural language processing techniques and human annotations. I employed Latent Dirichlet Allocation (LDA) [106] on all posts from all users in the compiled dataset. My goal was to obtain a set of topics spanning the content (tags) of the posts, some of which I suspected to be tied to increased MIS. This allowed me to go beyond simple tag-based MIS inference techniques — LDA would use all tags for topic inference including those that co-occur with known/unambiguous MIS tags. Moreover, LDA assumes each tag to be drawn from a mixture of topics instead of mapping each tag to a specific topic (MIS or otherwise). By using LDA, I was also able to obtain a posterior distribution of topics over posts of a user. This prevents assigning users to specific topics; instead I could model their posts as a distribution over content with varying levels of MIS.

My method proceeded as follows:

(1) Topic Inference. I built an LDA model on the posts of all 100,000 users. I first removed common English words and converted tags from each post to bag-of-words format. We trained an online LDA model, which is expected to converge quickly given relatively stationary topics (in this case, pro-ED related topics) without much drift over time. Specifically, I updated the LDA model for every chunk of 1 million posts for all 26M posts to obtain 100 topics.

(2) MIS Annotation. By reviewing the top 50 keywords of each topic provided by the LDA model, I then obtained human annotations of MIS on every topic. I defined MIS to span three levels — low (1), medium (2), or high (3).

The annotators included four researchers. Two annotators were trained clinical psychologists with specific expertise in eating disorders and experience interacting with eating disorder in-patients, and the other two had considerable social computing research experience. The researchers created a set of rules to annotate each topic. The raters first manually browsed pro-ED posts on Instagram by searching over all of the nine seed tags used for crawling pro-ED data[4] — so the raters could familiarize themselves with MIS manifested in pro-ED posts. Then, the raters collated a set of rules which were used for the annotation task. I call this the MIS scoring or rating system:

- High MIS (score of 3): included extreme weight control behaviors and "thinspiration" (*e.g.*, , "purge," "thinspo," "starve," "donteat"), self-injurious behavior (*e.g.*, , "cutting," "blades," "slit"), and suicidal ideation (*e.g.*, , "killme," "suicidal").
- Medium MIS (score of 2): included fat talk, self-deprecation, emotional instability, cognitive impairment, social isolation, and discussing eating disorders (*e.g.*, , "uglyandfat," "selfhate," "broken," "anorexia," "eatingdisorder," "mia"). In addition, manifestations of mental disorders but without any revealed vulnerability (*e.g.*, , "depression," "bpd," "anxiety") were also scored as medium MIS.
- Low MIS (score of 1): included tags not related to eating disorders or mental health (*e.g.*, , "nyc," "iphone," "biking," "cats," "fashion," "selfie").

Following annotation task, interrater reliability metric Fleiss' κ was very high (.91); further, between the novice (non-clinician) and the expert (clinician) ratings, there was high agreement. Discrepancies were resolved through mutual discussion. 5 topics out of the 100 given by the LDA model were annotated as high (3), 6 as medium (2) and 89 as low (1).

(3) Computing Users' MIS Rating. Given any post, I then use the topic probability distribution generated by the LDA model as weights and combine them with the annotated MIS levels of the topics, to obtain a weighted average MIS rating.

(4) Monthly MIS Rating Inference. To infer a user's MIS in a time slot (month), I obtained the discretized (rounded) mean MIS rating over all posts posted by the user on Instagram during the slot.

Evaluation of MIS Ratings. How effective is the LDA topic modeling approach combined with novice/expert annotations in identifying MIS in a user's content? To answer this question, I compared the algorithm-derived MIS ratings with MIS ratings in the same sample obtained from four human raters – two novices (non-clinicians) and two experts (clinicians). I first randomly sampled a set of 150 posts with equal numbers for the three MIS ratings 1 (low), 2 (medium), and 3 (high). For the sake of consistency, the same four researchers who labeled the LDA topics rated this sample of 150 posts for the three MIS ratings. The raters had high agreement (Fleiss' $\kappa = .86$) and resolved discrepancies mutually via discussion. Raters found 68 posts rated as low, 9 as medium, and 73 posts rated as high MIS.

Next, I compared the agreed upon set of human annotations with algorithm-derived MIS ratings on the 150 post sample. This gave high accuracy, precision, recall and F1 scores (mean accuracy $>$71%, mean precision $> 68\%$, and mean recall $> 70\%$). These scores were particularly high for MIS ratings 1 and 3; for these two classes respectively, precision was over 94% and recall above 64%. However, I observed that the human annotations and algorithm differ significantly in the case of MIS rating 2. A closer look revealed that, while recall is still very high for this class (66%), the low precision (37%) is responsible for the low F1 score. Manual inspection of such misclassified posts reveals that the disagreement arises due to the inherent ambiguity in the posts of actual MIS rating 2. I note that the MIS ratings 1 and 3 are perhaps the most important and distinctively defined classes of practical importance and constitute over 93% of the content (see Table 3.2).

Prediction of MIS. As with other types of regression, for the regularized multinomial logit method, there is no need for predictor variables to be statistically independent from each other (unlike, for example, in a Naïve Bayes classifier or an ordinary least squares model). Regularization helps us control for collinearity (*i.e.*, excessive correlation between MIS ratings that are temporally close) and sparsity (*i.e.*, users may not post in certain time slots, thus no MIS can be measured) in the data. Further regularization allows us to incorporate smoothness in my model — it is likely MIS changes smoothly across consecutive time slots for most users' content. In this case, I used the model implementation in the Python package scikit-learn. Below are the components of the regression model:

Response Variable. The MIS ratings of users' content (1=low; 2=medium; 3=high) at time slot t. Here, the time slot is taken as a month.

Predictor Variables. I define a sliding window of size w, and consider the monthly MIS rating of users' content over all time slots between $t - w$ to $t - 1$ as w predictor variables of the model.

The class sizes (response variables) in the dataset were unbalanced (ref. Table 3.2), hence I employed bagging and boosting to improve performance of the model. I used 90% of the data as training data (90K users); the remaining 10% of users was set aside as the held-out test set on which I report the prediction results. Specifically, I first generated B bootstrap samples of the training data using random sampling with replacement – in these samples, I selected users so that all three classes are balanced. I then trained my regularized multinomial logistic regression model on each of these bootstrap samples. Following training, in the boosting phase, I iteratively learned weak regression models using the Adaboost algorithm [107]. That is, I took the weighted sum of the coefficients of the model – this gave us a robust model that was not biased by class imbalance. This ensemble regression model given by Adaboost was finally applied on the 10% heldout set to report performance on predicting MIS rating.

3.2.3 Results

Description of High MIS Topics. I begin with a discussion of 2 of the high MIS topics derived from combining the LDA model and human annotations. Extended discussion of these topics is located in the paper. Recall, out of the 100 topics generated by the LDA model, five were rated by the clinicians and researchers to have high MIS. However there are systematic differences in these topics despite the presence of these tags.

Qualitative differences are noticeable across the high MIS topics. The first topic, topic 32, discusses expressions of a variety of different mental health disorders (*e.g.*, , "bpd," "ptsd," "social anxiety," "schizophrenic," "ocd," "panic"). The topic also contains mentions of stress and anxiety, some of the known concomitants of mental illnesses disorders like "addictions" and "insomnia," as well as some of the probable causes behind them (*e.g.*, , "abuse"). Additionally, the topic revolves around calls for support, desire or need for recovery, and mentions of treatment efforts ("medicine"). Note that while eating disorders as a distinct illness in the DSM-5 [8], many other forms of mental health disorders have high comorbidities — this explains the nature of content of this topic.

Next, topic 52 centers around commonly adopted harmful/dangerous habits of pro-ED lifestyles. For instance, I observed mentions of "binge," "purge," "starving," "blithe," "fast," "hungry". I also observe manifestation of self-loathing, *e.g.*, "ugly," "gross," "fat." These tags capture attitudes and thoughts that reinforce these lifestyles ("donteat") and seem to share motivations towards continuing to do so ("anatips.") Literature identifies such manifestation of self-identification with pro-ED lifestyles to be an indirect form of self-injurious behavior.

Dynamics of MIS. Next, I analyzed levels of MIS in the data as well as their change over time. Table 3.2 gives the proportion of posts with low (1), medium (2), and high (3) MIS. The majority of posts are low MIS (88.8%) and the rest span medium and high

MIS. While in general users who share pro-ED content are expected to show markers of high MIS through the sharing of content around physically or emotionally dangerous acts, I found that a notable fraction of these users also use the Instagram platform for sharing non-mental illness related content, as indicated by the low level of MIS in majority of the posts.

Table 3.2: Proportion of posts with different MIS rating.

MIS Rating	Post Count	Percentage
Low (1)	22,913,989	87.41%
Medium (2)	1,990,031	7.59%
High (3)	1,311,288	5.00%

However a deeper examination of the way the three levels of MIS change over time reveals that, while small in the proportion of posts, alarmingly, the *relative fraction* of users who share pro-ED content and show high MIS has been on the rise. Figure 3.1 presents the proportion of "active" users in the data with low, medium and high MIS per month (55 months in all: Oct 2010 through Mar 2015). The figure indicates that from month 18 (Mar 2012) to month 48 (Oct 2014), both medium and high MIS rating user proportions show a steep increase, whereas low MIS rating shows decline during the same period. In fact at its peak, as many as 10% of users in my data are inferred to express high MIS rating. A Wilcoxon rank sum test shows statistically significant differences between the fraction of users with high MIS rating in month 48 and in month 18 ($z = 4.19; p < 10^{-5}$).

I also compute a rate of change metric to capture the trend of the proportion of medium and high MIS rating users over time, known as *momentum* — a measure that has been used in social media analysis to observe changes in time series data. It is given as the mean ratio of the difference between medium/high MIS rating users at time t and that at $t - 1$ to the medium/high MIS rating users at time $t - 1$. The momentum gives a positive value (13%/year), which indicates that overall proportion of pro-ED Instagram users with medium/high MIS rating in their content has been monotonically increasing over time.

MIS Rating Prediction.

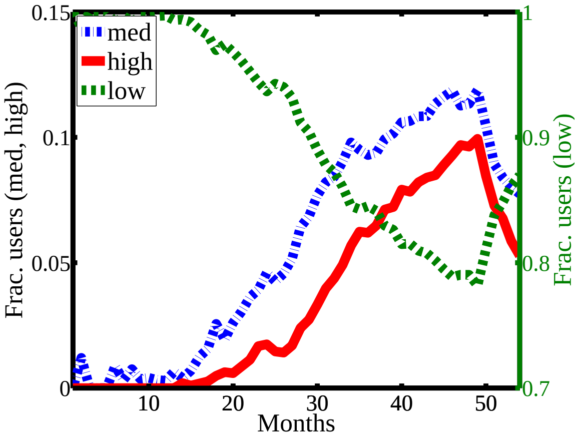


Figure 3.1: Fraction of users with low (1), medium (2), and high (3) MIS rating over time (in months). Here the fraction of users with a particular level of MIS in a certain month is given as the ratio between the number of "active" users with the particular MIS rating to the total number of "active" users during that month — "active" being defined as any user with at least one post during the month. Thus these fractions allow normalization and comparison of MIS ratings across months.

Fitting Models. In this final subsection, I report the measures of fitting the MIS rating prediction model on the bootstrap samples based on the regularized multinomial logistic regression framework. One of the aspects of this investigation was determining the appropriate sliding window size w for which I obtain the best model fit on the bootstrap samples as well as the one for which the prediction performance is optimal.

In Table 3.3, I present results on fitting a number of models to the bootstrap training samples with different values of the sliding window w. Compared to the null model, all versions of the regularized multinomial logit models (with different values of w) provide considerable explanatory power with significant improvements in deviances. The difference between the deviance of the null model and the deviances of the models approximately follows a χ^2 distribution, with degrees of freedom equal to the number of additional variables in the more comprehensive model. As an example, comparing the deviance of $w = 5$ model with that of the null model, I saw that the information provided by the MIS ratings over the past five months has significant explanatory power:

Table 3.3: Summary metrics of fitting the regularized multinomial logit model to the training data. Ten different models with different sliding window sizes w are reported, along with the null (intercept only) model.

Model	Deviance	χ^2	p-value
Null model	915.523		
$w = 1$	633.028	282.495	$p < 10^{-4}$
$w = 2$	533.195	382.328	$p < 10^{-7}$
$w = 3$	489.055	426.468	$p < 10^{-8}$
$w = 5$	320.625	594.898	$p < 10^{-8}$
$w = 7$	279.801	635.722	$p < 10^{-10}$
$w = 10$	297.81	617.713	$p < 10^{-10}$
$w = 13$	319.259	596.264	$p < 10^{-10}$
$w = 15$	327.505	588.018	$p < 10^{-8}$
$w = 17$	417.77	497.753	$p < 10^{-7}$
$w = 20$	643.989	271.534	$p < 10^{-5}$

Table 3.4: Predicting low, medium, high MIS ratings of users in heldout test set using the regularized multinomial logit model.

Model	Accuracy (%)	Precision	Recall	F1
$w = 1$	59.89	0.647	0.655	0.651
$w = 2$	62.76	0.674	0.706	0.690
$w = 3$	67.98	0.676	0.724	0.699
$w = 5$	69.32	0.723	0.797	0.758
$w = 7$	81.89	0.817	0.804	0.810
$w = 10$	73.99	0.808	0.790	0.799
$w = 13$	72.26	0.808	0.754	0.780
$w = 15$	67.42	0.774	0.728	0.751
$w = 17$	66.32	0.684	0.619	0.650
$w = 20$	61.50	0.638	0.617	0.627

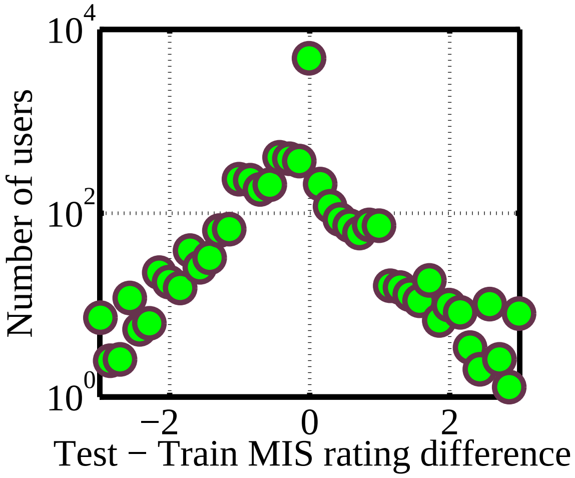
$\chi^2(5, N = 90K) = 915.523 - 320.625 = 594.898, p < 10^{-8}$. This comparison with the null model is statistically significant after Bonferroni correction for multiple testing ($\alpha = 0.005$ since I considered 10 different models corresponding to the 10 values of sliding window w). The best model fit (in terms of lowest deviance) is given by the $w = 7$ model ($\chi^2(7, N = 90K) = 915.523 - 279.801 = 635.722, p < 10^{-10}$), with best fits (low deviance) for models where w is closer to $w = 7$, and decreasing as w goes lower or higher.

Prediction on Heldout Data. In the next part of the MIS prediction analysis, I present the performance of my approach when tested on the heldout test dataset of 10K users. I summarize performance metrics of this multi-class classification via Table 3.4, reporting average accuracy, precision, recall and F1 measures across the 10 different sliding window choices of the logit model. As also observed in the results on model fit, the highest accuracy (82%) and F1 (82%) are given by model $w = 7$. Specifically, I obtained high accuracy and F1 for the 1 (low) and 3 (high) MIS ratings (accuracy: 89% and 87% respectively; and F1: 87% and 84% respectively). The performance is found to be relatively lower for MIS rating 2 (medium) (accuracy 70% and F1 71%), which I attributed to the ambiguity and subjectivity in such content. However all three classes perform above baseline models (majority vote models; since I used bagging and boosting, I can test against a baseline of 33% — equally split three classes), showing that past MIS rating of users measured from their Instagram postings, is indeed able to forecast with high confidence, future MIS.

Predicting low/medium MIS to high MIS Transitions. Next, I wanted to investigate if the best performing model ($w = 7$) can predict increase in MIS in the future in cases where past MIS may be relatively lower. To do so, I obtained the distribution of correctly classified users in the test set with respect to the difference between their predicted/test MIS rating and the mean MIS rating in historical/training data. Positive difference between predicted/test data MIS rating and that in historical/training data would indicate that the model predicted correctly the increase in MIS for the user's content.

As reported earlier (Table 3.4), in the set of 10K users in the test dataset, 8,189 users are correctly classified—I can infer MIS rating in their content in the eighth month correctly, based on ratings in the seven months before. I find that in the distribution for the bulk of these correctly classified users predicted future MIS rating is approximately the same as past (notice the spike near zero difference). However, there is a notable fraction of the users ($\sim 11\%$) for whom the model correctly predicted an increase in MIS rating despite comparatively lower MIS rating values in the past (notice the right side of the distribution). Similarly, for $\sim 17\%$ I inferred accurately a decrease in MIS in the future even though their MIS in the past was higher.

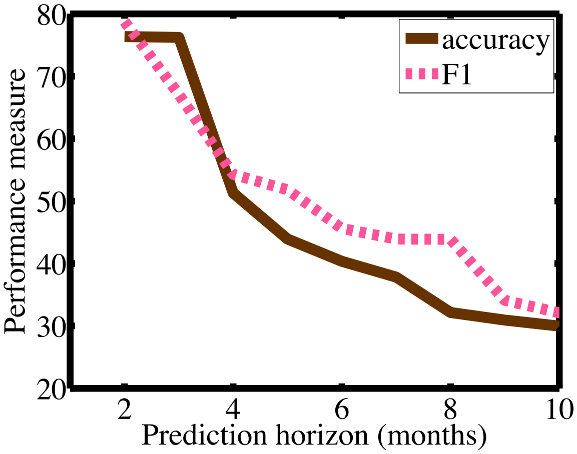
Figure 3.2: Distribution of differences between MIS ratings in test set and those in the training set for all correct predictions given by the $w = 7$ model (w is sliding window size). Positive difference (right side of the distribution) implies the model was able to predict that in the eighth month, the MIS rating measured from content of a user was *higher* than the mean MIS rating over the seven months preceding it.



Prediction Horizon. Finally, how far out into the future can we predict a user's MIS rating? In other words, so far, I have shown that a model trained on MIS rating data between time $t - w$ and $t - 1$ (w is the sliding window size) can predict MIS rating at time t. Using the same model, could I predict MIS rating at time $t + h$, where $h > 1$? Here, I define h as the "prediction horizon" and Figure 3.2.3 presents the accuracy and F1 measures of applying the $w = 7$ model (the best performing model from the previous paragraph) to predict MIS rating 2 months to 10 months out into the future ($h = 2, 3, ..., 10$).

Observations from Figure 3.2.3 indicate that, not surprisingly, performance as measured by accuracy and F1 steadily deteriorates as I attempt to predict MIS rating further out into the future. The best models, in other words, are those where predictions are attempted to be made closer in time to the last observed time of MIS rating in the model (the $h = 2$ model gives an accuracy of 78% and F1 of 80%). However, note that since bagging and boosting allows us to compare model performance at the baseline level (balanced class sizes, 33%), the predictions of MIS rating continue to be better than baseline through eight months into the future.

Figure 3.3: Change in accuracy and F1 score of the $w = 7$ MIS prediction model (w is sliding window size) for prediction horizons h between 2 months and 10 months. Here prediction horizon h implies the model is predicting MIS rating at month $t + h$ ($h > 1$) using MIS rating predictors from months $t - w$ through $t - 1$.



3.3 Recovery Amid Pro-Anorexia: Survival Analysis for Pro-Anorexia

In addition to having pro-ED communities across various social platforms, there also exist thriving "recovery" communities on social media sites. A recovery community is a group of users that discuss the health challenges of mental disorders, promote treatment options, and serve as support for users who are recovering from mental disorders. Studying both communities, the tensions that develop between them, and the users that frequent them is of interest to researchers drawn to improving anorexia recovery outcomes.

Our research question involves studying attempts at recovery amid the "anti-recovery" sentiment that pro-ED communities foster. The central research question, therefore, is to examine the role and efficacy of Tumblr as a platform for sustained recovery from anorexia after pro-ED participation.

This project makes the following three contributions

- Using a hybrid methodology that integrates text processing and human annotations, I identified users who shared anorexia-related content as well as showed signs of recovery in their social media posts.

- I developed a robust statistical model based on survival analysis to estimate recovery over time in the user population.

- Using survival analysis over a large dataset of over 10,000 Tumblr users and over 68 million posts, I identified a number of measures that are likely linked to anorexia recovery – body image concerns, behavior, cognition, and affectbased on the cognitive behavioral theory of anorexia. I observe that (a) the estimated time to begin recovery for half of the population is 45 months, and (b) over a six year trajectory, only 56% of the study body is estimated to show signs of recovery on Tumblr.

These results indicate that anorexia recovery on Tumblr is protracted; the data call into question the long-standing belief that health communities are universally beneficial at encouraging and sustaining positive health outcomes. An important implication of the work is that a challenging psychosocial disorder like anorexia may require significantly different approaches toward encouraging recovery, given the large and vocal presence of a pro-disorder community on social media. This project was published at CHI 2016 [31], and I summarize the data, methods, and results below.

3.3.1 Data

My investigation uses data from Tumblr, a microblogging service owned by Yahoo!, where users post text and multimedia content to a short-form blog.

Phase I: Collecting Pro-Anorexia Data. I first manually examined Tumblr blogs mentioning common eating disorders and their associated anorexia symptomatology tags. Based on a snowball sampling approach, I obtained an initial list of 28 tags. I examined the co-occurrence of other tags with these seed tags and applied filtering of generic tags (*e.g.*, , "fat.") This process expanded the tag list to 304 tags. Examples of these tags include "proana," "anorexia," "thighgap," "thinspiration,""thinspo." I used the Tumblr API to search for posts containing any of these tags. In the process, I collected 55,334 *public* English language posts generated by 18,923 unique users.

Table 3.5: Summary Statistics describing characteristics of the “Recovery” and the “Non-recovery” cohorts.

	All	**Recovery**	**Non-Recovery**
Total Users	13,317	2,353	10,964
Total Posts	68,380,375	25,710,069	42,670,306
Median Posts / User	1,701.50	4,515	1,276

Phase II: Obtaining Historical Data of Users. In the second phase, I started with the set of 18,923 unique users and retained only those who were still active in Tumblr (since the posts in the first phase were spread over 2008 to 2013, some users were no longer on the platform during the second phase). This left 13,317 active users. For each user, I crawled their entire Tumblr history along with their profile information. These include users’ activity information (total posts since account creation), the total likes on posts, whether post likes were set to be visible to any other Tumblr user, whether they set ‘ask’ (questions) to *yes* and ‘ask’ (questions) by anonymous users to *yes*, and whether their profile is set to share NSFW (“not suitable for work”) content. I also gathered post metadata: user id, tags, timestamp, the number of notes (or comments) it received, the number of likes, and users’ activity information from their profile (total posts since account creation).

Phase III: Identifying the Recovery Cohort. The final task involved identifying a candidate set of users who are likely to be recovering from anorexia. I leveraged findings from prior work on expression of recovery tendencies on social media—this literature has identified that social signal of posting to certain specific tags can be a strong indication that a user desires to recover [14, 50, 49]. I obtained a random sample of 1000 posts containing the regular expression “*recover*.” Next, two researchers manually went through the tag list to identify and compile a set of tags co-occurring with the recovery tags for these posts. This was to find only those co-occurring tags which had cues associated with anorexia recovery, *e.g.*, “fighting” and “edsoldier.” Table 3.6 lists example tags in the sample. Similarly, I compiled a set of tags with the regular expression “*relaps*” which has been found to be indicative of intent toward anorexia relapse [49]. Taken together, any user who used any of the recovery tags in at least five distinct posts but did not use any of the relapse tags were considered to belong to the “recovery cohort.” I refer to the rest of the users as the “non-recovery cohort.”

Table 3.6: Sample tags identifying the recovery cohort.

eating disorder recovery	anarecovery	chooserecovery
healthy recovery	pro recovery blog	reasons to recover
recovery fighter	recovery food	recovery intake
recovery record	recovery tips	recoveryisworthit
recoverywarriors	road to recovery	self recovery

Finally, I evaluated whether the users included in the recovery cohort shared Tumblr content containing signs of recovery. Two researchers familiar with Tumblr and pro-anorexia content online and a clinical psychologist with expertise in eating disorders and self-harm independently evaluated the correctness of the above method. They manually read through the posts of a random sample of 150 users from the recovery cohort, marking (yes/no) whether each user indeed posted content that signaled their desire to recover. Since a single post may not be indicative of the user’s inclination to recover, the researchers used a set of three posts per user in the set of 150 users. The raters had high agreement (Cohen’s κ=.83) and found that 81% of the users were correctly identified by my method to be in recovery.

Table 3.5 gives summary statistics of the data in the recovery and non-recovery cohorts. There were 2,353 users in the recovery cohort (25,710,069 posts) while 10,964 in non-recovery (42,670,306 posts). The posts across both cohorts were shared between Feb 20, 2007 and Aug 4, 2014.

3.3.2 Methods

Formally, survival analysis is a collection of statistical procedures for analyzing longitudinal data where the outcome of interest is *time until an event occurs* [108]. These techniques are well-suited for scenarios where subjects encounter the event of interest at varying times, cases where they might not even experience the event during the entire observation period, or subjects are lost during the study [109, 110].

Statistical Technique: To determine the rate of recovery, I used the Kaplan-Meir estimator for survival analysis [111]. This method provides an estimation of the survival function when the underlying data is censored. It estimates the probability of not having the recovery event (*i.e.*, to survive or be in the anorexia state) as a function of time. This is the same as finding the chronological sequence distribution of survival probabilities. The corresponding probability plot is called the survival curve while the tabular representation is referred to as the life table. The median survival time is the time at which one half of the entire cohort recovers.

To find the effect of various factors on the time to recovery, I used the Cox proportional hazards regression model [112]. It is a survival analysis regression method that describes the relationship between the event of interest (in this case, the “recovery event”) and the factors that affect the time to that event occurrence. It allows us to estimate the change in the survival probabilities with change in these potential factors. It especially fits the research because Cox modeling does not assume the survival times to follow any particular statistical distribution, unlike most other statistical models.

Measures. I offer a number of content and participation measures for the Cox proportional hazards regression model that predict anorexia recovery. I base these measures on observations in prior literature as well as attributes of the Tumblr platform. These measures are derived from psychological studies of language use [113] that indicate how different linguistic constructs capture diagnostic information about a wide range of psychological phenomena, ranging from psychiatric disorders, suicidal ideation to responses to a trauma-related upheaval [114, 115].

To characterize content and linguistic constructs of Tumblr content in a systematic and semantically interpretable way, I used the popular psycholinguistic lexicon LIWC (http://www.liwc.net) [116]. I frame the choice of content and participation measures in the light of the cognitive behavioral theory of anorexia nervosa [117].

I discuss two of four measures here. An extended discussion can be found in the paper.

1. *Body Image Concerns:* To capture attributes relating to idealized perceptions of body image and the desire for thinness among anorexia sufferers [117, 118], I included measures of the volume of *ingestion*, *body*, and *health* words in the content shared by users — these words and their corresponding categories were obtained from the LIWC dictionary.

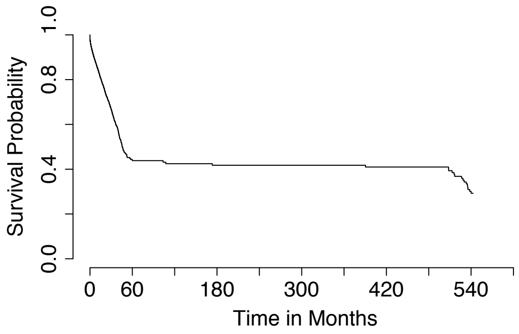
2. *Affect:* Finally, I measured affect to characterize emotional expression in Tumblr content. I represent affect as normalized positive affect (PA), computed as the ratio of LIWC words in the positive emotion category, to those in the *negative emotion*, *anger*, *anxiety*, *sadness* categories together.

3.3.3 Results

Figure 3.4 graphs the cumulative probability of experiencing recovery as a function of time. Using the Kaplan Meir estimator for survival analysis, the median time to recovery is 45.6 months. In other words, after 45.6 months, 50% of the user population have not recovered. Probabilities of recovery are also listed for periodic intervals up to 6 years in the life table in Table 3.7. I see that the probability of recovery 2 years after being on Tumblr is only 16%, while at year 6 it is at 56%. Table 3.7 underscores that the time course of recovery over the first several years is protracted (*i.e.*, , significant lengthening of the time to show signs of recovery in Tumblr content). In fact, as indicated in the survival curve, the likelihood of recovery beyond the 5 year mark ($\sim$60 months) is very low – the graph almost shows a flat trend.

Survival curves can estimate the likelihood that a Tumblr user who has not recovered at a specific time point will remain anorexic for an additional length of time (calculated by dividing the probability of not recovering at Time t_j by the probability of the same

Figure 3.4: Survival curve showing likelihood of experience of the "recovery" event in the Tumblr data sample of pro-recovery users.



at Time t_i, where $j > i$). For example, the probability that a user who did not discuss recovery on Tumblr by Year 2 would remain anorexic for another 2 years is 0.28/0.52 = 53.8%. If she does not recover in 3 years, the probability of remaining anorexic for another 3 years is 0.39/0.56 = 69.6%. As the time of not recovering increases, the likelihood of ever experiencing recovery decreases. This finding aligns with prior work in the anorexia literature where symptomatic recovery patterns had been examined [119].

Table 3.7: Cumulative probability of remaining in the non-recovery cohort. Median time to recovery is 45.6 months (shown as the shaded row). This is the time when 50% of the users are still expected to not gave recovered.

Time (Years)	Time (Months)	Survival Prob.	Cumulative Prob.	Std. Error
0	0.00	1.00	0.00	0.0003
1	12.00	0.84	0.16	0.0047
2	24.00	0.72	0.28	0.0083
3	36.03	0.61	0.39	0.0141
3.75	45.22	0.51	0.49	0.0235
4	48.01	0.48	0.52	0.0277
5	60.01	0.44	0.56	0.0364
6	72.44	0.44	0.56	0.0382

Recovery Models and Goodness of Fit. With the four different categories of measures identified in the Measures section, I reported on five models. The first four models correspond to the four measure categories, and the fifth includes all measures from all categories. I refer to them as: `BodyImage`, `Behavior`, `Cognition`, `Affect`, and `Full` models in the rest of this paper.

First, I evaluated the goodness of fits of all five of the Cox regression models with *deviance*. Deviance is a measure of the lack of fit to data—lower values are better. Compared to the Null model, the models provide considerable explanatory power with significant improvements in deviances. The difference between the deviance of the Null model and the deviances of the other models approximately follows a χ^2 distribution with degrees of freedom (df) equal to the number of additional variables in the more comprehensive model. As an example, comparing the deviance of the Behavior model with that of the Null model, the information provided by the corresponding variables has significant explanatory power: $\chi^2(10, N = 13,317) = 5294.57 - 739.55 = 4,555.02, p < 10^{-6}$. This comparison with the Null model is statistically significant after the Bonferroni correction for multiple testing ($\alpha = \frac{0.05}{5}$ as I consider five models). I observed similar deviance results for the BodyImage, Cognition, Affect and Full models, with the last model possessing the best fit and highest explanatory power ($\chi^2(26, N = 13,317) = 5294.57 - 214.15 = 5080.42, p < 10^{-10}$).

Table 3.8: Summary of different model fits. Null is the intercept-only model. All comparisons with the Null models are statistically significant after Bonferroni correction for multiple testing ($\alpha = \frac{0.05}{5}$).

Model	Deviance	df	χ^2	p-value
Null	5294.57	0		
BodyImage	328.63	3	4,965.94	$< 10^{-10}$
Behavior	739.55	10	4,555.02	$< 10^{-6}$
Cognition	424.92	12	4,869.65	$< 10^{-9}$
Affect	682.86	1	4611.71	$< 10^{-6}$
Full	214.15	26	5080.42	$< 10^{-10}$

Table 3.9 shows the overall Cox model fit by listing the likelihood ratio, Wald and chi-square statistics, and the concordance measure. The Full model showed the lowest deviance (refer Table 3.8), so I report expanded statistics on this model. The tests that generate these statistics are equivalent to the omnibus null hypothesis that all β coefficients are zero. Because the tests shown in Table 3.9 are statistically significant (Wald statistic $z =$

$567.4, p < 10^{-15}$) I reject the null hypothesis, indicating that the variables I consider in the `Full` model contribute towards significant explanatory power of estimating the likelihood of recovery.

Table 3.9: Summary of fit of the `Full` Cox regression model that estimates likelihood of experience of the recovery event based on the content and activity measures.

		df	p-value
Likelihood ratio test	511.6	26	$< 10^{-15}$
Wald test	567.4	26	$< 10^{-15}$
Score (logrank) test	526.6	26	$< 10^{-15}$
Concordance	0.658	(Std. Err. = 0.007)	
R^2	0.046	(max possible= 0.974)	

I also note that concordance is a generalization of the area under the receiver operating characteristic (ROC) curve and measures how well a model discriminates between different responses. Specifically, it is the fraction of the pairs of observations in the data, where the observation with the higher survival time has the higher probability of survival predicted by the model. A concordance of greater than 0.5 generally indicates a good prediction ability (the value of 0.5 denotes no predictive ability). The concordance for the *Full* model is very high (65.8%), and is higher by 14-26% compared to the other models.

What Predicts Recovery? In this section, I discuss how two of the four different classes—body image concerns and affect, relate to predictions of recovery.

Body Image. *Reduced body image concerns and increased focus on health and ingestion positively impact the likelihood of anorexia recovery.*

Increased use of *health* and *ingestion* words in Tumblr posts is linked to greater likelihood of recovery. The strongest predictor for recovery in the cohort is discussions about health. Users who mention *health* words increase their monthly hazard of recovery by a factor of 4.1×10^5 ($\beta = 12.9, p < 10^{-3}$). These users express awareness about the severity of anorexia, low body weight, and food restriction as a health risk. Recovering users thus largely use Tumblr as a way to publicly demonstrate their altered perceptions about self-starvation and positive attitudes towards resuming a healthy lifestyle in addition to exhibiting evidence of self-acceptance of their physical appearance:

> "Your weight is only *healthy* when you are." (recovered)

> "A while ago, a girl asked her friend why she smoked. When she answered 'i can become anorexic or keep smoking' and then laughed I just wanted to scream and make her realise that ANOREXIA IS REALLY THAT BITCH." (recovered)

Other users acknowledge the danger of anorexia and compared the danger of such illnesses to other conditions, like depression or diabetes.

> "I do not encourage any type of depression, eating disorder, bipolar disorder, this is just a blog to represent the mess in my brain. I AM NOT PRO-ANA OR PRO-MIA because this is the stupidest thing I ever heard, eating disorders are ILLNESSES, as well as Cancer, you wouldn't be PRO-CANCER would you? This is the same." (recovered)

Improved recovery was also found alongside discussions of ingesting or eating food. Words in this category increased the monthly hazard rate by a factor of 3.7×10^3 ($\beta = 8.2, p = 10^{-3}$). Posts with these words often show a preoccupation with food, nourishment, and eating. It may be surprising that preoccupation with food shows inclination to recovery. Often, however, individuals with anorexia are more focused on their bodies and a focus shift from their physical appearance to food shows movement into recovery [120].

> "Meal 1 at three pm Taco bell bean and cheese caloopa Meal 2 at midnight pasta Additional *foods* Half a chicken quesidia , chips and cheese , donuts Today i finally weighed in at 105 pounds!!! Only ten more to go till my goal of 115 . So proud of myself for coming this far ." (recovered)

> Yes, you should really *feed* yourself. No more of this sitting around for an hour with your cup of *coffee* trying to figure what you want. *Eating* is essential. There is never a time when skipping a *meal* is the right answer. Go *eat*, because you have a life to live. (recovered)

However, posts with *body* words that capture general “body talk” negatively correlate with recovery in the data ($\beta = -19.9, \exp(\beta) = 2.08 \times 10^{-9}(p < 10^{-3})$). These posts offer dietary strategies, fasting regimes, sharing ‘safe’ food or low-calorie foods that users have tried, ways to purge, diet pills, or diuretic abuse. Discussions of specific body parts or how the body looks can also be found here.

> I am on the scale and all what I can think of is how *fat* and ridiculous I am. So I don’t eat, and if I do, I’d purge. But I don’t really care [...] I want to be *skinny* and I don’t care if I die doing this. But at least I will be *skinny* in heaven if not on earth (did not recover)

> You do it because you think your *body* is cute for when your boyfriend picks you up, he feels the *thigh* gap not the *flesh*. But I do it because the voices in my *head* tell me I dont have any other choice. (did not recover)

Affect. *Increased positive affect and attitude increases the likelihood of recovery.*

Finally, likelihood of recovery is higher in users whose content exhibits a more pronounced hedonic focus on positive emotions and an objective outlook towards life, as indicated by the measure of normalized positive affect (PA) (hazard ratio $\exp(\beta)$ increases by a factor of 1.23, with $\beta = 2.5, p < .02$).

> I *love* a life full of eating out with my *loved* ones, sleeping in if that’s what I want, getting married and having a family and I can’t do these with ana. Let it go away and live your life to the *fullest* and *happiest*. (recovered)

3.4 Implications

In this section, I summarize some of the impacts of these two works on novel computational methods to understand pro-ED and the related pro-ana communities on Tumblr and Instagram. Taken together, these two methods provide new insights into the behaviors in these communities over time. I discuss the impacts these methods have on understanding complex mental health behaviors in online communities.

To begin, in "MIS," I demonstrated that, with a human-algorithmic hybrid approach, I can examine risk markers for mental illness severity (MIS) on Instagram over a period of 5 years. A significant finding of this work is the steady increase in the proportion of medium and high MIS users who share pro-ED content over time on Instagram. While pro-ED users would be expected to have a higher manifestation of MIS, I show that there is a relative increase in medium and high MIS expressed in content over time. Although causal relationships cannot be inferred without a more systematic investigation, examining why high MIS occurs opens up several avenues for future research. Are individual users showing markers of riskier behaviors because of contagion effects of the content they consume on Instagram? Are they leaving the community? Could increases in MIS be indicative of a shift in social norms in the community? Could high MIS tag usage reflect an adversarial response to Instagram's content moderation policies in 2012?

This conjecture springs from the observation that communities often adopt unprecedented and competing avenues to combat content regulation and moderation. For example, citizens of authoritarian regimes avoid censorship by embracing different forms of linguistic variation [121]. Similar adoption of unorthodox practices engender deviant communities engaging in cyberbullying and online harassment [122] as well as those involved in socially unacceptable or damaging activities (human trafficking, drug abuse, violence, or organized crime). MIS provides a new set of insights into these changes within the community.

Another impact of this work for social media research is exploring the nuances in how online health communities provide both support and resistance to the experiences of disorders, like the recovery experiences of anorexia. Through a survival analysis technique, I show that I can examine the linguistic and behavioral cues that may indicated the recovery from anorexia. Despite indexing for cues related to recovery through the Cognitive Behavior Theory of Anorexia Recovery [117], anorexia recovery is difficult, protracted, and often times accompanied by significant relapses. In "Recovery Amid Pro-Anorexia," I saw similar difficulties in maintaining recovery in the sample – 56% of the cohort remained in recovery through the study.

What could be some of the differences that leads to more positive or negative outcomes? The ecosystem of an otherwise general purpose social media Tumblr and the presence of the anorexia community brings to light the unique situation that these individuals face. In other illness support communities, few people, if any, come to promote the spread or maintenance of a condition like cancer, diabetes, or anxiety. In the context of anorexia, however, pro-anorexia communities exist and coexist in the same spaces as those interested in recovery. A tag as narrow as "anorexia" can contain posts ranging from graphic descriptions of self-harm to extreme diet ideas to positive affirmations and support.

This is not to say, however, that because this population fared no better than clinical results and there are barriers to perfect adoption that social media cannot assist in anorexia recovery. In fact, there are several contexts in which platforms like Tumblr may be facilitating recovery. One way is that the community provides emotional support in times of need and isolation. Social media platforms may be a first step to recovery by promoting a sense of togetherness and a place to openly share experiences and emotions with others. Tumblr also provides an emotional "safety valve." [12] Rather than taking more dangerous or drastic actions (*e.g.*, self-injurious behavior), those suffering can talk through their problems and avoid harmful behavior; social media like Tumblr offer a promising way to engage in disinhibiting and self-disclosing discourse.

CHAPTER 4
NORMS AND DEVIANT BEHAVIOR ONLINE

As discussed in the last chapter, pro-ED communities are considered deviant across a wide variety of viewpoints and levels of analysis. By using the lens of deviance and normative behavior to understand pro-ED communities, we can explore their motives, culture, and interactional patterns of these socially undesirable behaviors [20].

The first part of this chapter includes background literature for normative behavior, deviance, and these concepts' relationships to online communities. Then, I discuss two projects that quantitatively examine norms and deviant online in pro-ED communities.

The first project, "#thyghgapp: Instagram Content Moderation and Lexical Variation in Pro-Eating Disorder Communities," discusses how pro-ED communities subverted content warnings and bans on Instagram, and the implications for these communities sharing these behaviors. After content moderation, lexical variants emerged for all 17 pro-ED tags that underwent initial moderation in 2012. In fact, increasingly complex lexical variants have emerged over time. Communities that use lexical variants show increased participation and support of proED (15-30%). Finally, the tags associated with content on these variants express more toxic, self-harm, and vulnerable content. Despite Instagrams moderation strategies, pro-ED communities are active and thriving.

The second project, "Norms Matter: Contrasting Social Support Around Behavior Change in Online Weight Loss Communities," uses computational linguistic techniques and machine learning to understand norms of support in online weight loss communities on Reddit, I used computational linguistic methods to juxtapose similarities and differences in two Reddit weight loss communities, r/proED and r/loseit. I employ language modeling and find that word use in both communities is largely similar. Then, by building a word embedding model, I contrasted the context of word use and find differences that imply different behavior change goals in these OHCs.

Taken together, these projects use observational, computational analysis of the pro-ED community to understand the nuances in behaviors around disordered eating behaviors.

4.1 Related Work

4.1.1 Normative Behavior

Broadly conceived, social norms are understandings that guide behaviors that "serve as prevailing codes of conduct that either prescribe or pro-scribe behaviors that members of a group can enact." [123] Hogg and Reid define norms as "regularities in attitudes and behavior that characterize a social group and differentiate it from other social groups." [21] Said another way, social norms are guidelines for acceptable behavior within a community as well as what behaviors distinguish it from other groups. Norms may be descriptive, as in what people do, or prescriptive/injunctive, or what peoples' mental models of acceptable behavior should look like [124]. For new members of a community, assessing norms can be challenging as they do not know what is expected of them by others.

Norms are constructed and managed in multiple ways. Many theorists argue that norms are emergent from groups of people who seek balance [125], and that norms naturally emerge out of the shared negotiation as group members decide what constitutes acceptable and unacceptable behavior [126]. These norms may be explicitly codified through formal rules, laws, or codes of conduct; many social norms remain implicit in in peoples' understandings and mental models of the group [127].

Members of a group signal their belonging in a group by conforming to and performing these social norms. Individuals will "perform" what they believe to be socially acceptable, both as a way to indicate their personal belonging to a group [126, 21] and to persuade others that they belong in the group [128, 129]. Norms may be signaled in behavioral cues, such as wearing certain clothes or having certain kinds of body language [129]. In the offline and especially in the online setting, norms can also be situated in community language use [130, 131, 132, 133].

4.1.2 Deviant Behavior

Deviant behavior is actions that are socially outside the norms for a community or group of people [19]. Behaviors can be violations of explicit rules and laws or social standards and norms. Put another way, deviant behaviors are contextual—whether a behavior is considered "bad" is dependent on the person engaging in the behaviors, the circumstances, the community, and the timing of that particular activity [19]. Traditionally, deviant behavior is considered negative or undesirable – classic examples include illegal activity, like assault and murder, and culturally situated examples like having tattoos or cheating on exams. Not all deviant behaviors are necessarily negative; deviance can also be positive, like athletes overcommitting to the sportsmanship ethic [134].

There exist many theories on why individuals participate in deviant behavior, many of which root their analysis in the existence of crime. Functionalists/structuralists, like the seminal work in the space by Emile Durkheim, argue that individuals participate in deviant behavior because of cultural or structural issues that push them to violate norms [135, 136]. On the other hand, symbolic interactionists explain deviant behavior as oriented in human interaction, both that it may be learned from other individuals [137] or that deviance is identified and "labeled" by others [138]. Finally, deviant behavior can also be understood from the perspective of power, privilege, and hegemony, as is discussed by Marx in his critique of capitalism [139] or by feminist scholars in their critique of gendered expectations of behavior [140]. Each of these groups of theories of the origins of deviant behavior can explain the origins and reasons for the behaviors themselves. Deviant behavior, then, can be summarized as behaviors that conflict with norms of a community, either physical, imagined, or virtual.

4.1.3 Online Communities, Norms, and Deviant Behavior

Norms. In online communities, norms are created, negotiated, and managed in ways similar to offline counterparts. Norms may be explicitly written out in community guidelines and rules, Terms of Service agreements, and site-wide usage policies. However, many norms are also contained in the mental models people hold of acceptable community attitudes and implicit understandings of community behavior [127]. In addition, many stakeholders negotiate and iterate on these norms. This can include the platform owners of the site, moderators and administrators, the individual users themselves, local governing bodies, and even lurkers of a site [141, 142]. In HCI in particular, design theories have been proposed to facilitate and promote norm construction and online behavior [143].

Research in online communities has explored norms in a variety of different contexts and sites. Norms promote behaviors that help the community achieve its goals [143], like writing content on Wikipedia [144] or encouraging contributions to fan fiction archives [145]. Social norms can also facilitate newcomer participation [146, 147]. On Slashdot, Lampe and Resnick examined how distributed moderation systems harnessed community norms to find high-quality content [148].

Deviant Behavior. Deviant behavior and undesirable interactions on online platforms have also been extensively studied. Many studies have explored why deviant behavior exists in online communities, using psychological explanations of deviant behavior [149, 150, 20], perspectives on the community itself [128, 151], and managing negative behavior online [152].

In HCI, researchers have thoroughly examined different kinds of deviant behavior online. One prominent area of research has been trolling and anti-social behavior, studied by Cheng and co-authors [153, 154]. Other research in trolling considers moderator-labeled data [155]. Related to trolling, new research has begun to explore automated detection of hate speech and abusive behavior [156, 18]. Other instances of deviant behavior studied in HCI include managing vandalism on Wikipedia [157], aggressive harassment of newbies on the SomethingAwful forums [158], and online harassment [122].

4.2 Instagram Content Moderation and Lexical Variation

In April 2012, the photo-sharing social network Instagram banned some of the most common pro-ED hashtags in response to scrutiny about the existence of communities promoting eating disorders under hashtags like “#thinspo” and “#thighgap.”. Instagram also began issuing advisories for graphic content in their search feature when users searched for other hashtags as well, with the apparent goal of removing this content from their site. Their Terms of Service specifically regulated against, “

In response to this content moderation, the pro-ED community adopted tag variations to circumvent these restrictions, creating lexical variants like “anorexiaa,” “thynsporation,” and “thyghgapp.” To understand how pro-ED Instagram users circumvent these restrictions, I examined lexical, behavioral, and topical changes associated with the emergence of lexical variation in Instagram’s pro-ED communities. These adversarial responses by the pro-ED community are deviant because they go against the wishes of Instagram’s Terms of Service and seek to propagate such behaviors online.

In this project, I investigated the adoption of lexical variation in tags used by the pro-ED community before and after Instagram began moderating pro-ED content. To understand these emergent behaviors by the pro-ED community, I ask the following research questions:

- **RQ 1. Lexical Changes.** How do lexical variations of moderated pro-ED tags evolve over time?
- **RQ 2. Behavioral Changes.** How does posting activity and support manifested in pro-ED posts evolve as lexical variations are adopted?
- **RQ 3. Topical Changes.** What topics characterize posts with lexically variant tags, and how do they contrast to the set of posts with the moderated tags?

This study uses 2.5 million pro-ED Instagram posts from half a million users, shared between 2011 and 2014. After content moderation, lexical variants emerged for all 17 pro-ED tags that underwent initial moderation in 2012. Many lexical variants were adopted by the pro-ED community following the enforcement of content moderation - an average of almost 40 variants emerged corresponding to each moderated tag. Further, engagement on these variant tags through likes and comments was 15-30% higher compared to the original moderated tags. While the size of communities adopting the variations was often smaller and largely non-overlapping with the moderated tags, certain lexical variations reached dramatic sizes (2 to 40 times larger) relative to the initial tag. In fact, lexical variants of tags with content advisories grew by 22% following Instagrams moderation of pro-ED content. I also found that the content associated with lexical variants reflected heightened vulnerability to self-harm and isolation from the greater community of sufferers of eating disorders on Instagram.

This paper suggests that Instagrams current moderation practices are not effective at dispersing the pro-ED community or in controlling the propagation of pro-ED behavior on the platform. Moderation might in fact be amplifying the destructive power of pro-ED posts. My research offers insights into avoidance mechanisms of platform-imposed moderation for pro-ED communities, and whether moderation is a viable intervention mechanism for pro-ED. This project was published at CSCW 2016 as [28]. A summary of my findings follows below.

4.2.1 Data Collection

Instagram's API does not return any posts when queried with banned tags. My data gathering occurred in three steps to work around this limitation: sampling for pro-ED tags that co-occurred with banned tags in posts, a larger data collection, and creating a candidate pro-ED post set by removing noisy, ambiguous or irrelevant content.

First, I obtained post counts for nine “seed tags” related to eating disorders [49]. I collected all posts for these nine tags over 30 days. The resulting sample contained 434K posts with 234K unique tags. Sorting tags in order of decreasing probability of co-occurrence identified 222 tags with at least a 1% occurrence rate, collectively associated with tens of millions of posts dating back as far as January 2011.

I then manually curated the list to exclude tags semantically related to eating disorders and exclude the closely related communities of mental health and eating disorder recovery. I excluded tags too broad to be specific to eating disorders (#fat), tags related to other mental disorders (#depression), and obvious recovery tags (#anarecovery). This reduced the dataset from 222 tags to 72 known eating disorder tags.

Next, I collected the dataset, which contained all available posts tagged with any of these 72 tags from November 2014 as far back as January 2011. This dataset contained over 8 million posts.

Finally, I created a candidate set of posts from this raw set that I confirmed to be related to pro-ED behavior. I removed any posts with three tags ("mia," "ana," and"ed") that did not also contain another tag from my list of 72 tags. This filtering created a dataset of 6.5 million posts.

Defining and Identifying Lexical Variation

Before this project, there were no "gold standard" labels of hashtag variation of pro-ED communities. In this paper, I developed a set of pro-ED hashtags, both identifying the original hashtags that were banned or moderated by Instagram (root tags), the emergent variations on these tags (variant hashtags), and a mixed-methods approach to identify variants in the future.

My observations of Instagram shows that generally variant hashtags (*e.g.*, #thinspoo) emerged *after* the root tag (#thinspo) was moderated in some way. I also noticed that tags maintained semantic similarity, keeping similar meaning and structural components. However, I also noticed that not all permutations that may follow straightforward Levenshtein edit distance [159] were always semantically identical. Therefore, standard lemmatization, edit distance, or spell-correction techniques were not appropriate.

I consider a tag (t^j) to be a "lexical variant" of another tag (t^i) if:

1. t^j is lengthened by repeating any of t^i's characters or other newly added characters.

2. Some of the characters in t^i are permuted to create t^j.

3. Some of the characters in t^i are eliminated to create t^j.

4. One or more characters not in t^i (including alphanumeric characters) are added to or substituted in t^j.

5. A combination of the above criteria is used to create t^j

Finding Root Tags. Instagram does not publish a centralized resource for all moderated tags, and third-party sources on the same are scarce and only include banned tags, not the ones with content advisories.

First, I constructed a tag usage frequency distribution to identify frequent tags in all crawled posts (detailed in the next section). For the top 200 tags, two researchers manually checked for bans or content advisories on these tags. This produced 17 tags that uniquely characterized pro-ED content and have either a ban or content advisory placed by Instagram. These 17 tags served as the set of moderated root tags.

Finding Variants. I identified lexical variants of the 17 root tags in my dataset. For all root tags, I designed a matching regular expression in line with the rules stated earlier, designed to be broad to capture any potential variants. I used this regular expression to extract potential variants from the hashtags of 6.5 million posts.

Two researchers familiar with Instagram and pro-ED content independently participated in a binary rating task to remove spurious and unrelated variants. Each candidate variant was voted yes or no by the researchers who then pooled their responses. Cohens κ of inter-rater reliability was very high (.98), so I only included variants with unanimous agreement.

Table 4.1: A selection of the 17 root tags, total number of variants in each tag chain, and 3 example variants

Root Tag	Num. Variants	Example Variants
ana	9	anaa, anna, anaaa, annaaa
bulimia	49	bulimic, bulimc, bulimi
eatingdisorder	97	eatingdisorders, eatingdis, eatingdisorderr
promia	4	promiaa, promiaaaa, proomia
skinny	18	skiny, skiiny, skini, skynii
thighgap	107	thygap, thyghgapp, thegap
thinspo	40	thinspoooo, thynspo, thiinspo
Total Root Tags	17	
Total Variant Tags	672	

Table 4.1 provides an example of 7 of the 17 root tags and their lexical variants. From the 17 variants, I identified 672 total variants for those 17 tags.

4.2.2 Results

RQ1 - Lexical Changes. To answer RQ1, I investigated the pattern and evolution of lexical variations associated with the root tags, measured with Levenshtein edit distance [159].

For all chains, edit distance of a variant tag compared to the root increases over time a linear trend (least squares) fits to the edit distances of all variants for "anorexia," "eatingdisorder," and "thighgap" yield $R^2 = .2$ (p=.002), $R^2 = .27$ (p=.001), and $R^2 = .34$ (p=.0005), respectively. As newer variants emerged over time after the root, they were increasingly more syntactically distinct ("thighgap" to "thyghgapss").

Table 4.2: Variation patterns among tags in a chain, with respect to the root. Momentum indicates the rate of change of edit distance of variants over time of their emergence.

Tag Chain	Max edit distance	Mean Edit Distance	Momentum
ana	5	2.556 ± 1.257	1.281
bulimia	7	1.755 ± 0.980	1.203
eatingdisorder	5	1.629 ± 0.778	1.156
promia	3	1.750 ± 0.829	1.278
skinny	4	1.722 ± 0.870	1.221
thighgap	5	2.084 ± 1.006	1.218
thinspo	9	3.125 ± 1.952	1.383

I also explored the momentum of change in the tags and show the results of this momentum analysis in Table 4.2. All 17 tag chains show increased edit distance momentum of the variants with mean momentum of 1.3 across all chains (a value of 1 would indicate the rate of change is constant).

RQ2 - Behavioral Change.

Comparing Activity. Figure 3 shows the changes in normalized proportions of posts that correspond to six moderated root tags and the same for three of their most common variants. To determine this normalized proportion of posts, I divided the total number of users who posted on a root tag or any of its lexical variants by the number of users that posted on any tag during the same time slot.

After changing community policies and introducing content moderation in April 2012, posting activity changed in both varied and surprising ways. For the banned tags ("thighgap," "thinspo," and "thinspiration"), the proportion of posts sharply drops when Instagram reported changing its community policies. This is consistent across the other banned tags (not shown for brevity) the use of banned tags decreased 13-78% after April 2012 (mean 52%). However, for root tags with content advisories, I saw a surprising increase in the proportion of posts after the policy change ("ana," "mia," "eatingdisorder"). This increase ranges between 9 and 37% (mean 22%). The emergence and substantial adoption of lexical variant tags only happens after April 2012. While a causal effect may not be directly derived, I believe that this shows a deliberate strategy by the pro-ED community to circumvent content moderation policies and to continue to organize and sustain themselves.

Comparing Users and Support. I examined how the pro-ED community engages and supports posts in the root and variant tags. To measure engagement and support, I used mean likes and mean comments on root posts and variant posts (Table 4.3). There is a statistically significant increase in likes and comments for most variants when compared to the base tags. The mean number of likes in variant posts is higher by 30% compared to the root posts, while comments are 15% higher in variants.

RQ3 - Behavioral Change Finally, in RQ3, I investigated how the context of root

Table 4.3: Engagement (likes, comments) on the roots and their variants. Tag chains with most significant change in mean likes and comments are shown. Statistical significance is test-ed based on Mann Whitney U-tests. Bonferroni correction (/17), where α=.05 (*), .01 (***), and .001 (***), is adopted to control for familywise error rate.

Likes

Tag Chain	Mean	(Root)	Mean	(Variants)	z	
eatingdisorder	53	±55.28	44	± 72.87	-36.21	***
mia	44	±46.37	56	±46.42	32.79	***
thighgap	36	±39.02	52	±49.00	38.55	***
thinspiration	31	±26.35	58	±57.86	64.12	***
thinspo	33	±34.47	53	±50.58	87.16	***
Change in #likes in variant posts vs. root posts						30.6%

Comments

Tag Chain	Mean - Root		Mean - Variant		t	
eatingdisorder	2	±4.80	2	±4.01	-23.76	***
thighgap	1	±3.05	2	±3.97	27.85	***
thinspiration	1	±3.01	1	±3.62	24.50	***
thinspo	1	±3.22	2	±3.95	38.54	***
Change in #comments in variant posts vs. root posts						15.1%

tag use in posts differs from posts containing variant tags. In the data, there are 194,421 tags that co-occur with roots three or more times, while 225,282 tags co-occur with variants three or more times. Before I could compare topical content, I determined that the two sets of tags are considerably different. Mean normalized mutual information (NMI) between the two co-occurrence tag distributions is .32 (higher NMI implies higher correlation).Further, the frequencies of co-occurrence of the tags with roots and variants are also different Kendalls τ between the frequency distributions of the two sets is .28.

To examine the context differences further, I looked at clusters of topics in the set of tags co-occurring with the roots and those co-occurring with the variants using the normalized spectral clustering algorithm on two graphs constructed out of the two sets. The root co-occurrence tag graph $Gr(V, E)$ comprises the tags t^i as nodes, such that t^i co-occurs with one of the root tags in a post and $e^i j$ is in E if the tag t^i has co-occurred with the tag t^j at least five times in posts containing a root tag.

Extracting Themes in Co-Occurrence Tag Clusters. To examine the most dominant themes in the tag co-occurrence graphs, I analyzed the two clusters corresponding to the

Table 4.4: 15 most frequent tags in two dominant clusters extracted from the root and variant co-occurrence graphs.

Tags co-occurrent w/ roots		Tags co-occurrent w/ variants	
Topic I	Topic II	Topic I	Topic II
alone	bodycheck	suicide	smoke
alwayssad	nofood	selfharrrrm	failure
lifesucks	bones	selfmutilation	depression
pain	flatstomach	cutaddict	depressedquotes
unhappy	collarbones	cuts	deadinside
emptyfeeling	skinnyangels	harmingmyself	notgoodenough
anaxiety	thinstagram	scar	addiction
broken	mustbesmaller	razor	wishiweredead
emogirl	fat	bloodsecret123	abandon
sad	tiny	blades	paranoid
sadstagram	assbutt	cutting	callmemistaken
sadsmile	fatty	beautifulpain	useless
anxiety	hipbones	slicemywrists	letmeleave
sorry	beautiful	blood	lost
im_not_okay	pale	die	crying

first two eigenvalues of the Laplacian matrix given by spectral clustering (Table 4.4).

Two researchers familiar with pro-ED social media content and Instagram validated the set of tags in these clusters. They used an open coding approach to develop a codebook and extracted descriptive topical themes for the clusters (Cohens κ was observed to be .84). In Table 4.4, I reported a sample of the 15 most frequent tags in each of the two clusters for the root and variant cases.

The clusters of tags that appear with root tags depict negative emotions and feelings known to be associated with pro-ED. The first cluster of tags co-occurring with root tags depicts expression of sadness and pain ("alone," "alwayssad," "broken") and attributes of eating disorder and anorexia ("pain," "anaxiety," "sadstagram.") The second cluster is associated with thinness and body image depiction where users describe physical attributes of their body ("collarbones," "hipbones.")

The content of the variant tag clusters depict more vulnerable, toxic, and "triggering" content. The first cluster contains tags that bear a tone of self-loathing and self-harm ("suicide," "selfharmmm," "cuts.") These tags also describe depression and reduced self-esteem more dramatically than the other cluster (*e.g.*, “depression,” “deadinside,” “notgoodenough.”) These two distinctive clustering patterns show a tendency of the variant communities to adopt the lexical variations perhaps as a way to subvert Instagram attention on sharing of triggering, self-harm, and vulnerable content.

4.3 Norms Matter

Social support has been identified as a key factor in promoting behavior change, and online weight loss communities are no exception to this rule. Weight loss online health communities (OHCs) provide accountability through weekly weigh-ins, advice on navigating challenging situations, and celebration when hitting a new weight loss low [160]. Clinical research has overwhelmingly shown that OHCs help individuals lose weight with health outcomes comparable to offline groups [161, 162, 163].

However, weight loss can be used in less positive contexts in pro-ED communities. These communities outwardly discuss weight loss, but for physically destructive purposes; they share content promoting extreme calorie deficits, abuse of laxatives and prescriptions, and excessive exercise [4].

These two forms of online support for weight loss goals – one toward healthy behavior change [160] and the other towards harmful or “subversive behavior change” [28] – have surface similarities but are motivated by radically different intentions manifesting as distinct behaviors. The work is a case study juxtaposing norms in two weight loss OHCs: the subreddits r/loseit and r/proED on the social media platform Reddit. r/loseit is a subreddit that promotes “sustainable methods of weight loss.” [164] Conversely, r/proED defines itself as a “support subreddit for those who are suffering with...disordered eating behaviors but are not ready for recovery.” [165]

Using comments on these communities as a proxy for support, I addressed three research questions:

RQ1: What **content** differences in norms characterize social support on r/loseit and r/proED?

RQ2: What **context** differences in norms characterize social support in these two communities?

RQ3: Can I algorithmically predict support to be healthy or subversive, that is, characteristic of the norms prevalent in r/loseit or r/proED, based on their content, context, or both?

I developed a novel computational framework to address these RQs and explore differences in linguistic norms. The framework uses several probabilistic language modeling techniques derived from deep neural networks to understand support. I distinguished between content, or the specific linguistic cues in support, from context, or the meaning and use of these cues in specific ways. Surprisingly, I found that these two communities show similarity in their linguistic content. However, by exploring the context of these linguistic cues, dramatically different behaviors around the seemingly common goal of weight loss emerge in the two communities. Finally, I showed that these content and context norms predict with high accuracy (78%).

These results show that norms matter in how different OHCs direct support for health and well-being goals, and also in understanding how support encourages healthy or subversive behavior change. This project was published at CHI 2018 [33], and a summary of the results is below.

4.3.1 Data

To gather data from r/loseit and r/proED, I queried archived Reddit data through Google's BigQuery in October 2016. From r/loseit, I gathered over 2.3 million comments from July 2010 to September 2016, shared on over 164K posts. r/proED had over 123K comments that ranged from May 2015 to September 2016 from nearly 8.5K posts. Summary statistics for this data are given in Table 4.5.

Table 4.5: Summary statistics of r/loseit and r/proED data.

	r/loseit	**r/proED**
Total posts	164,745	8,468
Post Authors	60,599	1,423
Total Comments	2,301,766	123,407
Comment Authors	172,685	4,067
Total Users	184,109	4,253
Average comments per user	1.601	1.472
Median comments per user	1.0	1.0
Std. dev. of comments per user	3.174	1.233
Average score per comment	2.944	3.036
Median score per comment	2.0	2.0
Std. dev. of score per comment	10.066	3.346

4.3.2 Methods

RQ1: Characterizing Content. To prepare the data for this analysis, I first employ n-gram language modeling to tokenize all lowercased comments, followed by stop word and punctuation removal.

TF-IDF Analysis: Term-frequency inverse-document-frequency (TF-IDF) is a statistical measure that balances for the appearance of the word in a document to its overall frequency in the entire corpus. The importance of a token increases proportionally as its frequency in a document (comment) increases, but is offset by the frequency of the token in the entire corpus (all comments in a community).

Log Likelihood Ratio (LLR) Analysis: For the comments from r/loseit and r/proED, LLR is calculated as the logarithm of the ratio of the probability of a word's occurrence in r/loseit to the probability it appears in r/proED. LLRs range from -1 to 1. Therefore, large positive values imply that the word is more frequent in r/loseit, whereas negative values show the word appears more frequently in r/proED. A value of 0 shows the word is equally frequent in both sources. I normalized the LLRs by the raw number of words in each dataset to prevent r/loseit's larger size of the comment corpus from biasing my ratios.

RQ2: Characterizing Context. To examine the context of specific linguistic tokens, my computational framework borrows from recent advancements in deep neural probabilistic language modeling [166] through word embedding analysis [167].

Word embeddings capture the idea that "a word is characterized by the company it keeps," popularized by Firth [168]. Unlike traditional vector space language models, word embeddings go beyond simple linguistic co-occurrence analysis and reveal latent contextual cues of language use not observable directly in the data. They do so by projecting similar words into a continuous vector space of lower dimension. I used word2vec [167] with the skip-gram neural network architecture to best model word associations to nearby words, with a minimum count of 50 for all words to remove most misspellings. Although word2vec provides pre-trained embeddings, I built and trained the embeddings from scratch due to the uniqueness of the comment data.

RQ3: Prediction Task. I used supervised learning by training regularized logistic regression models. Prior work has shown that these models, due to their high interpretability and ability to handle collinearity and sparsity in data, are well-suited for problems like ours [30]. I fit three models with different predictor variables:

`Content Model`: This model uses the top 1000 linguistic tokens TF-IDF weights as the independent variables.

`Context Model`: This model uses the outcomes of the word embedding analysis performed on the linguistic tokens with the largest TF-IDF weights. They specifically include the 50 most similar words (based on cosine similarity) given by the embeddings corresponding to the 600 largest TF-IDF tokens, taken from both communities.

`Content + Context Model`: This final model combines the independent variables of the above two: the top 100 TF-IDF weights (from the `Content Model`) and the top 50 most similar words corresponding to the 600 tokens with the largest TF-IDF weights (from the `Context Model`).

The response variable for all models is a binary variable, indicating whether a post belongs to r/proED (0) or r/loseit (1). In all cases, I balanced the class sizes of the r/proED and r/loseit datasets by randomly sampling from r/loseit to match the total number of comments from r/proED. After removing deletions and removals, the class size is 115,921 with 231,842 total examples. I used 80% of the data for training, parameter tuning, and reporting goodness of fit; the remaining 20% were held out for testing and assessing model performance.

4.3.3 Results

RQ1: Content Analysis

TF-IDF Analysis

Table 4.6: Left two columns: top 25 most frequent linguistic tokens and their weights in descending order from the TF-IDF analysis. Right two columns: top 25 linguistic tokens with the most positive (left column), and most negative (right column) LLR values across the comments in both communities.

r/loseit		**r/proED**		**r/loseit > r/proED**		**r/proED > r/loseit**	
Token	**Weight**	**Token**	**Weight**	**Token**	**LLR**	**Token**	**LLR**
weight	0.307	like	0.358	faq	0.927	lw	-0.997
just	0.292	just	0.314	myfitnesspal	0.885	thinspo	-0.995
like	0.241	really	0.186	logging	0.766	bronkaid	-0.990
calories	0.186	feel	0.174	wife	0.753	ugw	-0.990
day	0.178	weight	0.168	victory	0.75	wl	-0.986
eat	0.175	eat	0.163	paleo	0.747	hw	-0.970
good	0.165	day	0.154	journey	0.746	eds	-0.969
really	0.145	think	0.146	guide	0.744	ec	-0.968
time	0.136	know	0.146	lifestyle	0.741	purge	-0.954
eating	0.126	good	0.127	action	0.731	purging	-0.948
food	0.115	calories	0.126	cheat	0.716	expression	-0.946
think	0.112	food	0.125	5k	0.710	ephedrine	-0.937
know	0.112	want	0.122	concerns	0.683	ed	-0.934
going	0.110	eating	0.121	jogging	0.670	stack	-0.906
people	0.109	people	0.121	wiki	0.670	restricting	-0.897
want	0.108	time	0.115	fitness	0.667	binged	-0.846
make	0.106	make	0.097	program	0.647	restrict	-0.840
week	0.102	going	0.090	index	0.643	105	-0.831
work	0.102	look	0.086	machines	0.637	idk	-0.811
feel	0.101	way	0.085	sustainable	0.637	disordered	-0.784
lose	0.100	lot	0.083	wagon	0.635	broth	-0.778
way	0.095	try	0.080	tracking	0.629	binges	-0.767
fat	0.095	love	0.080	discouraged	0.629	underweight	-0.764
great	0.090	fat	0.080	success	0.622	binging	-0.755
diet	0.088	water	0.074	trainer	0.619	anorexia	-0.750
look	0.087	body	0.074	mfp	0.618	gender	-0.747

First, I showed the top 25 linguistic tokens sorted by their TF-IDF weights from both communities in Table 4.6.

There is very little discernible quantitative or qualitative difference in the most frequent tokens of either community. To begin, across the tokens listed in Table 4.6 I find 3%-80% (mean: 26.4%) difference in use across the two communities from the TF-IDF weights; this difference is not found to be statistically significant based on a two-tailed Mann Whitney U-test ($U = 311; z = 0.48, p = 0.63$).

Qualitatively, I saw the use of similar function words across both communities, such as "like," "good," "really," and "make." I also observed that words related to weight loss are used very similarly in both communities. I saw the appearance of tokens about regulating food intake, like "calories," "eat," "eating," "food," and "diet" in the comments of both r/loseit and r/proED with nearly the same TF-IDF weights. I also saw similar use of the word "weight" as well as "lose" in both communities.

Log-Likelihood Ratio (LLR) Analysis. Next, I reported the outcomes of the log-likelihood ratio (LLR) analysis. Table 4.6 shows three categories of tokens and their associated LLR values. I did not show LLRs closest to 0 because the results from this analysis align with and repeat the findings from the TF-IDF analysis.

Tokens more frequent in r/loseit (positive LLR) discuss the methods and techniques of weight loss. Users talk about strategies for food tracking ("myfitnesspal," "logging," "paleo," "cheat," "program") as well as new fitness and exercise habits they may be adopting or promoting to others ("5k," "jogging," "fitness," "machines," "trainer.") They also seem to discuss weight loss as a struggle with lifestyle changes ("journey," "guide," "discouraged," "lifestyle," "action," "sustainable," "wagon") as well as celebrating their own or others' achievement of desired weight loss targets and goals ("victory," "success.")

Conversely, in r/proED with tokens with negative LLR, I saw an emphasis on weight loss goals suggesting extreme or dangerous approaches. This includes the use of appetite suppressants ("ec [short for ephredrine/caffeine]," "ephredrine," "bronkaid," "stack," and "broth [commonly used to suppress appetite during fasting])," and symptoms of eating disorders ("purge," "purging," "binged," "restrict," "binging.")

RQ2: Context Analysis

In this section, I present the results of the word embeddings analysis for r/loseit and r/proED. Recall that I assembled a word embedding for each community from the comments, one for r/proED and one for r/loseit. For more details of the most 20 similar words to the TF-IDF and LLR, please reference the paper [33].

To understand these embeddings, I present a discussion of selected quotes from comments where the tokens and similar words from the word embedding analysis are both present. To select comments for consideration, I used the following inclusion criteria. First, the quote must contain a token from the top 25 lists for TF-IDF or LLR close to 0. Second, to analyze high-quality quotes the community endorses as good behavior – a signal of the community's norms – the quote must have a score of `median score + stdev`. For r/proED, the comments have a score of +6 or higher, and in r/loseit, +12.

I explored one token in-depth: *diet*. Quotes and scores have been lightly edited to protect privacy. The numbers after the quotes indicates the net votes (upvotes minus downvotes) it received in its community.

Diet. I saw similar occurrence of the token "diet" between the comments of r/loseit and r/proED, per their respective TF-IDF weights in both the communities. However, this implied similarity of usage in content is not supported by the word embeddings.

In r/loseit comments, "diet" refers to two meanings. The first is specific dietary choices or plans. This is captured in words like "vegetarianism," "lowcarb," "keto," "highcarb," "ketogenic," "veganism," "vlc (very low calorie, a medically supervised extremely low calorie diet)," "slowcarb," and "paleo".

> Where I worked about 5 years ago, I started doing a low-carb *diet*. I basically ate grilled chicken and broccoli all day (and lost over 100 lbs). (r/loseit, +39)

The other use of diet in r/loseit is an abstracted notion of diet, closer to an overall theory of nutrition. This is captured in related words like “calorie-restricted,” “regime,” “regimen,” “dietary,” “lifestyle,” “fad,” “restriction,” and “intake.”

> It’s a crappy cycle. You are overweight and struggling, and in order to succeed you need to change your life, your lifestyle, your *diet*. Everything! (r/loseit, +15)

In comments on r/proED, however, the token “diet” is used very frequently with low or no-calorie drinks, especially sodas – “coke,” “mountain dew,” or “doctor pepper”:

> Dinner: Strawberries and cream (44), chicken alfredo lean cuisine (250), diet coke, tootsie roll (22) =316 (r/proED, +7)

In these two highly upvoted comments, diet soda choices are discussed mostly for the daily food logging threads that happen on r/proED. In these threads, users are encouraged to state a daily calorie goal and report the food they eat. Diet sodas are frequently discussed because of their low-to-no calorie status as well as the appetite suppressing qualities of caffeine.

> Always have a diet soda or bottle of water in your hands so your holding something. (r/proED, +17)

Here, the comment author is advising someone to have a drink in hand at social events to appear normal in eating patterns.

RQ3: Classification Tasks

First, I present the goodness of fit measures of all three models in Table 4.7. Compared to the `Null` models, all three models provide considerable explanatory power with significant reductions in deviances. The best fitting model, the `Content + Context Model` fits

Table 4.7: Summary of model fits. Comparisons with the `Null model` are statistically significant after Bonferroni correction for multiple testing ($\alpha = \frac{0.05}{3}$).

Model	Deviance	df	χ^2	p-value
`Null`	128560	0		
`Content`	106320	999	22240	$< 10^{-15}$
`Context`	93449	1849	35111	$< 10^{-15}$
`Content + Context`	87309	2849	41251	$< 10^{-15}$

out data the best. The difference between the `Null` and the deviance of this model approximately follows a X^2 distribution: X^2(2849, N=263K) = 128560 - 87309 = 41251, $p < 10^{-15}$. Expectedly, I find the second best model to be the `Context Model` that gives $\chi^2 = 3.5 \times 10^4$.

Next, I analyzed performance of the models on the 20% heldout dataset, beginning with the results of the `Content Model`. The `Content Model`'s confusion matrix and results are given in Table 4.8. Using the TF-IDF weights as features, this model has an overall accuracy of 69%, and an average precision/recall/F-1 at 69%. The performance of this model is very good, with a 19% improvement over baseline, a chance model where all test data points are labeled with the larger class's label. Next, my results for the `Context Model` are given in Table 4.8. It outperforms the `Content Model` noticeably. The accuracy of this model is 75%, 6% higher than the `Content Model`. It also gives precision/recall/F-1 values as .74/.74/.74, respectively, which are an improvement over baseline by 25% and over `Content Model` by 6%.

Next, I present an extended analysis of the `Content + Context Model`. This model's confusion matrix and results are given in Table 4.8. I observed that this final model outperforms both the `Content Model` and `Context Model` substantially with a mean accuracy of 78% and precision/recall/F-1 of .78/.77/.78, respectively. The area under the curve (AUC) is .779. Overall, this model improves over baseline by 28%.

Finally, I present an analysis of 20 of the top 50 independent variables/features with the largest positive and negative coefficients (β weights from the logistic regression `Content + Context Model`) – see Table 4.9. Here, positive β values indicate that the presence

Content Model

Actual/Predicted	Class 0	Class 1	Total
Class 0	16362	6715	23077
Class 1	8080	15212	23292
Accuracy	68%	68%	68% (mean)
Precision	.68	.68	.68
Recall	.69	.67	.68
F-1	.68	.68	.68
AUC	.681		

Context Model

Actual/Predicted	Class 0	Class 1	Total
Class 0	17030	6065	23007
Class 1	5907	17367	23292
Accuracy	75%	75%	75% (mean)
Precision	.74	.74	.74
Recall	.74	.75	.74
F-1	.74	.74	.74
AUC	.747		

Content + Context Model

Actual/Predicted	Class 0	Class 1	Total
Class 0	17529	5548	23007
Class 1	4968	18324	23292
Accuracy	78%	79%	78% (mean)
Precision	.78	.77	.78
Recall	.76	.79	.77
F-1	.77	.78	.78
AUC	.779		

Table 4.8: Performance of the classifiers on 20% heldout dataset.

of the corresponding token in a comment increases its likelihood of belonging to r/loseit (Class 1). Negative β values increase the likelihood that the comment will be from r/proED (Class 0).

The positive variables most predictive of whether a post will promote healthy support, *i.e.*, come from r/loseit, overwhelmingly relate to behavior changes associated with long-term weight loss. This includes “myfitnesspal,” “counting,” “moderation,” “c25k (Couch to 5K, a beginner running program)”, and “5k.” I also see the appearance of the contextual meaning of words like “cardio” being more predictive for healthy lifestyle changes. In contrast, beta values that increase the likelihood of a post containing subversive support, or

Table 4.9: Selected features with the largest positive/negative coefficients (β) given by the Content+Context Model.

Feature	β	Feature	β
myfitnesspal	5.01	restricting	-10.86
journey	4.89	thinspo	-9.59
c25k	4.38	purge	-6.30
counting	3.95	bronkaid	-5.37
logging	3.67	laxatives	-5.35
success	3.40	underweight	-5.27
diet	3.36	restriction	-5.01
moderation	3.30	electrolytes	-4.17
wagon	2.92	idk	-4.05
healthier	2.92	free-embeds	-3.84
cardio-embeds	2.63	mean-embeds	-3.81
learning	2.51	broth	-3.47
confidence	2.47	pants-embeds	-3.44
started	2.34	thin	-3.37
self-embeds	2.31	fast	-3.36
12-embeds	2.27	bmi	-3.25
5k	2.19	boyfriend	-3.15
sustainable	2.19	anxious	-2.98
eaten-embeds	2.16	treatment	-2.97
1200-embeds	2.12	gap	-2.97
110-embeds	2.09	cal	-2.93
victory	2.05	recommend-embeds	-2.74
overweight	2.04	world-embeds	-2.63
awesome	2.02	watch-embeds	-2.63

coming from r/proED, match to behaviors related to disordered eating. I see words related to binging and purging cycles, such as “restricting,” “purge,” “laxatives,” and “electrolytes.” I also see a preoccupation with thinness and low bodyweight throughout, in words like “underweight,” “bmi,” “thin,” “thinspo,” and “gap (referring to a gap between the thighs).”

4.4 Implications

Given these two projects about normative behavior and deviance, I present several implications of this research.

These computational methods can also augment current methods to improve understanding of desirable normative behavior in communities. Examples of existing methods include distributed scoring/voting systems [148]. However, for scoring moderation systems in communities like r/proED, scores may perpetuate norms of subversive behavior change [30, 29], not of “high-quality” or “good” support. While qualitative insights are important to characterize normative support behaviors, scaling these insights to large, dynamic, or growing OHCs may be challenging. Using the proposed methods, health and social computing researchers can both understand community norms and support, but also understand the complex ecosystem of healthy and subversive behavior change.

These projects are some of the first that use at-scale, computational approaches to understanding how norms and deviance influence behavior in online communities. For r/loseit and r/proED, I examined how linguistic community norms encourage healthy or subversive behavior change outcomes around weight loss. For Instagram, norms and deviance are crucial to understanding how lexical variants emerged and were adopted by communities that avoided Instagram’s content bans. My approach highlights the importance of considering normative behavior and deviant behaviors of avoiding platform restrictions on behaviors.

Studying pro-ED communities through the lens of norms and deviance is very useful because of their subversive and adversarial relationships with many stakeholders in this space. Pro-ED communities have adversarial relationships with many communities and groups: to the pro-recovery and other mental health communities [14], to platforms and social networks [28], and to society because of their views of body image and thinness [4]. In related contexts, language signatures through communities are one way to examine deviance and their manifestation through adversarial relationships. I expect these adversarial relationships to play out through language, as they have played out in other contexts, like avoiding censorship of authoritarian regimes [121] and through abusive content [18].

Through these two studies, my work demonstrates demonstrate that computational understandings of language is one way to measure and index these shifts in normative behavior and the construction of the pro-ED community in online space. For “#thyghgapp’ and the pro-ED community on Instagram, moderation in April 2012 was tied to the emergence of lexical variants of tags that had undergone content moderation. I developed a rigorous computational method of identifying these hashtag shifts in these communities.

When applied to the pro-ED community, I noticed that lexical variation showed a monotonic increase over time, indicating a desire of the community to avoid outside attention (and moderation) from the platform in RQ1. Although the sizes of these communities adopting lexical variant tags were smaller relative to the corresponding root tags, some lexical variation communities disproportionally increased in size, and also increased social participation and engagement in RQ2. These variants were extensively used to share information encouraging adoption and maintenance of pro-ED lifestyles, often to also share more triggering, vulnerable, and self-harm content.

My computational approach here was able to systematically identify the emergence of lexical variants on the platform and demonstrate changes in community participation after the effects of the bans. While social support and cohesion are linked to improved well-being, norms in pro-ED communities in the aftermath of enforcement of moderation situated such social cohesion for strengthening harmful or deviant attitudes towards body and health. Thus, I concluded that content moderation has been mostly ineffective at decelerating the dissemination and proliferation of pro-ED behavior on the platform.

In a complementary project for "Norms Matter," contextual shifts in language in pro-ED communities indicate distinctive behaviors from healthier weight loss goals and the construction of a community of support for subversive behaviors. The similarity in content between the comments in r/loseit and r/proED was unanticipated in RQ1 and therefore somewhat surprising. We as a community know that weight loss and disordered eating behaviors present very differently, both clinically and behaviorally [6]. Using straightforward linguistic techniques, like TF-IDF, LLR, or other bag-of-words approaches, therefore, would not adequately distinguish these two communities.

Yet, the comments on r/loseit and r/proED used very similar linguistic cues while engaging with support seekers. By looking more strongly at normative behaviors around language use through RQ2 and the use of word embeddings, I could examine *how* these linguistic cues were used in context in the comments. Despite similar “surface goals” of weight loss, these communities actually perpetuated distinct norms and behavior change goals, that I then demonstrate through automatic prediction in RQ3.

These two studies emphasize the need to decipher community norms, social support, and the role of support in behavior change in relationship to the findings of computational studies of these communities.

CHAPTER 5

CONTENT MODERATION AND MANAGEMENT OF ONLINE COMMUNITIES

Finally, I move to the last theme of my thesis: moderation and management of pro-ED communities. There are many challenges finding and labeling this content, questions about moderator load and management, and debates on whether this content should be moderated at all and its implications on society. These are all complex questions, and I hope that my studies on pro-ED communities can shed light into these debates.

In this section, I will describe my prior work with content moderation and management of online communities, focusing on how to identify pro-ED content that is likely a candidate for moderation. I will begin with an overview of related work to content moderation and management of online communities. Then, I will discuss my two previous studies on moderation and management of online communities. Finally, I will discuss some of the implications of this work on moderation practices, and social computing.

The first, "'This Post Will Just Get Taken Down': Characterizing Removed Pro-Eating Disorder Social Media Content" looks at content removed from Instagram by both users and moderators. This work shows that straightforward signals can be found in deleted content that distinguish them from other posts, and that the implications of such classification are immense. I built a classifier that compares public pro-ED posts with this removed content that achieves moderate accuracy of 69%. I also analyzed the characteristics in content in each of these post categories and find that removed content reflects more dangerous actions, self-harm tendencies, and vulnerability than posts that remain public.

The second, "Multimodal Classification of Moderated Online Pro-Eating Disorder Content" uses deep learning techniques to make guesses about multimodal (text and image) content that violates Tumblr Community Guidelines. I developed a deep learning classifier that jointly models textual and visual characteristics of pro-eating disorder content that violates community guidelines. Using a million Tumblr photo posts, the classifier discovers deviant content efficiently while also maintaining high recall (85%). This approach uses human sensitivity throughout to guide the creation, curation, and understanding of this approach to challenging, deviant content.

5.1 Related Work

5.1.1 Online Content Moderation

Since the emergence of online communities, moderation strategies have been a critical area of research [169, 170, 143, 171]. One moderation style allows users to moderate their own content. On one extreme, 4chan content is unmoderated, and the community has informal moderation of good content by "bumping" desirable threads and using the "sage" command to comment without bumping [172]. Many sites build explicit social moderation systems, where users can publicly vote for or against content on the site. This is used on sites like Reddit, MetaFilter, Yik Yak, Stack Exchange, and Slashdot. The effectiveness of these social moderation strategies are mixed and are often dependent on user engagement with these systems [173, 148].

Another style of moderation uses outright banning strategies. In this case, platforms either ban keywords about deviant behaviors, ban users, or delete entire communities. In many cases, outright banning of keywords limits the dissemination of certain kinds of content, like spam or pornography. However, these kind of moderation strategies can have unintended consequences. My prior work found that users developed new lexical variants to avoid these keyword bans in pro-ED Instagram communities [28]. On Reddit, the site banned several hateful subreddits, r/fatpeoplehate and r/coontown (a subreddit dedicated to degrading black people), and the community reacted strongly both for and against the banning [174], even though the ban was found to be mostly effective at eradicating this hateful content on the site [156].

Finally, another strategy uses human moderators to evaluate content. Almost all sites have volunteer or paid moderators to police content. These moderators make decisions about whether content aligns with the rules and community guidelines of social communities. Not all moderator actions are negative—some approaches help editors find beautiful weather photos [175], where others help find good social media content [176] and high-quality news comments [177, 178].

Many communities use combinations of these three approaches. For instance, Reddit has subreddit-wide volunteer moderators that enforce local rules of conduct, subreddits can create auto-moderator rules to automatically pull down content, paid sitewide admins to ban content that breaks its Terms of Service guidelines, and allows for users to self-moderate through its voting system.

5.1.2 Automated Detection of Content for Moderation

In the last few years, there has been abundant interest in designing computational tools to detect content that may need to be moderated on social media. One area of recent interest has been in abusive or toxic content detection. Wulcyzn *et al.* used crowd-sourced annotations to develop a dataset of abusive, personal attacks on Wikipedia data [179]. Chandrasekharan *et al.* built a system that draws on pre-existing data from known communities to [18]. In a related area, cyberbullying detection research has also become prominent – Soni and Singh developed a multimodal representation of Vine videos to detect cyberbullying in both the video and the comments [180]. Research has been appearing around hate speech detection [181, 182, 183].

5.1.3 Challenges in Moderation

For all kinds of moderation, there exist challenges that permeate managing deviant content on social communities.

Modality Challenges: Most state-of-the art automated detection systems process text content to identify eligible content. For example, automated approaches on Reddit and

Facebook often use text analysis to identify word matches, like the automoderator keyword lists on Reddit. Although recent strides have been made in computer vision technology in the last few years, these technologies are not generally adopted in the mainstream by social networks. Exceptions such as YouTube's Content ID[1] for copyright infringement and Yahoo's deep learning system for NSFW images[2] exist but are not commonplace across all types of content to be moderated, nor are they necessarily deployed universally across social media.

Correctly Identifying Behavior: In the case of content areas of concern, classification methods need to be designed for each sub-group of content. Clasifiers must be created to identify trolling vs. abusive behavior, specific kinds of self-harm content, spam, etc. Given the complexities of each kind of deviant behavior and the various representations across social sites with very different rules, it is not surprising that a "one size fits all" approach works for identifying content.

Finding Content for Consideration: Another challenge is finding content that may violate rules. Without automated systems, moderators often rely on reports or alerts that a content violates the community's given rules. Especially in large-scale social computing systems like Facebook or Instagram, manually combing for inappropriate content is impossible for moderators. This may mean content removal biases towards what is reported to moderators by the other users of the site.

For inappropriate content published in deviant communities, community members have little incentive to report this content to moderators or administrators because this is the content they seek. For example, in pro-ED communities on Tumblr, the rate of reporting disordered eating posts was very rare (600 in 3 months time span) compared to the hundreds of thousands of pro-ED posts found on the site. Looking for support in maintaining or engaging with such content, this content will likely not be flagged to moderators for review. So in many cases, the "worst of the worst" content will not be reported to moderator attention.

[1] https://support.google.com/youtube/answer/2797370?hl=en

[2] https://yahooeng.tumblr.com/post/151148689421/open-sourcing-a-deep-learning-solution-for

Labor and Staffing Issues: Human staffing and employment trends also influence moderator effectiveness. Most social networks do not have enough labor to curate the entire site, let alone answer the volume of requests that come in through manual report systems [15]. Moderator labor is often out-sourced, and employee turnover is often in the span of months, making re-training for subtle cues of deviant behavior challenging. These two forces – not enough people to moderate on top of high turnover – make it challenging to develop a persistent and consistently effective moderator pool to judge and manage content.

Emotionally Draining Content: In many cases where human moderators must intervene, graphic content in various forms presents another hurdle for moderation [3]. Moderators are often undertrained and ill-equipped to deal with disturbing content, and in some companies, employees turn over rapidly – some as fast as 6 weeks – on the content moderation teams [15]. Moderated content can be graphic, depicting violent and graphic scenes, like beheadings [15]. Researchers have identified this need for an "emotional shield" to desensitize those who interact with this content on a daily or near daily basis [54]. Without this shield, moderators is likely to be emotionally traumatizing.

Broader Platform Challenges: In addition to challenges with moderating for the individuals in charge of this, there are also broader platform challenges to consider. Platforms may be resistant to change, as was the case with Reddit when it changed community managers [174]. Finally, it is important to note that rules are not always clear, and new content that pushes the boundaries of those rules frequently appear on social networking sites.

5.2 This Post Will Just Get Taken Down

Instagram prohibits self-harm and pro-ED content on its site. Additionally, users can also remove their own posts that they feel are not appropriate for the platform or that they no longer want to have visible to others. Whatever the reason, I consider these two groups of posts to be "deviant content" because they do not conform to the personal or collective

[3] https://www.theguardian.com/news/2017/may/21/facebook-moderators-quick-guide-job-challenges

norms of the platform.

This project examines characteristics of pro-ED posts removed from Instagram using machine learning techniques. I built a supervised learning approach, a binary logistic regression classifier, to distinguish between the content of public pro-ED posts and removed posts. I built the classifier on a sample of over 30,000 deviant and an equal number of public pro-ED posts. I found that the two classes of posts can be distinguished with satisfactory performance (accuracy of 69% and an area-under-curve measure of 76%). I found that deviant pro-ED posts indeed show heightened vulnerability compared to public posts, and these content markers provide important insights into deviant behavior. This project was published at CHI 2016 [30], and a summary of the project is provided below.

5.2.1 Dataset

To allow enough removal time for deviant pro-ED posts, my data collection occurred in three phases over ten months. In all phases, I used the tools in the official Instagram API.

Phase I: Obtaining Pro-ED Data. Using the strategy I adopted in my #thyghgapp project (ref. Chapter 2, Instagram Content Moderation and Lexical Variation), in November 2014, I obtained a sample of 6.5 million posts relating to pro-ED.

Phase II: Gathering Pro-ED Users and Post Timelines. In February 2015, I obtained a random sample of 100,000 active users from the authors of the 6.5 million posts above. I gathered the public timelines of each of the 100,000 users. This set contains over 26 million posts from 100,000 users, with posts shared between October 2010 and March 2015.

Phase II: Gathering Deviant Pro-ED data. In August 2015, I used the Instagram API to check whether the posts from Phase II were still publicly accessible. I randomly sorted the posts and gathered the first 31,000 posts where the post was no longer available on the platform but the user's account still existed. Note that checking whether the user's account was still active was an important step to prevent confounding resulting from a post being deleted because the user removed their Instagram account altogether.

These 31,000 posts represent deviant pro-ED posts for the classification task. To construct an equivalent still-public dataset of pro-ED posts, I gathered the first 31,000 random public posts from Phase II. This gives us 62,000 total posts in the dataset.

5.2.2 Methods

Next, I developed several logistic regression models to learn the characteristics of deviant pro-ED posts as well as to automatically distinguish them from content that remains public. I used a regularized logistic regression to help control for collinearity and sparsity in the data, using the python package `statsmodels` for the logistic regressions.

The response variable is the binary indicator of whether a post is deviant or still-public. For the predictor variables, I considered four different sets and build a model for each set. These variables capture straightforward linguistic constructs in the post's text.

`TagCt`: uses the frequency of tags in a post as predictor variables; I consider all tags which occur 200 or more times, which gives 614 predictor variables

`TagCo`: uses the 500 most frequent pairwise tag co-occurrences in posts, and the 614 tags with frequencies over 200 (from `TTagCt`). Tags or unigrams co-occur if they are used together in a post.

`TagCtCo`: uses 500 most frequent pairwise unigram co-occurrences in the captions of the posts, the 500 most frequent pairwise tag co-occurrences in posts, and the 614 tags with frequencies over 200 (from `TTagCt`).

`TagCtCoCap`: uses 1,000 most frequent pairwise unigram co-occurrences in the captions of the posts, the 1,000 most frequent pairwise tag co-occurrences in posts, and the 614 tags with frequencies over 200 (from `TagCt`).

The model used 80% of the data for training purposes and parameter tuning; the remaining 20% were heldout for testing.

5.2.3 Results

On the training data, I first present the goodness of fit of the four models and how they fared against the Null model (Table 5.1). Compared to the Null model, all models provide considerable explanatory power (statistically significant based on Bonferroni correction) with significant reduction in deviances. Particularly, the TagCtCoCap model (that uses tag frequencies, tag co-occurrences and unigram and bigram tokens in post captions) yields the best fit. I found that the difference between the deviance of the Null model and the deviance of this model approximately follows a χ^2 distribution, with degrees of freedom equal to the number of additional variables in the latter model: $\chi^2(2613, N = 62K) = 107387 - 54487 = 5.29 \times 10^4, p < 10^{-10}$.

Table 5.1: Summary of different model fits. Null is the intercept-only model. All comparisons with the Null model are statistically significant after Bonferroni correction for multiple testing ($\alpha = \frac{0.05}{4}$).

Model	Deviance	df	χ^2	p-value
Null	107387	0		
TagCt	61029	613	9.43e+03	$< 10^{-7}$
TagCo	58266	1113	6.67e+03	$< 10^{-8}$
TagCtCo	57632	1613	6.03e+03	$< 10^{-8}$
TagCtCoCap	54487	2613	2.89e+03	$< 10^{-10}$

Next, I report the results on the 20% heldout dataset. I only report expanded performance metrics on the model with the best performance (the TagCtCoCap model) in Table 5.2.3. In the confusion matrix, class 1 is deviant posts and class 0 is public posts. I find that the TagCtCoCap model gives satisfactory accuracy in classifying both deviant and public pro-ED posts, with a mean precision, recall, and F-1 score of .69 each. The accuracy of the model is 69%, an improvement of 19% over a chance model (50% baseline accuracy due to balanced class sizes).

I also present 20 of the top 50 positive and negative β values of the TagCtCoCap model in Table 5.2.3. The positive and negative β values indicate increased likelihood of a post to be deviant or public, respectively. The variables that are most predictive of de-

Table 5.2: Performance of the `TagCtCoCap` model in distinguishing deviant and public posts.

Actual/Predicted	Class 0	Class 1	Total
Class 0	4433	1913	6346
Class 1	2018	4328	6346
Accuracy	69.85%	68.2%	69.03% (mean)
Precision	.69	.69	.69 (mean)
Recall	.70	.68	.69 (mean)
F-1	.69	.69	.69 (mean)

viant posts are overwhelmingly associated with attitudes and behaviors that reinforce pro-ED lifestyles as well as self-injurious behaviors. These include “cutting,”, “bodycheck” (where users invite others to suggest improvements for their body), and the desire to look or be skinny. There are also indicators of high vulnerability and threats to personal safety (“worthless,” “suicidal,” “razor.”) In contrast, predictor variables that increase the likelihood of a post remaining public are closest to those reaching out to the eating disorder recovery community [14]. Public posts also emphasize a larger variety of emotions, cognitions, and confessions (“gourgeous,” “angry,” “misunderstood.”)

5.3 Multimodal Removal on Tumblr

Tumblr is also challenged by the existence of pro-ED communities as well. Tumblr’s community guidelines prohibit the glorification of self-harm, including promoting eating disorders and their accompanying lifestyles. This includes “content that urges or encourages others to...cut or injure themselves; [or to] embrace anorexia, bulimia, or other eating disorders.” [25]

In particular, Tumblr provides a unique case study for understanding multifaceted and mixed modality content. Not all content in pro-ED communities is dangerous nor do all posts qualify for removal. However, for the content that is dangerous, psychological research shows that there are unique social contagion-like effects on those who are exposed to this content [4].

captionSelected 20 out of the top 50 features' positive and negative beta weights in the `TagCtCoCap` model. There are 3 types of features: TC (tag with at least 200 occurrences), TP (co-occurring tag pair), and CP (co-occurring unigram pair in caption)

Type / Content	β	**Type / Content**	β
TC/#different	2.22	TP/#edrecovery&#beated	-1.33
TC/#energy	1.95	TC/#eatittobeatit	-1.27
TC/#depressedteen	1.30	TC/#toned	-1.23
TP/#ana & #anorexia	1.10	TC/#nevergoodenough	-1.04
TC/#ptsd	1.08	TC/#angry	-0.97
TP/#anorexia & #anorexianervosa	1.00	TP/#ana & #edsoldiers	-0.92
TP/#cutting & #crying	0.98	TC/#misunderstood	-0.92
TP/#depression & #blade	0.94	CP/today & strong	-0.85
TC/#skinnyplease	0.90	CP/fat & depressed	-0.84
TP/#blade & #suicide	0.87	TC/#gourgeous	-0.80
CP/personal & account	0.86	TP/#anarecovery&#edfamily	-0.78
TP/#ana & #weightloss	0.85	TC/#edarmy	-0.77
TP/#sue&#anorexia	0.82	TP/#prorecovery &#anorexiarecovery	-0.71
TC/#harm	0.80	TC/#nourishnotpunish	-0.71
TP/#anxiety &#depressionquotes	0.80	TC/#eathealthy	-0.71
TC/#anagirl	0.79	TC/#ednosrecovery	-0.70
TC/#selfharmmm	0.79	CP/got & think	-0.70
TC/#bodycheck	0.79	CP/know & friend	-0.69
CP/want & look	0.77	TC/#prorecovery	-0.69

Compounding these concerns about deviant behavior are the use of multimodality, or combinations of text and images, in pro-ED communities. These images show explicit negative emotions and graphic content of thinness, motivation for starvation, and even pictures of self-injury [6]. This presents two unique challenges for existing moderation strategies. First, state-of-the-art approaches in automated detection of deviant content primarily rely on textual signals [154, 153]. Second, for human moderators who must interact with this emotionally challenging and sensitive content, it may require domain-specific knowledge as well as an "emotional shield." [54, 15]

In this project, I propose, develop, and evaluate a supervised learning model that distinguishes between deviant pro-ED content and acceptable content from Tumbler data. The proposed model is *multimodal*—it uses both textual and visual characteristics of pro-ED content. I leveraged recent advances in computer vision and large-scale text mining. The model directly incorporates *human assessments* and works to support moderation pipelines with domain knowledge and reduced emotional stress.

To build this multimodal human-machine hybrid classifier, I first curate a dataset of nearly a million pro-ED posts on Tumblr. I then bring in human sensitivity by deploying an iterative expert rating task that identifies thousands of pro-ED posts as potential community guideline violations. Using this annotated data for training, I then evaluate a Deep Neural Network classifier against a Support Vector Machine. This classifier has high performance in detecting deviant pro-ED content with an accuracy of 89% and F1: 65%. I show that this model performs comparably well in classifying the expert rated posts and the posts actually removed by moderators for breaking community guidelines. This work was published as a full paper at CHI 2017 [30], and a condensed version is provided below.

5.3.1 Data Overview

The master dataset contains about 877,000 public photo posts from the Tumblr pro-ED community between November 2015 and August 2016. To assemble this dataset, I used the following approach:

Seed Data Collection. First, I crawled one month of public Tumblr posts to create a set of eating disorder and pro-ED tags. I referred to the list of tag suggestions made in my prior work [28] and by De Choudhury [49]. I developed a short list of 12 known eating disorder and pro-ED tags[4]. This initial crawl with 12 seed tags returned about 100,000 seed posts from May 2016.

Snowball Sampling. Using the seed tags, I used the methods in #thyghgapp to snowball and select additional eating disorder and pro-ED tags in the 100,000 photo posts from May 2016. From this list, I filtered for all tags that co-occurred with the initial set of 12 seed tags above a certain probability threshold (2%). I excluded tags related to recovery

[4] The 12 seed tags were #anorexia, #anoreixa, #eating disorder, #eatingdisorder, #ednos, #proana, #pro ana, #thinspo, #thynspo, #thighgap, #thigh gap

(*e.g.*, #ed recovery) or that were too general (*e.g.*, #ugly, #fat or #sad teen). This produced 31 unique tags for the final data collection.

Final Data Collection. Next, I used the 31 tags to get a sample of pro-ED photo posts shared between November 2015 and August 2016. I ran a list-based filter on these posts to exclude posts with salacious tags (*e.g.*, #boobs) or recovery-related tags. This master dataset contains about 877,000 photo posts from November 2015 to August 2016. For each post, I gathered its metadata and image.

Table 5.3: Summary statistics of the master dataset of 877,998 Tumblr posts. This includes how many posts Tumblr blogs generate as well as the use of tags per post.

Unique Blogs	118106	Unique Tags	297,597
Average Posts/Blogs	7.43	Average Tags/Post	8.85
Median Posts/Blogs	1	Median Tags/Post	6
Standard Deviation	84.47	Standard Deviation	7.56

Table 5.4: Top 25 tags in the master dataset.

thinspo	skinny	ana	thin	anorexia
thinspiration	mia	ed	bulimia	eating disorder
depression	fitspo	thigh gap	anorexic	depressed
suicide	proana	sad	anxiety	suicidal
self harm	motivation	pro ana	goals	tw

5.3.2 Compiling Posts Removed By Tumblr Moderators

The second step of the data collection is finding photo posts removed by Tumblr moderators for violating Tumblr's community guidelines. Tumblr community guidelines prohibit the glorification of self-harm and eating disorders like anorexia or bulimia [25]. When Tumblr staff removes a post for breaking community guidelines, they overwrite the original image with a new image that says that Tumblr removed the original for violating community guidelines seen in Figure 5.1. This action indicates if a post was removed by a moderator.

To gather moderator-removed posts, I first downloaded all images in the master dataset. This download happened in June and August 2016 (to get July and August's photo data). Using a strategy similar to the one I adopt in [30], I re-downloaded the same set of images in September 2016. To detect images that were taken down for violating community guidelines, I tracked images with changes in file size and used a fast visual similarity approach to find images identical to the default image in Figure 5.1. I found 569 posts removed by Tumblr moderators between June and September 2016.

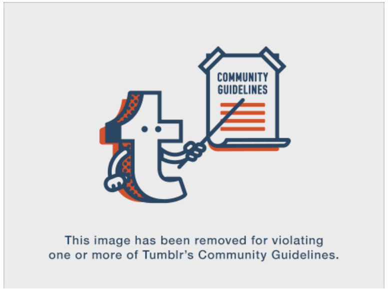

COMMUNITY GUIDELINES
This image has been removed for violating one or more of Tumblr's Community Guidelines.

Figure 5.1: Default image used by Tumblr to substitute images in posts that violate community guidelines.

I found that the proportion of photo posts removed due to breaking the guidelines is low. I have several hypotheses that might explain this phenomenon. Even with the most aggressive moderator removal, social media platforms cannot examine all content that should be removed from their platforms. Most platforms rely on reported data from users to drive their curation and content moderation efforts [15]. Second, in communities like pro-ED, posts glorifying self-harm behaviors like purging or extreme food restriction are less likely to be flagged by community members who are deliberately seeking this content. These communities are often "hidden in plain sight" where anyone can find them, but their tags prevent most outside parties from discovering the community [28].

Collecting Non Pro-ED Data. Finally, I collected a sample of Tumblr photo posts that are semantically close to pro-ED images but are topically unrelated. I first manually inspected a sample of 200 posts in the master dataset that were not removed due to community guideline violations. I obtained tags in these posts unrelated to deviant pro-ED behaviors [30]. Examples of these tags include: #outfit, #fashion, #selfie, #fitness, and #fitchicks.

Using these tags, I gathered a large set of over seven million posts from August 2016 and excluded posts with any of the 31 pro-ED tags from the master dataset. From this, I randomly sampled about 10,000 public photo posts.

5.3.3 Methods

Qualitative Labeling Task

Developing a Rationale for Assessing Guideline Violations. I wanted to objectively capture what might constitute removal from Tumblr's platform for "promoting or glorifying eating disorders or self harm." [25]. As the guidelines specify, this kind of content must, "urge or encourage others to: cut or injure themselves; embrace anorexia, bulimia, or other eating disorders; or commit suicide." [25] Therefore, I holistically evaluated a post to decide if it might be deviant content. I interpreted promotion either as encouraging the maintenance of an eating disorder or promoting related actions. The researchers used the posts' tags, image, the caption, and the author's username.

In the first iteration, three raters independently rated 250 public Tumblr posts randomly sampled from the master dataset (ref. Data section). Ratings were based on a three-point scale: photo posts that would potentially be removed due to guideline violation (rating 3); posts that might be a challenging case of moderation but are actually not guideline violations (rating 2); and posts that should continue to remain on Tumblr (rating 1). Then, the raters resolved rating differences and designed a shared rulebook that included their rationale to assess posts. For the sake of space, details of this rulebook can be found in the paper.

With this rulebook, the researchers rated an additional set of 50 posts to test the convergence of their rating system. Using Fleiss' interrater reliability metric κ, I found high agreement ($\kappa = 0.7$) between the three raters. Additionally, I measured how well the ratings identified posts that should be removed (rating 3) and the posts with clear signals of guideline conformity (rating 1) and found that interrater reliability was higher ($\kappa = .8$).

Another 5000 posts were randomly selected from the master dataset (without replacement from the previous rating samples), and two raters assigned a single score to every post. Prior work shows that the most dangerous content in these communities is relatively uncommon [29] and I predicted it would be the smallest category. To boost the size of this category for the classification task, I identified the top 25 tags on other posts labeled as a 3 in the initial set of 5000. Using these tags to filter, I identified a final sample of 960 photo posts from the master dataset—this narrowed the search space of finding posts most likely to be a 3 so I could quickly bulk up the set of posts with a rating of 3. The same raters rated these additional posts. In total, 5960 posts were rated.

Constructing Training, Validation, and Test Datasets

Bringing the data collection and rated posts together, I discuss the construction of the training, validation, and test sets.

Positive Examples/Class 1: The positive examples (deviant content) for *training* and *validation* included posts that rated as a 3—posts that potentially violate the community guidelines (referred to as PGV or Potential Guideline Violations now on). I split this dataset 80% for *training* and 20% for *validation*. The positive examples for *testing* included posts that were taken down from Tumblr by the moderators (referred to as GV, or Guideline Violations).

Negative Examples/Class 0: The negative examples contained three sets of data: (1) photo posts gathered in the Data section unrelated to eating disorders; (2) the annotated posts that were rated a 1; (3) the annotated posts rated a 2. Because the first two sets of negative posts are likely to have distinctive visual and textual markers in contrast to pro-ED posts, I refer to them as GC-S (or Guideline Conforming—Simple). The third set of posts are challenging or contextually complex, so I refer to them as GC-H (or Guideline Conforming—Hard). I randomly split this combined set of negative examples into 70% for *training* and *validation* and 30% for *testing*.

Representing Images with Convolutional Neural Networks

I extracted visual features of the image posts by starting with the publicly available pre-trained AlexNet model from the Caffe deep learning framework. I experimented with two types of features from this CNN - the $4k$-dimensional $fc7$ features, *i.e.*, the features after the second fully connected layer. Limited by the amount of available annotated data for the problem and in pursuit of a more compact representation, I use dimensionality reduction to get the features down to $d = 128$ dimensions. To reduce dimensionality, I use principle component analysis, learned using $fc7$ features from a public 100 Million YFCC100M image set [184]. I define the visual features for an image i as $\mathcal{V}_i \in \mathbb{R}^D$, where $D = \{128, 4096\}$.

Text Embeddings

To extract text features, I used the *skip-gram* model and *word2vec* training [167]. The principal force behind the skip-gram model is the use of *context as supervision*. As a learning objective in a generic text representation scenario, skip-gram tries to maximize classification of a word based on another word in the same sentence. Here, I do not target tag or word prediction but choose to *aggregate* the individual learned tag representations of posts given by word2vec into a single compact post representation.

I create the tag contexts through *tag co-occurrences* and form the supervision signal from all possible pairs of tags that appear in a post. Given a set $T_i = \{t_1, \ldots, t_T\}$ of T_i tags for post i I construct the set of co-occurring tag pairs $\mathcal{P}_i : \{(t_j, t_k) \forall j, k \in T_i\}$ with the objective to maximize the average log probability spanning all tag pairs in i [167]. I form the tag dictionary $\mathcal{D}$ using the most common (top-$40K$) tags from the training dataset. After learning the embeddings, each tag in the dictionary is now represented by a dense compact vector, that lies in a space $\mathcal{E} \in \mathbb{R}^E$ of dimensionality $E = 128$, where tags that often co-occur are *close*. The word2vec model produces semantically relevant tags in the dataset for ED-related terms: for example, #anorexic co-occurs most strongly with #starving and

#anamia, two tags related to pro-ED actions and ideations.

Learning a Deep Multimodal Neural Network (DNN)

Given a post i with visual features $\mathcal{V}_i$ and textual features $\mathcal{T}_i$, I want to use both to learn whether or not this is deviant pro-ED content. Formulating this as a classification problem, I jointly learn from both modalities. The joint model outputs a function $f(\mathcal{V}_i, \mathcal{T}_i)$ that models the probability $p(y|i)$ of whether a post is deviant.

Layers in the CNN block follow the AlexNet model [185] while all added layers are fully connected. Although the model can be trained end-to-end, I chose to only learn the last layers of each modality jointly with all subsequent multimodal layers.

I trained the model using the Adagrad [186] optimizer and the softmax cross entropy as the loss function. Although the depth and width choices for the model are limited due to limited ground truth data, I experimented with deep architectures with up to 3 joint layers.

5.3.4 Results

I compare the Deep Neural Network (DNN) to text-only, image-only, and multimodal Support Vector Machines (SVMs). For the multimodal SVM, the two are first concatenated. I experimented with both linear and Radial Basis Function (RBF) kernels and found performance to be similar. I report results on the classifier trained with the Radial Basis Function (RBF) kernel (C=100, g=0.01). I also report results using the 4K-dimensional visual features for the SVM.

I experimented with different configurations for the DNN, varying the trainable parameters both through the depth (*i.e.*, the number of layers added) and width (*i.e.*, the size of each layer) of the model. For brevity, I will only present results for the top performing configurations at each depth in Tables 5.5 and 5.6. I refer to different configurations of the model as *DNN-lX-Y*, where X stands for the number of joint layers and Y for their size. During training I used a batch size of 64 and a learning rate of 0.01 in most cases. I set the

Table 5.5: Results for the validation dataset. The best performing SVM and DNN are bolded.

Method	**Tags**	**Vis**	**Metrics**				
			Acc	*P*	*R*	*AUC*	*F1*
SVM	✓		0.89	0.72	0.41	0.61	0.53
SVM		✓	0.85	0.53	0.36	0.49	0.43
SVM	✓	✓	**0.86**	**0.54**	**0.81**	**0.69**	**0.65**
SVM-4k	✓	✓	0.86	0.56	0.50	0.57	0.53
DNN-l1-256	✓	✓	0.90	0.67	0.74	0.72	0.70
DNN-l2-128	✓	✓	0.88	0.57	0.82	0.71	0.67
DNN l2-256	✓	✓	**0.90**	**0.62**	**0.85**	**0.75**	**0.72**
DNN-l2-512	✓	✓	0.90	0.66	0.75	0.73	0.70
DNN l3-256	✓	✓	0.90	0.65	0.75	0.72	0.69

Table 5.6: Results for the test dataset. The best performing SVM and DNN are bolded.

Method	**Tags**	**Vis**	**Metrics**				
			Acc	*P*	*R*	*AUC*	*F1*
SVM	✓		0.88	0.50	0.25	0.42	0.33
SVM		✓	0.73	0.30	0.90	0.60	0.45
SVM	✓	✓	**0.86**	**0.46**	**0.85**	**0.66**	**0.59**
SVM-4k	✓	✓	0.88	0.50	0.46	0.51	0.48
DNN-l1-256	✓	✓	0.90	0.57	0.70	0.65	0.62
DNN-l2-128	✓	✓	0.87	0.49	0.82	0.67	0.61
DNN-l2-256	✓	✓	**0.89**	**0.52**	**0.85**	**0.70**	**0.65**
DNN-l2-512	✓	✓	0.88	0.52	0.60	0.58	0.56
DNN l3-256	✓	✓	0.89	0.54	0.71	0.64	0.61

probability threshold of the classifier to $p = 0.5$. I report accuracy (A), precision (P), recall (R), F1-measure and area under the curve (AUC), also known as average precision.

Classifier Performance on Validation and Test Data

In Tables 5.5 and 5.6, I present results for the SVMs and the DNN over the validation and test sets. I varied model parameters for all methods in the validation set and applied the best performing models at the test set. In this discussion, I report the best-performing models for the multimodal SVM and the DNN-l2-256 models measured by AUC.

In the validation step, the multimodal SVM outperforms both unimodal ones by 8–20%, and the deep models outperform the multimodal SVM by 2–6%. Interestingly, higher dimensionality features (the 4K SVM) did not result in higher performance for the SVM—I hypothesize that this model might be overfitting on the training data with the large feature space. However, increasing the deep model's parameter set size makes the model perform better on both the validation and the test sets (3–12% improvement over all other deep models).

I report results on the test set. From Table 5.6, the DNN outperforms the state-of-the-art SVM. the DNN l2–256 model achieves an accuracy of 0.89, a precision of 0.52, a recall of 0.85, and an AUC of 0.7. This is a 3% improvement in accuracy over the best SVM (0.86) and a 4% increase in AUC (SVM at 0.66). Recall is particularly high, indicating that the method is robust against false negatives. What is contributing to the improvement of AUC is a 6% increase in precision, a significant step in precision. That is, the DNN is able to capture fewer false positives while maintaining recall.

Finally, between the validation and test sets, I see one interesting result: for almost all of the SVM and the deep models, relative performance is preserved between the validation and testing phases. For the best performing SVM, the validation AUC is 0.69 ($F1 = 0.65$) and testing AUC is 0.66 ($F1 = 0.59$). For the DNN, the validation AUC is 0.75 ($F1 = 0.72$) and testing AUC is 0.70 ($F1 = 0.65$).

5.4 Implications of Moderation for Pro-ED

In both of these studies, I examined factors that may predict if a post would be taken off a platform, whether that be Instagram or Tumblr. There are numerous pragmatic applications of this work to detecting content that could be candidates for removal. Automated systems could be designed to help identify posts that were not reported to moderators because of the self-reinforcement within pro-ED communities, or develop online systems to spot pro-ED content as it is uploaded to the social network (I discuss these design implications more thoroughly in Chapter 9, "Discussion.") However, this work underscores an important ethical and social question—do social networks have an obligation to curate this content on their platforms? Social networks, and communities more broadly, are inherently defined both by the content they permit on the platform as well as the content they remove.

One might argue that social networks should allow as much speech as possible, and value that over overly aggressive moderation policies. For issues like copyright infringement or sexual exploitation, social networks have a legal obligation to remove that content; however, for socially contentious subjects like pro-ED, social networks might not necessarily choose to ban them in alignment with their values.

The extreme examples of this in social networks are Voat and Gab, which allow nearly all kinds of content to proliferate on their sites. In some cases, the community or platform may feel that it is better for these people to identify and express themselves as a safety valve against these behaviors. There is also promising research that suggests discussing dangerous ideas might help people disinhibit themselves from self-harm [12]. Some social networks, like Reddit, have struck a balance with this approach—they ban or remove the most outright offensive or dangerous content and quarantine, or isolate these communities from being exposed to

On the other hand, there are those that value moderation for its ability to constrain sentiments that might harm individuals, communities, or larger groups. In particular, pro-ED behaviors show contagion-like effects [187], and research has shown that this discourse can encourage others to maintain or continue their dangerous behaviors [26]. Some might argue that, because the social network can shape discourse, this kind of content should be taken down because it is “toxic” to the community [188]. Recent research has also confirmed these findings, showing that by banning abusive content, hateful Reddit communities disbanded and did not spread to the rest of the site [156]. Another argument for moderation is for user engagement—if the negative content causes people to leave or stop participating, the content should be removed. Other work has also shown that, after banning certain tags on Instagram, the pro-ED community became more insular and focused on more dangerous ideas [28]

Moderation practices pose a risk of these communities moving to the periphery of social networks where any intervention techniques will be increasingly difficult to implement.

It is also important to balance these public health impacts alongside privacy concerns. To what extent can I notify trusted friends, family, and clinicians that someone may be suffering from an eating disorder? I would expect that these suggested strategies would be implemented with privacy-protecting standards in mind. Interventions must be delicately crafted to balance the needs of those who are being targeted – in this case, individuals who participate in pro-ED communities. But, at what point do interventions on social media become counterproductive or possibly manipulative? There is a necessary balance between providing dignity and respect to those who suffer from EDs and/or engage in pro-ED behaviors and violating an individual's privacy to deploy an intervention which could, in some scenarios, save their life.

Detection and intervention will always be reactionary to new trends of deviant communities to avoid detection and hide in plain sight. Thus, any kind of intervention technique is a "game of cat and mouse" for many social networks—pro-ED is only one example of a community strategically avoiding oversight. This is by no means an exhaustive analysis of the benefits and drawbacks to content moderation, nor do I try and decide whether these platforms have these social obligations. In fact, this only scratches the surface of a complicated set of issues around content moderation, social networks, and deviant behavior like pro-ED. It will take collaborations from industry professionals, researchers, designers, psychologists, and other stakeholders to make decisions in this area.

CHAPTER 6
A TAXONOMY OF EMERGENT ETHICAL AND METHODS CHALLENGES

Last year, Facebook unveiled automated tools to identify individuals contemplating suicide or self-injury [189, 190]. The company claims that they "use pattern recognition technology to help identify posts and live streams as likely to be expressing thoughts of suicide," which then can deploy resources to assist the person in crisis [189]. Reactions to Facebook's suicide prevention artificial intelligence (AI) are mixed, with some concerned about the use of AI to detect suicidal ideation as well as potential privacy violations [191]. Other suicide prevention AIs, however, have been met with stronger public backlash. Samaritan's Radar, an app that scanned a person's friends for concerning Twitter posts, was pulled from production, citing concerns for data collection without user permission [192], as well as enabling harassers to intervene when someone was vulnerable [193].

Since 2013, a new area of computer science has emerged that harnesses social media data to understand expression related to mental disorders. These algorithms are powerful enough to infer with high accuracy whether an individual might be suffering from disorders such as major depression [16, 58, 75, 85, 73], postpartum depression [59, 71], post-traumatic stress [17], schizophrenia [79, 78], and suicidality [194, 195]. These algorithms can also reveal symptomatology linked to psychiatric challenges, such as self-harm [89], severity of distress [29], or cognitive distortions [86]. Together, I use the term predicting **mental health status** to describe these mental disorders and related symptomatology.

Computer Science (CS) researchers and clinicians are now poised to learn more about the earliest manifestations of psychiatric disorders through social media data. New insights could prevent the development of latent conditions, mitigate the impact of emerging disorders, or as exemplified by Facebook's new suicide AI, new opportunities to intervene with life-saving assistance. With the rising prevalence of mental disorders [196], many researchers see the benefits of better screening, identification, and intervention assisting to promote better health and well-being worldwide.

However, the examples of suicide prevention AIs demonstrate major concerns for algorithmic development and their implications. This includes new concerns about consent into monitoring or intervention systems and privacy and data management questions. Ethics boards do not have standards for managing social media research, and the prediction of mental health status raises new questions about consent, vulnerable populations, and online communities. There are also methodological concerns of data collection and bias, validity of these results for clinical assessment, and the application of machine learning methods to predicting mental health status. Furthermore, the lack of consistency with methods across this research space makes this problem more troubling. For implications, actors with many motivations can misuse data and predictions, and amplify the harms of algorithms in reproducing unfair stereotypes and discrimination of individuals with mental disorders.

As these technologies are developed to detect mental health status, these concerns will grow unless we as a field rectify these problems. We stand to gain much from this research – in better understanding and making interventions in mental health. Addressing these concerns will resolve questions around rigorous science in the area, benefit clinical research, and safeguard well-being for individuals and society. Many of these concerns are not limited to just mental health and social media and apply to other application domains of these technologies that touch on sensitive issues.

In this section, I present the first taxonomy of issues in algorithmic prediction of mental health status on social media data. In answering these questions, I offer insight into questions on how to ethically and rigorously apply machine learning and AI to sensitive domains such as mental health. First, I discuss the gap between ethics committees and participants in such research, on what can be sensitive and sometimes stigmatizing data. Second, I identify tensions in methods and analysis, such as construct validity and bias, interpretability of algorithmic output, and privacy. Finally, I examine implications of this research in benefiting mental health research, challenges faced by key stakeholders, and the risks of designing interventions. This work was published at ACM FAT* in 2019 [34], and a summary of the taxonomy follows.

6.1 Related Work

Complementary to research in predicting mental health status (see Chapter 3, Related Work for an overview) is a long history investigating the ethics of computing technology on broader domains. In fact, some of the gaps we note above, such as participant consent, role of ethics boards, and challenges to autonomy and privacy, have been discussed at length in these works. Given the growing significance of machine learning and algorithms in different domains, this field has received renewed attention both within the FAT* community [197, 198] as well as the field of "critical algorithms. " [199, 200, 201] I provide a brief overview of relevant research in three spaces: social media research ethics, public health research, and critical data studies.

6.1.1 Social Media Research Ethics.

Ample research has addressed issues in social media and ethics, as early as 2004 [202]. Moving into the age of "big data," scholars are considering how new methods and data aggregation techniques impact individuals involved in this research. Hargittai analyzed the snowballing effects that of unintended biased sampling of social media data on big data analyses [203]. Zimmer has examined ethical use of Facebook data [204] and proposed a topology of ethical issues from Twitter research [205]. Finally, Olteanu *et al.* considered the methodological challenges of mining social media for information, including issues of internal and external validity, data curation, and methods [206].

6.1.2 Public Health and Ethics

Second, I look to the history of public health research, social media, and ethics for population-scale predictions of disease and disorders. Dredze and Paul consider social media research for public health, focusing on end-to-end consideration of study design, identifying target conditions, methods, and ethics [207]. Next, Conway and Connor address advances and ethics of population-scale predictions of mental health, providing an overview of the field and reflecting on how "big data" methods like machine learning and NLP facilitate surveillance of mental health for populations [208]. Metaphors for social surveillance of public health have been proposed, like Vayena *et al.* 's "digital epidemiology" to understand ethical obligations of researchers using public data [209]. Horvitz and Mulligan analyzed the potential legal, privacy, and data protection issues of big data analysis for well-being [210]. Norval and Henderson unpack various theories of privacy to analyze whether informed consent should be gathered in social media health research for patient information [211], while Mikal *et al.* used focus groups to understand users perceptions of social media data use for mental health research [212].

In NLP, Benton *et al.* recently considered the protocols for ethical social media health research from their own experiences in the field [213]. Their work discusses the ethical contention surrounding the use of public social media data for population health inference and its exemption from review by U.S. Institutional Review Boards (IRBs). Stylistically, this work is closest to my position, although the ethical guidelines provided by Benton *et al.* are geared toward public health needs, not individualized predictions

6.1.3 Critical Data Studies

. Finally, the intersection of critical technology research and big data has led to "critical data studies," providing useful metaphors and case studies on the impacts of big data research. In an early work, boyd and Crawford push the new field of data science to critically consider its methods [199]. In response to the failure of Google Flu Trends, Lazer *et* *al*. cautions researchers to be cautious in applying predictive techniques [214]. Foucault-Welles brings light to the discriminatory impacts of aggregating analysis of social data that erases differences of minority groups [215]. Metcalf and Crawford discuss the difficulties of using other research relationship metaphors (such as the physician-patient metaphor) to illuminate how data researchers could conceptualize their users as more than just data sources [200].

These three perspectives discuss important concerns: participant consent [213, 200] and contextual data integrity [209, 215]; data protection, anonymization, and privacy [216, 210, 205, 204]; methodological rigor [206, 57, 214]; bias and validity [203, 206]; and implications of the research for different stakeholders [199, 208]. Drawing from these two larger domains – the state-of-the-art on mental health status prediction and surrounding discussion – I identify three areas of tension that encapsulate concerns in this research area.

6.2 Participants and Research Oversight

Reacting to unethical behavior in medical and psychological experiments in the 1940s and 1950s, many countries have adopted ethical research standards for human subjects research. These standards manifest in an ethics committee, whether that be an Institutional Research Board (IRB), Federalwide Assurance-certified ethics board, European Union ethics committees, and corporate internal review committees. Researchers and clinicians must also follow legal requirements to protect the dignity and privacy of individuals. In the United States, the Belmont Report and accompanying Common Rule legislation set protocols for human subject research which receives federal funding [217]. Further, the Health Insurance Portability and Accountability Act (HIPAA) protects privacy of patients in clinical relationships with doctors in the U.S and privacy rights of medical records [218], with similar protections in other countries [219].

Guided by the principles of respect, beneficence, and justice, US IRBs deliberately transform people into "research subjects" in scientific inquiry; this transformation prescribes people with certain rights, protections, and obligations that must be protected [200]. In clinical studies, this obligation is at the forefront of experimental design [220].

Is predicting mental health status on social media human subjects research? How do we assess harm of this mental health research without the oversight of an ethics committee? In this section, I discuss challenges of predicting mental health status outside a clinical setting using data-driven algorithm, and impacts to participants.

6.2.1 The Unclear Role of Ethics Committees

Analysis of publicly visible social media data is often exempt from research protections provided to subjects through ethics committees. These studies are exempted for two primary reasons: one, in large-scale data analyses, there is no interaction or intervention with subjects because the research is observational; two, the data being used was publicly available when collected. Many ethics boards consider social media to be public space synonymous with gathering publicly available data that might be stored in Census records or courthouses.

I find this interpretation consistent across different countries and in different research environments [17, 75, 59]. Researchers will often cite one or both of these principles in their data collection sections – there exists no relationship between researcher and social media user, nor a doctor-patient relationship that would mandate medical privacy guidelines come into play. Studies that do interact with subjects, through surveys of crowdworkers [85] or individuals recruited through word of mouth, advertisements [85] or through apps [70], tend to declare appropriate ethics board approval.

However, predicting mental health states using public social media data emphasizes whether this research should be exempt from ethics committee oversight. Unlike in public health [213], predicting mental health states, even if with public data, borders on medical diagnosis, such as predicting the presence of schizophrenia. Research is more than just the “sum of its parts,” and extensive secondary analysis can be done from traces of social media data [17, 85]. Mental health is a complex and sensitive area that can be isolating and stigmatizing [221], and harm can be difficult to evaluate, especially in second-order impacts. Is this research human subjects research? How should ethics boards handle this new research paradigm?

6.2.2 Consent at Scale

In traditional human subjects research, participant pools rarely exceeds several hundred. This is because inference about mental health states could only be learned through clinician-patient relationships or lab studies that naturally limits the subject pool. By consenting into this research, participants are aware that they are part of research and therefore being surveilled. Consent could meaningfully be gathered from participants, and served as an important signal for participation.

Unlike clinical mental health studies, social media datasets can contain millions of public posts [88], and user accounts regularly exceed the hundreds of thousands [29] – obtaining consent at this scale is pragmatically impossible. However, there are tensions between the infeasibility of obtaining consent and conducting analysis about mood and well-being on social media. This emerged in scrutinized experimental studies of Facebook data [222], where researchers manipulated the mood of millions of Facebook users without consent. In fact, a recent survey study, though not specific to the mental health domain, found that few social media users were aware that their public content could be used by researchers, and the majority felt that researchers should not be able to use tweets without consent [223]. Essentially, passively collecting data transforms its initial purpose, and we miss essential details of individuals’ experiences and symptomatology that may be gained from clinical relationships. Is consent necessary in these contexts, and if so, what is meaningful positive or negative consent?

6.2.3 Vulnerable Populations and Risk

Vulnerable populations, such as prisoners, expectant mothers, and minors require additional protocol to protect participants in the U.S. IRB system [224]. Even riskier research topics, such illegal behaviors, are protected with additional scrutiny. For example, the National Institutes of Health releases certificates of confidentiality that prevents research data from release to anyone, including government authorities [224].

No restrictions exist for studies of public social media users, no matter how vulnerable the population may be. For example, the median age of onset for eating disorders is between 18 and 21 [225]. Given that demographic attributes such as age are inferrable from social media language [226], should we research online eating disorder communities, knowing a large subset of these individuals are likely minors [29, 31]? When should data scientists consider vulnerable populations, and how should we protect this data?

Additionally, ethics boards mandate that researchers take actions to protect against risks that a study may cause for mental health. Many clinical studies include a risk management protocol, where participants identified by the research team to be at an elevated mental health risk can be directed to appropriate forms of help and support resources. Researchers can also intervene to stop participation in scientific research if the subject or research team believe the harms outweigh the benefits.

Even in studies without directed interventions, the presence of researchers in communities could be triggering for individuals with mental disorders. For example, individuals dealing with schizophrenia and fear of mass surveillance may be upset by the knowledge that researchers are tracking their behaviors, even if for beneficial outcomes. Protocols for risk management and drop outs are missing or unimplemented in social media research on mental health. There is no insight into what happens when users "drop out" of social media participation [31], which is a close proxy to withdrawing consent. Are they switching accounts, exiting the platform entirely, or is their mental health state dire? Should we provide information to participants who may be in a dire mental health state?

6.2.4 Contextual Integrity of Communities

Although online communities may post publicly to find support for anxiety [77] to suicidality [81], it is unclear whether social media users understand if their data can be surveilled as they discuss sensitive issues. Behavior in these communities indicate that these groups may have no intention of being discovered by others [29], and they may outright refuse participation in research [202]. When asked directly if users were comfortable with predicting depression with their Twitter profiles, comfort with such research is decisively mixed [212, 223].

Are we violating community norms with these observations? I draw from the notion of "contextual integrity" proposed by Helen Nissenbaum in understanding privacy violations [227], and a related follow-up by Zimmer about contextual gaps in big data research [228]. Zimmer argues that these gaps cause violations of "normative bounds of how information flows within specific contexts." [228] Is is appropriate to observe online health communities for research if it violates this contextual integrity? What about benign discussions on personal social media accounts?

As Bruckman recommends, one way to resolve this contextual gap is by asking for permission through community leaders [229], which is feasible for Reddit or public Facebook groups. However, most research is done on Twitter data, where no formalized community structure exists, and those that do (like hashtags) are amorphous. Must we ask for consent in these scenarios to maintain contextual integrity, and if so, how would we do this?

6.3 Validity, Interpretability, and Methods

The diversity of fields this research pulls from as well as the venues it publishes in brings many methods questions to the forefront of this work. However, there are documented inconsistencies and unanswered questions in this space (ref. section 2). In this section, I discuss ethical tensions arising from the validity and rigor (or the lack thereof) of new algorithms that infer mental health state.

6.3.1 Construct Validity

The American Psychiatric Association's Diagnostic and Statistical Manual of Mental Disorders (DSM) is the best resource for identifying psychiatric symptoms and classifying mental disorders [8]. With over 60 years of empirical support, the DSM guides clinicians and researchers to make accurate psychiatric diagnoses using tested and validated constellations of symptoms and experiences obtained through clinical interviews.

Moreover, clinically and psychometrically validated scales measure the presence and severity of mental disorders, such as the Patient Health Questionnaire (PHQ-9) or the Generalized Anxiety Disorder scale (GAD-7). It is unclear if mapping these scales to digital contexts validly reproduces results. Further, the complexities of patient-clinician interactions make rote application of DSM guidelines to online social media data unclear—DSM guideline for diagnostic criteria of certain illness may be misinterpreted, exaggerated, or even lied about on social profiles.

As technology can sense psychiatric symptoms, identify, and potentially diagnose mental illness, we must consider how best to incorporate these tools into clinical practice. How do we map symptom assessment techniques to social media data in a way that preserves its validity? Is it ethical to use mappings of traditional symptomatology or non-traditional ways to predict mental health?

Related to this is valid gold standard labels of mental health status, or "ground truth." For prediction tasks in this space, gathering ground truth data measure the target/predictor variable (mental health status); it is therefore a crucial part of the research process and impacts the quality of the algorithms that are built. There are several standard approaches in the research on assessing ground truth of mental health status, including self-disclosure of mental health state [58, 17, 230], specific hashtag use [88, 31], and community participation [77]. Other styles directly recruit participants and administer screeners, then collect social media data of these participants [16, 85]. Most studies do not include clinical annotation; however, new approaches incorporate clinicians directly in labeling ground truth [78] or validating the accuracy of other sources [29]. These approaches will vary, depending on the research question and study design.

However, there is no guidance on how to select the correct ground truth collection procedure, or whether clinicians are necessary to this process. Are we measuring the phenomenon we argue we are measuring? Are certain kinds of measurement more appropriate for different scenarios? To prevent misinterpretation of the inferences, must we involve clinicians to assess ground truth states?

Data Bias. Bias is a concern for any project; for mental health status prediction, bias is worrisome for the perceived validity and quality of research output. I focus on population biases in datasets (for an excellent analysis of bias, see Olteanu *et al.* 's survey [206]).

Population bias refers to differences in characteristics between samples in a dataset and those of the target population we intend to measure [206]. The individuals in our datasets (those with a certain mental health status *on social media*) are a subset of the target population (those with a certain mental health status). By gathering data from social media, we bias our data to those who use social media, meaning it is likely a younger and more technologically literate sample than the population as a whole [206].

For mental health status, this bias can manifest in unique ways, leading to ethical lapses and challenges. One well-grounded source of ground truth data is self-reported, diagnosed mental health status (*e.g.*, "I was diagnosed with schizophrenia.") This was pioneered by Coppersmith *et al.* to unobtrusively identify those with mental disorders [58], and has been validated and used in subsequent projects [17, 79, 78]. By sampling those who publicly self-disclose their mental health diagnoses, this subsample has at least two biases. First, these individuals have (likely) been diagnosed with a mental disorder, meaning they are likely to have sought professional treatment to receive those diagnoses. Second, they are comfortable enough to disclose their mental health status to others, meaning that their forms of sharing could be different from others.

I acknowledge that bias is impossible to avoid in any sampled dataset; however, unaccounted bias can cause latent problems, especially when inferences are incorporated in real life situations. How should we sample and correct for bias? How do we handle these biases in generalizing our results to new mental health statuses, social networks, or contexts?

Algorithmic Interpretability. Next, I discuss ethical challenges arising from a need for algorithmic interpretability and performance [231]. On one end of the spectrum are interpretable models, as in many types of regression models like generalized linear or logistic regressions. As input, these models take intuitive features, derived from social media behavior, known symptomatology [113], or innovations in sub-domains like character n-grams in NLP [232, 195]. As output, these models produce easy-to-understand metrics of model fit and coefficients and probabilities of salient predictors. A strength of these models is that they are easily interpreted by clinicians and stakeholders who may not have technical expertise in algorithmic interpretation, especially when matched to known symptomatology. However, interpretable models have been known to suffer from poor performance [83, 75, 16]. Regressions and similar algorithms are also limited by data modality, as they do not handle image and video data without extensive preprocessing. Sacrificing performance in the name of interpretability limits applications to applied research. Simply discovering relationships between predictors and outcomes, such as risk to a certain mental illness, can be insightful to stakeholders like clinicians; however, it remains unclear how imprecise insights can be actionable during risky situations.

On the other hand, deep learning techniques have emerged as state of the art for powerful and accurate models in prediction tasks. Trained on millions of data points, these algorithms can "effortlessly" outperform other models, handle images and audio, and can intuit features out of the data without human supervision. Performance using deep learning techniques has seen noticeable improvements in predictive power in this space [88, 83]. However, deep learning has a key limitation – they do not produce intelligible feature sets for human understanding [233]. These algorithms are “black boxes,” producing impressive results but providing little insight into how the algorithm made its decision. This can make relevant stakeholders in the process, concerned about adopting these algorithms into practical scenarios. Opaque models runs the risk of not only misconstrued and biased conclusions on sensitive data, but also can lead to poor accountability to abide by ethical research principles as well as correcting algorithms when they fail to predict correct outcomes.

These models also challenge human interpretability of their outcomes. How do we handle results that might not align with our clinically-grounded understanding of mental health? These insights might propel research into new areas of signs of mental illness; but, they may also be *red herrings*, providing false hope when in fact the algorithm has latched onto qualities of a particular training set. Multiclass predictions complicate this when they discretize mental health in mutually exclusive binaries (*e.g.*, anxiety or not; depression or not) [58]. The clinical literature overwhelmingly points to mental disorders as frequently co-morbid, and disorders can manifest over a continuous spectrum instead of clearly delineated outputs [8]. Existing algorithmic approaches are often not subtle enough to model this continuum or incorporate interactions between disorders and self-reported symptoms, leading to “artificial” notions of risk.

Performance Tradeoffs. Risks of error in predicting mental health status should be addressed, especially when these algorithms may be used in consumer-facing intervention systems.

False positives, or incorrectly identifying the presence of a mental health status, could cause dramatic consequences for individuals who are the subject of such errors. Many mental disorders are stigmatizing and embarrassing, and being labeled as “disordered” can damage someone’s self-esteem, employment prospects, and reputation [221], as was the case of Samaritan’s Radar [192]. Depending on implementation, false positives can also cause undue stress on individuals who may now believe something is wrong with them, perhaps stifling their sharing on social platforms in the future. When used in scenarios like content moderation or engagement with a clinician, many false positives may overburden key stakeholders with too many requests to deploy assistance.

On the other hand, a false negative means that mental health status was incorrectly labeled as not having a certain mental health status. Pragmatically, this means no intervention is triggered and no risks for interaction take place. However, in practical use of these systems, false negatives mean that mental health status is missed and may go untreated, as mentioned in prior work [86, 234]. These risks become more concerning when dealing with grave mental health statuses, such as suicidality and psychosis. False negatives also raise responsibility and accountability questions for the results of these algorithms. If being used in functional or practical scenarios, which metric is more important to prioritize? If these algorithms "miss" someone, who is responsible for not intervening? Does this reduce clinician accountability in these scenarios?

Data Sharing and Protection. Even after careful data analysis come risks to privacy for participants. I focus in this section on the risks of data sharing and publication of sensitive information (for excellent overviews of privacy risks, please see Zimmer and Proferes [205], and Horvitz and Mulligan [210]).

Scientists share datasets for reproducibility and consistent benchmarking of new algorithms. However, sharing datasets is complicated by mental health research goals. These datasets are collected under specific circumstances, and users may find issues with context changes. Second, datasets are rarely cleaned for deleted or removed data. In the case of mental health discussions, deleted or removed data could have particularly sensitive data, or data that does not reflect the public perception a person wants to have. How do we manage the joint goals of promoting scientific reproducibility while also protecting participants? What does a benchmarking dataset look like for mental health?

Second is publication of sensitive information such as names, locations, and other personally identifying information. When processing textual social media data, algorithms can occasionally latch onto predictive textual cues; this is amplified when sample size is small. To combat this, researchers have various levels of privacy preservation techniques, such as removing usernames from data before analysis [195] or de-identify algorithmic output later [29]. When should we curate our datasets – pre or post-processing? What are appropriate ways to de-identify data to preserve individual privacy, while maintaining data integrity to promote good science?

A related risk comes from using exemplary social media postings/quotes in papers. Recent work by Ayers *et al.* found that, of papers that use quotations in papers, over 80% of participants in datasets are able to be reidentified [235]. Other methods, like interview studies, have guidelines on modifying quotations in publications to protect participant identity [229], and I ask similarly: are quotes necessary for demonstrating the validity of the results of the paper? If quotes are needed, what protections can be used for privacy?

6.4 Implications for Stakeholders

Using the perspectives of relevant stakeholders, the final section deals with numerous implications in this research area. I focus on the impacts to researchers in this space, the individuals who are the target of predictions, as well as social networks.

Emotional Vulnerability. Researchers and practitioners, especially those from CS, are not often taught how to manage complex emotions when engaging with mental health content. Mental health content can contain graphic and disturbing content, like pictures of self-harm, or detailed discussions of suicide plans [29, 81]. Those who engage with this content can be traumatized by these encounters, and traditional approaches to research design do not take into account the researcher's own emotional well-being [236]. For those who are rarely taught to handle sensitive or emotionally-laden information when annotating and interacting with data, how do we train CS and data scientists to handle the weight of this work?

Skillset Mismatches. There are unique challenges in recognizing and rectifying skill gaps in interdisciplinary research collaborations. Both sets of domain experts must actively work to communicate their research processes and decision-making guidelines this work. As mentioned before (ref. ”Algorithmic Interpretability”), algorithmic output can be complex and inscrutable to outsiders. CS researchers are often experts in data collection, feature engineering and model tuning, and performance enhancement. This information needs to be made interpretable to clinicians and other stakeholders with insights into the process. Likewise, CS researchers may lack training in the skills that clinicians traditionally possess. This may be in assessing valid signals of mental health, acquiring ethics board approval, and interpreting signal in datasets.

Some of these decisions may compromise the performance of models—if a clinician suggests removing a highly predictive feature because it is not clinically relevant to predicting depression, the research team will need to negotiate how to proceed. For these partnerships to blossom, both sets of researchers have to be mindful of making such interpretations accessible to build trust and reliability between collaborators.

Role of the Clinician. Data collected passively/actively or continuously/intermittently may imply different responsibilities for clinicians involved in this research. After entering into a physician-patient relationship, clinicians are bound by the “duty to treat,” where they must provide treatment in accordance with their best judgment to their patients. Failing to act on this knowledge would be unethical and potentially illegal. For example, a physician who discovers expressions of suicidal ideation by examining their patient’s social media may be bound to treat and therefore intervene.

However, in this field, data is both passive and actively gathered. Information gathered and analyzed passively may not necessarily imply such a strong ethical responsibility for the duty to treat. For example, a clinician annotates an algorithmically-gathered dataset for intent to self-injure and discovers someone states that they plan to commit suicide at a specific date and time – does a clinician have an obligation to intervene? The obligation for intervention here may be weaker, because there is no relationship developed between clinician and social media user.

However, there also exists the “duty to rescue,” where a bystander has an obligation to rescue another party in peril. Unlike the duty to treat, the duty to rescue has far more varied interpretations. Does the duty to act or rescue vary depending on the type of professional on the project? In many cases, mental health professionals and computer scientists work in tandem - but what when they work separately? Are computer scientists bound by the duty to rescue if they see someone that says they will harm themselves?

Another question is incorporate these new technologies effectively and ethically into clinical care. How the data is collected, monitored, and presented to the clinical team will alter responsibilities and expectations for clinicians and researchers. For example, research in this space often suggests that insights from this data could be given to clinical care teams [31]. How do we design data interfaces that make sense of these algorithmic predictions for effective insights? How do we not overburden clinicians with large amounts of data and direct their efforts?

Designing Interventions. Another implication is the ability to design interventions, one of the most mentioned applications of this technology in the literature [79, 59, 85]. With suitable performance, the results of these algorithms could provide alerts to help identify moments of crisis, assist in the early identification of mental illness, or avoid risky episodes. The potential for great societal benefit of these prediction algorithms is rooted in these interventions; however, design and implementation of interventions remain a key concern. Outside of clinical interventions, numerous stakeholders are cited as potentially invested in this work, ranging from social networks, crisis hotlines, caregivers, and individual friends to family members. If we detect that a person might be suicidal, should we alert experts or close family members? The automated use of such technologies has been controversial when deployed for Samaritan’s Radar [192], but has been better received when driven by human intervention systems on Facebook [189].

There is also risk in alerting individuals of their own mental health status – a piece of information that is inferred algorithmically from passively shared social media data. Are we doing more harm than good by making individuals who are not in a research study aware that they might be suffering from depression or anxiety, thereby alerting them that we have gathered and analyzed their (public) data? These concerns are also connected to issues of managing false positives and false negatives as an important performance tradeoff.

Bad Actors and Fairness/Discrimination. Another issue involves misuse of algorithmic inferences beyond the interests of the individuals themselves by other actors. In one case, the actor has benevolent intentions but misuse the data, or violate the context of what data was gathered. Samaritan's Radar had good intentions of decreasing suicidality, but was poorly received because it enabled other actors to harass or stalk those when they were at their most vulnerable [193]. This can also be seen in automatic screening and text processing systems, like advertising recommendations, which could scan Twitter posts for self-reported diagnoses of mental disorders [58, 17] and send advertisements for prescription drugs. Is this a desirable outcome?

However, researchers have also identified the risks of malintentioned actors using and reproducing the findings in these papers for unsavory purposes [87, 59]. One example could be the use of this research by health insurance agencies to deny coverage for medical care or raise premiums if an individual is detected as "having" postpartum depression yet never sought treatment. Other applications of these algorithms to other prediction systems, like determining credit worthiness for loans or ability to maintain employment status, are possible. In some countries, these predictions are illegal because mental health is a protected class; however, in other cases, this information is not safeguarded or cleverly designed proxy variables can be engineered to get this information. Can researchers in this space safeguard against bad actors or mitigate these risks?

A related result of these algorithms is discriminatory output – it is possible that the algorithms have a strong sampling bias towards certain groups of people, independent of their mental health status. As mentioned above, social media researchers may be sampling for younger and possibly more affluent audiences by sampling from certain social media data [206]. In their paper about postpartum depression, De Choudhury *et al.* note that they over-sample Caucasian, affluent women for their data collection and interviews [59], which makes generalizability of this algorithm to other demographics challenging. If we extrapolate our algorithms to these groups, how will we manage unintended biases that might lead to negative and discriminatory repercussions? What impact does this sampling have on predicting on different groups of people, such as those with lower socioeconomic status who do not use social media sites, or older adults with lower rates of social media adoption? Do these algorithms only help the proverbial "rich get richer" by predicting mental health status on groups already likely to seek treatment?

6.5 Summary

Social media provides a unique perspective into individuals behaviors and moods. In this research, I discussed emerging research in using social media data to predict an individuals mental health state. I covered the state-of-the-art in the field and discussed three areas of ethical tension: consent and individuals' involvement with the research; methods and study design challenges; and implications of this research. I hope that interdisciplinary researchers act on these ideas, and begin to work on solving these pressing challenges in methods, ethics, privacy, and consent.

CHAPTER 7
SYSTEMATIC LITERATURE REVIEW

The taxonomy I presented in the last chapter of emergent challenges is one step towards coming to consensus on practices, methods, and ethics in the broader realm of predicting mental health status from social media data. In addition to proposing gaps and potential ways to ameliorate them, it is also important to explore the emergent practices in the scientific research space.

The use of predictive techniques like machine learning and regression modeling to predict mental health status is a nascent area of research and has highly variable methods and standards across areas. Caused in part by the interdisciplinary intersection of data science, machine learning, psychology, and human-centered computing, unanswered questions emerge around the role of the individual in predictions and the methods and ethics decisions to conduct this work. Prior work has explored the broad ethical issues of conducting public health research on social media data [208, 213, 237]. Only one has looked at methods [238]; yet these study does not assess *the entire research process*, its process in various fields of computer science, and the roles and implications of this work on the "research subjects" within this work.

One way to explore these questions is through an empirical, systematic literature review (SLR) of research within this area. Systematic literature reviews are a crucial practice within science, providing necessary summary of areas of research, identifying gaps in understandings or methods, and offering reflection for many fields. Explicit focus on these concerns can drive critical reflection in a nascent area to identify best practices in HCI and CSCW [239, 240, 241, 242] and related domains [237] and, possibly, support redirection [243]. As this area emerges as a field with extensive academic and popular media attention, it is an opportune time to step back and assess its trajectory.

This chapter focuses on two studies conducted from a systematic literature review (SLR), inspired by some of the questions that my taxonomy raised [34]. I identify a corpus of 55 papers related to predicting mental health and social media between 2013-2017. Then I conduct two distinct research studies:

- First, I critically explore the methods, study design and algorithmic choices of studies in predicting mental health status on social media data. I report on patterns of data annotation and collection, data bias management, pre-processing and feature selection, and model selection and validation. This paper is under review for CSCW 2019.

- Second, I examine the conceptualizations and representations of the humans who provide data and are the beneficiaries of this work. I ask **"who is the human in human-centered machine learning?"** to explore these representations and framings of human research participants in a new interdisciplinary space.

In this chapter, I briefly describe my methods for gathering a corpus of literature for the SLR. Then, I describe these two studies, focusing on the methods then the results of the two works. After each study, I briefly provide some implications of the work.

7.1 Gathering a Corpus for an SLR

I adopted the tools of a SLR to gather studies in my area of interest: predicting mental health status on social media data. A literature review provides insights into the models of representation in scientific research. My overall approach was informed by standards for literature and meta reviews [244] as well as others in HCI and CSCW [245, 241, 242].

Constructing a corpus across disciplinary boundaries in CS is difficult. I could not use a single professional organization's search database (ACM or IEEE); however, most scholarly indexing services, like Web of Science or Scopus, do not consistently index CS conference proceedings. When I tested the initial search strategy through these services,

Table 7.1: Keywords for literature search

Category	**Keywords**
Mental health (1)	mental health, mental disorder, mental wellness, suicide, psychosis, stress depression, anxiety, obsessive compulsive disorder, post traumatic stress disorder, bipolar disorder, eating disorder, anorexia, bulimia, schizophrenia, borderline personality disorder
Social media (2)	social media, social network, social networking site, sns, facebook, twitter instagram, forum
Search term	**(1) AND (2)**

journal entries were robustly indexed; yet, there were large gaps in the coverage of conference proceedings known to be important in these subfields (*e.g.*, AAAI, ACL, CHI, NIPS/NeurIPS, DH, AMIA). Initial experiments with keyword searches through engines like Google Scholar yielded an intractable number of results (over 200,000 candidate papers before deduplication).

For my approach, I iteratively generated a list of 41 venues (both conferences and journals) that "seeded" the search. I then used keywords to filter in each of these venues, then identified candidate papers through this list. I then sampled the references of these papers once to identify missing papers from the first pass. This produced 55 papers in total. A summary of my methods is provided below.

7.1.1 Searching the Datasets

First, I searched the literature in May 2018 for articles published between 2008 and 2017, dovetailing with the emergence of academic research on social media [246].

I developed two sets of keywords to search in pair-wise fashion: those for mental health and those for social media. These were inspired from meta-reviews on social media and mental health [238, 247] and my expertise in the area. I experimented with other social networks (Reddit, Sina Weibo), but found that these keywords added no additional coverage. A list of keywords can be found in Table 7.1.

Next, I searched for these keywords across 41 English venues in the interdisciplinary intersection of prediction of mental health through social media. These were inspired, again, by expertise in the field as well as from the results of previous literature reviews in the space [238, 247]. This includes venues across professional societies (ACM, IEEE), the Association for Computational Linguistics (ACL), independent conferences (*e.g.*, NeurIPS/NIPS, AMIA), and journals. These are displayed in Table 7.2.

Table 7.2: The venues to identify documents related to mental health and social media research

Topic Area of Interest	**Conferences and Journals**
General Interest	Science, Nature, PLoS One, PNAS
Data Science and Data Mining	KDD, WebSci, WSDM, HT, WWW, MM, TOKDD, TWEB, EPJ Data Science
Health, Medicine, & Health Informatics	JAMA, DH, AMIA, PervasiveHealth, bmj, JMIR, JMIR Mental Health
HCI and Social Computing	CHI, CSCW/ PACM HCI, GROUP, ASONAM, SocInfo, TOCHI, ICHI
Natural Language Processing	ACL, EACL, NAACL, EMNLP, CLPsych
Machine Learning & Computer Vision	NIPS/NeurIPS, CVPR, ECCV, ICML, ICCV
Artificial Intelligence	AAAI, IJCAI
Other	ICWSM, UbiComp/IMWUT

I used three search engines to ensure robust coverage across these venues (see above motivations). I used the ACM Digital Library for ACM journals and conferences, Google Scholar using the Publish or Perish software [248] for other conference publications, and Web of Science for journals[1]. One venue (CLPsych) was not indexed correctly by any search engine, so I manually searched the proceedings for matching keywords in the title and abstract. I identified 4,420 manuscripts that matched these keyword pairs in these publication venues.

7.1.2 Filtering Strategy

I first filtered the manuscripts to include peer-reviewed, full-scale archival studies published between 2008 and 2017, deduplicating entries as I went. I honor the home community

[1] www.webofknowledge.com

standards to assess archival status[2] Additionally, I included studies that conduct full-scale research as primary sources. This removes meta reviews and literature reviews, news reports, case studies, panel proposals, and shared tasks. After deduplication and filtering, this produced 2344 manuscripts.

Next, I manually filtered by title and abstract, removing spurious items obviously not related to mental health or social media data. Examples of mismatches included other health conditions, such as cancer or diabetes, or data sources such as electronic health records. This reduced my corpus from 2344 to 87 papers. Finally, I read and fully screened all 87 papers, using the following criteria for inclusion in my analysis of the corpus:

1. They must address mental health in clinically specific ways. This meant studying a mood or psychosocial disorder (*e.g.*, depression, anxiety, schizophrenia), symptomatology from the DSM-V [8] about disorders (*e.g.*, suicidality, psychosis), or the severity of mental disorders (*e.g.*, moderate vs. severe depression). I excluded subjective mood, well-being, happiness, or general emotions not directly related to mental disorder diagnosis. I also excluded papers about mental disorders and conditions that are not mood or psychosocially-oriented (*e.g.*, ADHD, autism spectrum disorder) [8].

2. The paper's method must focus on quantitative prediction through ML techniques from social media data. This included regression analysis, machine learning, and time series analysis.

3. The paper must study social media data from social networking sites, blogs, or forums. I excluded other digital data traces, such as search engines or app use (if not related to social media apps).

4. Finally, the prediction must be made on an individual. I excluded papers that made

[2] In CHI, workshop proceedings are not considered archived; however, in ACL, workshop proceedings that appear inside the ACL Anthology are archived.

predictions on groups or communities. If a paper made predictions on individuals later aggregated for another purpose, I included these.

This process generated 44 papers that matched all of the constraints. Finally, I conducted an iterative pass, sampling related papers to the 44 identified from the bibliographic details of the citations. I then undertook the same screening and filtering process above. I identified 519 candidates; after deduplication, this produced 253 unique papers. After filtering for date, year, and archive status, there were 200 left. After screening the title and abstract and deduplicating these citations against the 44 entries, there were 20 unique papers. Finally, after a full paper screen, I identified 11 new papers for analysis. Additional snowballs through these papers did not return substantially new results.

This produced 55 papers (44 from initial analysis + 11 from iteration) included in this analysis. The full list of all 55 papers is provided in the Appendix.

7.2 Analysis of Methods

In my first study, I critically examined the practices and strategies for conducting such work. This project is under preparation for submission to npj digital medicine.

7.2.1 Motivations

Grounded research describing appropriate techniques for building algorithms to predict mental health status in social media data has been lacking, as noted in recent research [34, 249]. Given the nascence of this field, I see incredible value in identifying trends in the research methods, identifying important gaps before they systemically emerge. These issues are important not only as they reflect scholarly research quality, but more importantly the implications predicting mental health status can have on individuals who are the object of such predictions in clinical care, social media settings, and more broadly society.

7.2.2 Analysis Technique

I developed *a priori* a rubric for analyzing the manuscripts that included both descriptive, quantitative, and qualitative criteria, influenced by prior work [206, 238, 247, 244] and my expertise in this space. This rubric had 75 items, including data collection methods and pre-processing strategies, accuracy and baseline thresholds, results reporting mechanisms, and the presence of commentary on certain study design choices and implications of the research. Before beginning the analysis, I randomly selected four manuscripts of the corpus to annotate, adjusting the rubric for additional reporting categories based on the results of the trial annotation. The entire rubric can be found in Supplementary Materials.

I then conducted a close reading of all 55 papers in the corpus, annotating the elements of the rubric. I also recorded qualitative notes for analytical insights and thematic observations while conducting the close readings. The entire dataset was read and coded twice to standardize the coding process, each time in a random order.

After the close reading, I used inductive thematic analysis of the annotated dataset to identify relevant themes, gaps, and other information from the dataset [250]. The first high-level theme I noticed was interesting gaps in the practices, study design, and methods to conduct this research, which constitutes this first study.

7.2.3 Descriptive Overview of Corpus

Table 7.3 shows the years of activity in publication. The first papers in this field were published in 2013, with 8 papers in total [16, 251, 252, 59, 76, 253, 251, 70, 254]. In 2017, the most recent year of analysis, this work is growing at 17 papers in total [255, 256, 230, 257, 258, 194, 87, 259, 234, 234, 78, 86, 260, 85, 77, 261, 88, 230]. [3]

I also identified the social media platforms of interest in these studies. The most popular social media site for this analysis was Twitter, with more than half (28/55) of the

[3] I verified that no errors were made in calculating 2016 data by double-checking the raw source data from the searches as well as the manual curation procedure. Many papers around social media and mental health were published in 2016, though most did not incorporate predictive aspects into their work.

Table 7.3: Years included in corpus

2013	2014	2015	2016	2017
8	10	13	7	17

corpus studying this site, *e.g.*, [262, 263]. Other popular sites include Sina Weibo (9) [83, 264, 265, 194, 84, 230, 266, 82, 267], Reddit (4) [88, 259, 81, 77], Tumblr (3) [80, 86, 31], Facebook (3) [73, 71, 70], and Instagram (3) [85, 29, 258]. Other sites with a single study include Flickr [230], PTT [251], mixi [252], LiveJournal [268], and TOBYO Toshoshitsu [72]. Year-over-year, Twitter was the dominant social media site examined in the corpus.

In Table 7.4, I show the representation of languages in the corpus. The majority of studies are done on English data (38) (*e.g.*, [268]), followed by Chinese (10) [267, 82, 266, 251, 76, 84, 83, 265, 194, 264], Japanese (4) [252, 72, 253, 75], Spanish and Portuguese (1) [269], and two that were not easily identified [261, 258].

Table 7.4: Languages present in the corpus

English	Chinese	Japanese	Spanish/Portuguese	Not described
38	10	4	1	2

Disorders and Symptomatology

I examined the disorders and symptomatology in each of the 55 papers. 8 papers study more than 1 condition in the paper [58, 270, 257, 258, 232, 256, 88, 262], so I report the counts the unique disorders, symptomatology, and other disorders examined in Table 7.5.

Nearly half (26/55) of papers examine depression. Examples of this includes study of depression generally [251, 72, 271, 58], major depressive disorder [16], postpartum depression [71, 59], degree or severity of depression [73], and depression as a symptom for other mental health risks, like suicidality [194].

I also find that 17 papers examine suicidality [252, 195, 264, 194, 265, 84, 81, 256, 272, 263, 273, 274, 268, 82]. Examples of this includes examining whether someone has suicidal ideation/is suicidal [252], will attempt suicide [195, 264, 256] or may shift to suicidal ideation in the future [81], risk factors for suicide [272, 194], and distinguishing between suicidal ideation and other discussions of suicide [274].

7 studies examine eating disorders [257, 258, 80, 31, 230, 269, 232], in the general case [230, 269, 232, 258, 257] as well as specifically focusing on anorexia [80, 31]. 7 studies examine schizophrenia. [78, 79, 275, 232, 257, 256, 88]. Other disorders/conditions examined in the corpus include bipolar disorder (6) [262, 232, 58, 257, 270, 88], anxiety (5) [77, 257, 232, 256, 88], post-traumatic stress disorder (PTSD) (5) [17, 58, 232, 257, 270], borderline personality disorder (2) [232, 262], and panic disorder (1) [257]

I also found that papers also examined symptomatology related to mental disorders. This primarily focused on predicting stress (5/55) [266, 267, 83, 259, 194], self harm (2) [89, 88], panic attacks [256], cognitive distortions [86], mood instability [87], and mental illness severity [29].

Table 7.5: Counts of Disorders and Symptomatology Studied in the Corpus. Note that, because 8 papers study more than 1 condition, this sums to more than 55.

Disorder or Symptomatology	**Count**
Depression	26
Suicide	17
Eating Disorder	7
Schizophrenia	7
Anxiety	5
Stress (symp)	5
PTSD	5
Bipolar Disorder	4
Borderline Personality Disorder	2
Panic Attack Disorder	1
Self-harm (symp)	2
Panic attacks (1) (symp)	1
Seasonal Affective Disorder	1
Cognitive distortions (symp)	1
Mood instability (symp)	1
Mental illness severity (symp)	1

Results of Literature Review

Most studies (35/55) examine the individual/user as the prediction task, such as predicting suicide risk of a person [194] or if someone is depressed [16]. 17 studies predicted mental health status per post/aggregated posts [89, 29, 258, 259, 84, 80, 86, 77, 85, 269, 272, 88, 273, 274, 268, 254, 267], such as detecting anxiety [261], depression [268] and then, by proxy, inferring the mental health status of the owner of those accounts. One paper examined both [234]. There was no distinction on what condition or status was examined at a post versus a user level.

In the corpus, almost all papers (50) conceptualized their problems as a classification problem, primarily through binary classification (47/50), such as predicting whether someone suffers from depression or not [72], and the distinction between high and low stress [87]. 3 papers use multi-class schema instead of binary classification [274, 29, 88]. 4 papers use a model that predicts continuous or discrete values [265, 73, 253, 258].

7.2.4 Construct Validity and Establishing Truth

How does the state-of-the-art identify positive and negative examples of what is considered mental health status? In this section, I examine questions of construct validity, or how the studies in the literature review validate that they indeed are examining mental health status without direct clinical appraisal.

Establishing Ground Truth for Positive Annotation

I identified 6 methods of annotation for the positive sign of mental health status.

- **Community or Network Affiliations (17).** Researchers look for community and network participation [252, 89, 79, 29, 258, 259, 80, 86, 81, 275, 77, 251, 88, 31, 230, 268, 262]. Community participation is used as signal in social networks with formal communities, such as participating in communities about mental health on LiveJournal [268], Reddit [259, 77, 88], or posting in a suicide crisis community/forum [81]. These measures also include network signals such as following another account on Twitter [275, 262]. Other studies use the signal of hashtags on social media with amorphous community structures on Instagram [29, 258] or Flickr [89].

- **Keyword Use (8).** Another approach uses the presence of keywords or phrases [72, 255, 234, 269, 251, 273, 266, 262]. Researchers use dictionaries connecting to suicide [273] or stress [266]. They also used symptom words and names of disorders on Twitter posts or profiles [269, 234], behaviors associated with disorders (like "ultimate goal weight," [230]) or if they use phrases associated with life events (like childbirth) [59].

- **Self-Disclosure (15).** Self-disclosure looks for individuals to state that they suffer from a specific condition, illness, or are engaging in behaviors indicative of mental health status [58, 195, 59, 255, 83, 78, 232, 275, 256, 261, 71, 87, 17, 266, 267]. These are triangulated with specific expressions, like "I was diagnosed with..." [58, 17]. Positive annotation includes stating that have a specific illness, like depression [261, 72], PTSD [58], or schizophrenia [78]. Work also looks for reports of anti-depressant medication usage [255], attempts to take their own life [195], or self-describes as being stressed or relaxed [83].

- **Administering Screening Questionnaires (11).** Another popular technique for validating mental health status is by administering screening tools and questionnaires to consenting participants [194, 75, 16, 265, 253, 260, 85, 73, 263, 70, 82]. These includes various screeners that can measure depression, including CES-D [85, 254] and/or BDI [16, 75, 70], PHQ-9 [71], and Zung's SDS [253]. This is also used for other mental health status, such as suicidality [82, 263].

- **Human Assessments (26).** Many papers bring humans to annotate the presence of mental health conditions. Domain experts, such as practicing clinicians or psychologists, are often asked to annotate or label data [258, 84] – one study assessed depression through clinical interviews [76]. Other research uses CS researchers to conduct the annotations, to annotate if someone suffers from a genuine mental health condition [257, 234]. Often, both domain experts and CS researchers partner to do the annotations [78, 272]. Finally, some researchers use crowdworkers from sites such as Amazon Mechanical Turk to identify status [274] or verify the veracity of status downstream after another protocol has been used [59].

- **Acquired Annotations (3).** Several papers acquire annotations from previously published research [270, 271, 257].

- **News Reports (2).** Two studies looked at news reports of who had died by suicide to identify victim's names, then find social media data on these individuals [84, 264].

Some papers (21/55) take the results of the initial proxy assessments at face value [252, 264, 72, 259, 194, 265, 83, 76, 253, 77, 85, 73, 272, 230, 88, 87, 274, 268, 254, 82, 267], such as identifying who suffers from anxiety by their posting in an anxiety community [77] and other mental health subreddits on Reddit [88].

However, most studies (32/55) combine two approaches listed above to acquire a precise sample. Human annotation is a popular follow-up assessment technique, with the validity of initial results of keyword matching screened by humans others [273, 262]. Other approaches use human verification to ensure that self-disclosure was genuine, and not flippant, funny, or otherwise not related [79, 234, 78]. Two papers mesh together three ground truth assessments [195, 263]. There was no reflection in the dataset on what ground truth approach was appropriate for establishing construct validity, nor how many approaches combined together would be appropriate.

Source of Control Data/Negative Examples

Papers also source and design negative/control data for predictive tasks.

- **Random Selection of Control Users (20).** Many studies construct a negative/control dataset from randomly sampled data on the social media platform [252, 89, 58, 79, 257, 195, 264, 72, 259, 255, 78, 84, 80, 232, 275, 77, 256, 261, 230, 17]. This random sampling can come from historical samples of data, like the YFCC100m (Yahoo! Flickr Creative Commons 100 million) dataset [89] or other collections [58]. Others gather randomly throughout the whole platform, like from Twitter's streaming data [261], random Tumblr users [80], or the front page of Reddit [259].

- **Matching Strategies (6).** These studies took randomly sampled users and constructed matched samples along demographic/behavioral characteristics to characteristics or traits of the positively identified users [79, 195, 232, 275, 256, 230]. This includes matching on inferred traits, like age and gender [195, 232, 230], or time-matching controls [256].

- **Lack of Mental Health Disclosure (15).** These studies use a lack of disclosure of mental health status as source for negative data [252, 257, 264, 258, 72, 234, 81, 275, 260, 272, 261, 251, 88, 268, 262]. This may mean sampling people who do not meet or disclose having a condition [262, 275] or not participating in any communities related to mental health [251, 268].

- **Validated No Mental Health Status (24).** Many studies engineered ways to validate that the negative dataset did not contain people with the mental health status of interest [29, 59, 194, 75, 16, 265, 83, 78, 76, 86, 85, 269, 73, 263, 71, 31, 87, 273, 70, 266, 274, 254, 82, 267]. This often was taking the lower bounds of cutoff from screening participants with screeners [75, 70]. Other approaches used an expert to validate that there was an absence of the mental health status in question, such as schizophrenia [78], or took ground-truth statements of relaxed status to signify "not stressed." [83]

Data Source

Finally, I noticed three methods of sourcing data for research.

Gathered Data. Most of the papers (32) gathered the data themselves with no interactions with the subjects/individuals of analysis. 24 papers gathered data through public application programming interfaces (APIs) [275, 255, 86, 79, 270, 195, 58, 264, 59, 78, 261, 232, 234, 274, 230, 269, 31, 273, 262, 266, 258, 83, 29], searching for individuals who post on social media through keywords, hashtags, or profile information [195, 264, 31]. 7 papers used community structures as the data source for their analysis [81, 80, 259, 88, 268, 251, 72], including LiveJournal communities [268], subreddits [259, 88], or forums [251]. Finally, one paper mentioned gathered data from a blog portal under the topic of "depression." [72]

Solicited Data. Next, I identified 18 papers that sourced some portion of their data from direct contact with or identifying specific individuals to search for [253, 16, 194, 71, 260, 85, 263, 76, 253, 265, 267, 254, 82, 87, 70, 84, 264, 79]. The most common technique (13 studies) solicits data from participants by screening them for their mental health status and asking them to allow access to their social media data [253, 16, 194, 71, 260, 85, 263, 76, 253, 265, 267, 254, 82]. 2 papers used more involved interventions in participant behavior, capturing data over time through mobile phone apps about mental health [87, 70]. Two papers identified people who had died by suicide through news articles, and find their social media profiles for analysis [84, 264]. One study solicited data through an initiative called ourdatahelps.org, where individuals knowingly donate data to be studied for mental health and well-being [79].

Acquired Data/Archival Data. 4 papers acquired the datasets from other published research projects [257, 272, 73, 271], using both the dataset itself as well as provided annotations. Two papers acquired their datasets from corporate sources without annotations [89, 252]. One study used both acquired data as well as sourced new data from Twitter [17].

7.2.5 Managing Data Quality

41/55 papers include filtering to the dataset to manage issues of data bias or quality, and some studies used multiple approaches to filter data.

- **Platform Behavior Thresholds. (15)** Researchers described removing data for not meeting minimum content or engagement thresholds [252, 194, 75, 265, 232, 85, 73, 268, 262, 16, 71, 275, 274, 260, 17]. This included behaviors such as having an account on the site of interest [71, 16]. Primarily, these studies had minimum activity thresholds threshold, like the number of posts [262, 58]. Others looked for a minimum number of friends/relationships [252] or platform engagement in a certain time period [256, 252].

- **Legitimate Mental Health Mentions (17)**. These studies took additional steps to validate disclosures of mental health status or assembled more rigorous characteristics after the dataset was assembled [58, 79, 29, 195, 258, 81, 269, 272, 71, 273, 274, 254, 230, 259, 31, 230, 275] Some had strict thresholds on the precision of positive mental health status [269, 230] or the time frame in which certain behaviors could occur [81]. For instance, one study looked for suicide attempts with discernable dates [195]. Others removed participants for participating in eating disorder recovery communities [29, 31].

- **Restriction on Participant Characteristics (11).** These studies excluded individuals based on certain characteristics or traits. [82, 194, 267, 256, 87, 273, 70, 260, 85, 263, 16], such as age [194, 267]. Other studies filtered participants on crowdworking sites based on overall approval ratings or a minimum number of previous tasks completed [260, 85].

- **Quality Control During Online Surveys (6).** One threshold was removing participants for not passing quality control measures on the surveys, especially on surveys given through crowdworking sites such as Amazon Mechanical Turk or Crowdflower [254, 265, 85, 263, 70, 82]. This includes filtering surveys completed too fast [254, 265], who did not pass attention checks during the survey [85, 263], or did not finish the survey [70, 82].

- **Removing Spurious Data (6)**. Other studies removed spurious data [72, 263, 275, 194, 273, 82], such as duplicate survey responses [194] or gibberish [273]. One study mentioned removing advertisements [72], and 2 removed spam [72, 275].

I did not notice any larger dataset adjustments to account for other kinds of biases, as noted by Olteanu *et al.* [206]. I looked for whether studies adjusted for sampling bias issues with limited access APIs, adjusted for other clinically-relevant signals (such as demographics), took alternative data sampling strategies (such as selective rather than random sampling), or removed adversarial content, bots, outlier large accounts (such as organizations or celebrities). Other than matching samples and 2 papers that removed spam and advertisements [72, 275], I did not notice any corrections in the dataset for these factors.

7.2.6 Feature Engineering

I present the results of how researchers engineer the features/variables in the prediction task. 31/55 studies reported the number of features. Of those 31 papers, the range of the number of features was 7 [252, 29], to over 15,000 [81]. I identified 6 categories of feature types:

- **Language Features.** (50/55)
  - **Structural/Syntactic.** (21) I found features that describe the structural or syntactic composition of social media posts [89, 72, 194, 75, 16, 255, 83, 234, 84, 80, 76, 81, 275, 260, 73, 71, 31, 87, 274, 82], like the length of the post/chunk of interest [194, 81], ratios of part-of-speech tagging [274], and modality tagging

in other languages [72]. I also saw counts of specific characters, like emoticons [275]. One study used the length and number of numeric characters in the domain name of a blogging site [82].

- **Character and Word Models.** (24) These representations of language draw on probabilistic representations of character and word patterns within text [89, 58, 79, 257, 195, 264, 258, 259, 75, 78, 84, 232, 253, 77, 269, 73, 251, 88, 17, 273, 274, 254, 262]. This includes word n-gram use [272], character modeling [264], bag of words models [75], term-frequency-inverse document frequency [251], and word embeddings [258].
- **Topical.** (9) Other papers use topic modeling to identify meaningful connections between concepts in datasets [58, 270, 75, 265, 84, 73, 272, 261, 271]. This includes the popular Latent Dirichlet Allocation (LDA) topic model [270, 271], and Brown clustering [79].
- **Linguistic Style.** (15). Linguistic style and content was measured [89, 58, 79, 59, 16, 234, 78, 80, 81, 260, 71, 87, 268, 254, 267]. This used style categories from the Linguistic Inquiry and Word Count (LIWC) dictionaries [268, 254]. I also noticed the examination of readability, coherence, and perplexity measures [79, 234].
- **Domain-Specific.** (11). This approach designed domain-specific linguistic features to evaluate in text documents [89, 58, 72, 16, 234, 81, 275, 261, 266, 274, 268]. This includes making dictionaries or lexicons related to depression [234, 82], self-harm [89], suicide [274], and stress [266]. This also includes assessing user-generated mood tags taken from LiveJournal [268] as well as explicit mentions of taking certain kinds of medication [16, 77].
- **General Language Measures.** (10) Papers also described generic language measures [79, 270, 194, 265, 86, 77, 73, 263, 230, 274], using the LIWC library

in its entirety.

- **Behavior.** (29/55)
  - **Activity.** (27) Some features examine behavioral activity of the individual of interest [252, 89, 58, 195, 264, 59, 75, 16, 255, 83, 234, 84, 80, 76, 81, 275, 260, 85, 261, 251, 71, 31, 230, 70, 254, 82, 262]. Posting frequencies are a source of interest [195], including volume of posts [81], rates of posting [262], and temporal distributions of posting history [251]. Studies also examined platform-specific features, like geo-tagged posts [70] and Twitter favorites [234].
  - **Interaction.** (25) Interactions with others on the platform were another common feature source [252, 89, 58, 59, 75, 16, 255, 83, 234, 84, 80, 76, 81, 275, 260, 85, 261, 71, 230, 70, 274, 254, 82, 267, 262]. This included number of connections, including uni-directional follower/followee relationships [275, 261] and bi-directional friendships [70]. Papers also examined community membership/affiliation or participation [77, 230], platform affordances like Twitter's retweet/quote or mentions/replies features [262]; Facebook's friend requests [70] and posts/shares to others [71].
  - **Network**. (4) Several papers examined the network or graph structures [16, 255, 252, 267], including graph density [16, 255], clustering coefficients and homophily [252], and network size and depth [267, 16].
  - **Domain-Specific.** (5) In addition to activity behaviors, studies also engineered domain-specific features of activity [82, 58, 252, 16, 76]. This focused on measuring posting between the night hours, quantified as the "insomnia index." [16] Another paper examined suicide homophily, or the number of friends who had died by suicide [252].
- **Emotion and Cognition.** (28/55)

  - **Sentiment, Affect, and Valence.** (26) Many papers examined peoples' expressed moods, sentiment, and intensity of emotion [89, 58, 270, 195, 264, 59, 259, 75, 16, 255, 83, 234, 78, 80, 80, 76, 260, 256, 76, 71, 31, 268, 254, 82, 267, 262]. This was measured with sentiment scoring mechanisms like ANEW [268], LIWC [71, 31], and VADER [256]. Other studies examined affect and intensity [59], or counted the number of positive and negative emoticons [267, 230].
  - **Psycholinguistic.** (9) Researchers also use other psycholinguistic evaluations of emotional status [58, 83, 78, 80, 260, 31, 87, 268, 82]. This includes using certain categories of emotional speech (such as *anger* or *anxiety* in LIWC) [58, 268], common topics around stressors in daily life [83],
  - **Domain-specific.** (3) Domain-specific applications of the emotion and cognition measurements include measuring personality traits via Big 5 [270], categories related to behavior theories of anorexia recovery [31], and other Tweets annotated to be related to depression [234].

- **Demographic Features.** (9) Papers also incorporated data about personal demographics [270, 257, 258, 16, 71, 70, 82, 262]. This includes age and gender [82, 262], and other factors like education, income, and/or relationship status [16, 261]. Some of these were not reported or gathered from individuals in the dataset; rather, they were inferred via computational means [261, 270].

- **Image Features.** (5) Researchers extract information from the image data of the post [83, 85, 261, 89, 258]. This includes color themes/Hue-Saturation-Value (HSV) values [83, 85], if the image includes a face [85], brightness and saturation values [261], and the types of colors used [261]. This also includes data extracted from a convolutional analysis of the images [89, 258].

For feature reduction or selection techniques, 20/55 describe reducing features to salient ones [72, 89, 79, 59, 194, 259, 75, 16, 255, 234, 78, 86, 253, 275, 269, 73, 261, 274, 268, 254]. The most popular feature reduction technique was dimensionality reduction in the form of Principal Component Analysis [73, 275]. Other feature selection methods include experimentally removing features [234], feature ablation [79], stepwise regression [194], and taking the k-best features [78].

7.2.7 Algorithm Selection

All but one paper (54/55) paper report what algorithm they used to construct a predictive model of mental health status.

There is high diversity in the dataset about the selection of the algorithms. The most popular predictive algorithm is Support Vector Machines, used by 19 papers as their algorithm of choice [80, 79, 264, 59, 72, 194, 259, 75, 16, 234, 84, 80, 275, 272, 251, 87, 248, 230, 273, 254, 271]10 papers used logistic regression as their predictive algorithm of choice [252, 82, 269, 29, 195, 86, 81, 71, 268, 82, 267]. Next was Random Forest at 7 papers in the corpus [262, 78, 82, 260, 85, 256, 274], which included one paper who used a Rotation Forest (a boosted version of Random Forest) [274]. I also see the use of decision trees (2) [255, 263], and Naïve Bayes (2) [76, 269].

Deep learning has been a recent trend in the corpus, with 7 papers using this technique [89, 257, 258, 83, 77, 261, 88]. Some papers used a more straightforward deep neural network [230, 83, 77] or convolutional neural networks [88], whereas others focused on a multitask neural network to share information between prediction tasks [257, 266].

The other kind of technique used were regression techniques (6) [17, 70, 265, 73, 17, 31]. This includes the use of linear regressions [265, 73], log-linear regression [17, 58], and survival analysis/Cox regression [31].

How were these algorithms selected for use? 30/55 papers described their process for selecting their algorithm of choice. The vast majority (22/30) were selected because they performed the best [77, 266, 252, 264, 59, 16, 234, 78, 86, 275, 260, 77, 85, 269, 73, 263, 261, 88, 87, 230, 266, 254, 262], experimentally chosen across several algorithmic options [234, 254]. Another performance related concern in addition maximizing F1 or accuracy was the trade-off between precision and recall [271]. Some other reasons for selecting the algorithm were the suitability of the model to the task of interest, such as sharing knowledge between multiple tasks [257], interpretable features for clinicians or other stakeholders [263], or issues of dropout impacting the use of standard regression techniques [31]. Others drew from theoretical and practical reasons to select their models [274], such as the "no free lunch theorem." [86]

7.2.8 Validating Algorithms and Reporting Performance

54/55 papers reported how they validated the models. The most popular method of validating model performance is using k-fold cross validation. 40 papers use this technique, with a k ranging from 5 [87], 10 [269], to leave-one-out [194, 263]. Another common technique (13/55) was holding out blind data as a test set and reporting performance [89, 29, 195, 234, 78, 81, 275, 73, 272, 88, 273, 271, 267]; heldout dataset size ranged from 10% [273] to 30-40% [265, 89]. The next most common method to manage model validation is multiple experimental runs of the model (10/55) [59, 75, 16, 83, 80, 260, 85, 261, 254, 82], ranging from 5 [260], 10 [261], 100 [75] to 1000 [80]. Other papers (3) used model fit measures to validate the fit of the model, such as deviance measures for regression fit [31, 29, 252] and feature relevance or curation techniques like stepwise regression to prevent overfitting [253, 252].

Many papers took multiple approaches, the most common of which was cross-validating their test data and reporting results on a heldout dataset (*e.g.*, [59, 273]) or pairing cross-validation with multiple experimental runs (*e.g.*, [82, 254]).

Essential Reporting of Prediction Technique Details

Last, I examine reporting of essential information required to reproduce a predictive algorithm, which are *de facto* minimum standards within modeling to evaluate an algorithm. Additional areas of research, such as biomedical informatics [276], require more extensive documentation; however, I focus on 5 elements that are essential to running any regression model or machine learning approach. These are: the number of samples/data points, number of variables/features, the predictive approach of choice (either a specific algorithm, regression type, *etc*), a method for validation, and the metric used to evaluate performance. I then counted the number of papers that explicitly reported on these 5 items:

- 51/55 number of samples/data points
- 31/55 number of variables/features
- 54/55 algorithm or regression of choice
- 54/55 at least one validation method
- 50/55 explicit performance metrics

If each paper is examined for the presence of each of these 5 traits, only 20/55 papers, or 36.3%, report all 5 measures. If we look for 4 of 5 criteria, 41/55 papers, or 74.5%, report on at least 4 of these criteria.

7.2.9 Implications

In this section, I reflect on these results and identify two key areas of improvement that research in this field can make going forward.

Precision of Mental Health Status to Signal

A concerning trend was the imprecise identification and prediction of mental health status throughout the majority of the corpus in the dataset. There is no reflection, critical analysis, or justification given for the methods chosen to predict mental health status.

I see evidence of this trend throughout the research process, beginning at the descriptions of the mental health status of interest. Within many papers of the corpus, it is often unclear what condition is the object of study. One example of this is Shen and Rudzicz [77] – they use Reddit data to identify "anxiety." However, anxiety is a vague and overloaded term; it is a category of nervous disorders (generalized anxiety disorder), symptomatology that can influence other mental disorders, an transient emotion that people experience (anxiety around an exam), and lay usage to referring to emotional states and traits of a person (an anxious person). None of the studies that examine anxiety make this distinction. Similar issues emerge when examining the conceptualization of depression – few studies explicitly acknowledge what kind of depressive disorder they examine. It is assumed that the authors are referring to depression as in major depressive disorder, and not dysthymia or depressed mood. However, the reader is left to assume the precision of the signal that is measured through the study.

These ambiguities emerge again when establishing construct validity, or standards for "gold standard"/ground truth labels for building predictive systems. As discussed in the results of the study, different signals for ground truth have different levels of precision - consider the precision of community participation for depression compared to triangulating depression with the CES-D and BDI screeners, which have been robustly tested for their validity. Papers in this corpus do not reflect on the use of ground truth signals or their ability to measure their characteristics of interest. Recent research has highlighted this gap in schizophrenia prediction [249], and I find it more broadly in the corpus.

Finally, this lack of precision permeates through the experimental design, into the selection criteria for ground truth annotation methods, designing negative/control dataset, and designing and selecting models. Rarely is reflection or justification provided that explain the selection and reduction of variables/features, data bias corrections, or algorithm selection. As noted, there was inconsistency in reporting standards of key details of algorithms to facilitate reproducible science. Only 20 of 55 papers reported explicitly five minimum standards for reproducing these algorithms.

Different mental health statuses require different levels of precision for their identification and measurement. The needs of any given research project are necessarily determined by the specificity at hand – using social media to predict the presence of major depressive disorder would need higher levels of precision than identifying stress. These difference are caused by numerous factors: what mental health status is being identified; whether it is intended to be diagnostic or speculative; the specificity of the status; information available from a given platform; and the intended use of said signals in other practices. As specificity of diagnostic signal increases, the proxy for which to make that assessment becomes more strict.

What are the risks of unclear reporting standards around mental health status, methods, and study design decisions? These unclear reporting standards jeopardize the reproducibility of these studies for future work. For review and evaluation of studies, these omitted details make it difficult for reviewers and readers to understand what was done. For those unfamiliar with the predictive technique, these gaps can imply that undisclosed researcher discretion guided the decision-making process, when in fact there are standards that guide the number of features compared to the number of samples. There is also no ability to benchmark or validate these approaches against work that reproduces the findings – for instance, without clear performance metrics (like R^2, RMSE, F-1, or accuracy), we cannot know if we have improved on state-of-the-art.

I hypothesize a few reasons why reporting these basic components of the algorithms is missing. The interdisciplinarity of the field has not yet normalized reporting standards for machine learning, in part because this research happens across ML, NLP, HCI, and AI conferences. This is amplified by reviewer unevenness for conferences — finding one expert who can tend to machine learning or statistics, mental health, and online communities is challenging. A pool of reviewers may not satisfy the necessary expertise needed to capture these problems. Another reason for these gaps may also be that certain conferences prioritize certain characteristics in reporting, and authors accommodate their information to both those styles as well as page limits.

Being explicit about each component of the methods process, from naming precisely the condition of interest to the motivations for selecting a specific variable reduction technique, makes for better science. Without these details, these studies cannot be carefully reproduced because of the number of assumptions that would need to be made. Further, the results from studies with these gaps cannot generalize to other social networks, and also limit the application of these methods for clinical or interventionist approaches. Finally, these gaps can cause bad or erroneous conclusions to be drawn from the results, which can have downstream impacts on adoption into clinical or moderation practices.

Despite these challenges, there are promising examples of papers found that both address precision of the measured signal and consider the choices of methods throughout the research process. De Choudhury *et al.* hones in on major depressive disorder as the object of study and assess it via administering clinically grounded screeners like CES-D to participants [16]. Burnap *et al.* address this problem head-on by building classifiers to distinguish between six kinds of Tweets about suicide, ranging from those indicating legitimate disclosures of suicidality to awareness campaigns [274]. For images, Wang *et al.* build a classifier that can identify the presence of self-harm photos on Flickr, engineering their approach around a more generalized symptoms [230] – their ground truth dataset annotates for a presented behavior on the platform, using the photos as evidence of sharing self-harm content.

7.3 Representations of the Human in Predicting Mental Health Status with Social Media

7.3.1 Motivations

Interdisciplinary research, such as the human-centered work that I do, focuses on impacts to humans and society, made explicit by its contributions to human-centered domains and challenges [277]. However, this interest causes tensions in researcher responsibilities to the "human subjects" within their datasets [200, 237]. In machine learning and related areas, there are few protocols for managing researcher relationships to data [204, 237] – ML has historically relied on large publicly available benchmark sets (*e.g.*, ImageNet [278]). Now, data comes from sources much "closer" to the human.

In my work, social media provides a large, unobtrusively collected source of data over time about peoples' thoughts, feelings, moods, and experiences. As data-driven research moves closer to paradigms of human subjects research traditionally protected through ethics boards [217], this new proximity complicates the traditional representations of individuals involved in such work from either ML or human subjects research perspectives. These *representations of the human* affect downstream consequences on how research is conducted and reported, as well as on the influence the work can have on society.

The impacts of these representations go beyond abstract notions of roles or responsibilities. Predictions that are central to many questions can both support decision-making [279, 68] yet have impacts on peoples' lives [214]. Representations have meaningful consequences on research methods [280, 281, 282] and practical risks in increasing stigma [283], reproducing stereotypes and discriminatory practices [284, 285], and harming the individuals and communities of interest [286, 287]. Explicit focus on these concerns can drive critical reflection in a nascent area to identify best practices [245, 240, 237] and, possibly, support redirection [288].

In this study, I ask **"who is the human in predicting mental health status on social**

media data?" to explore these representations and framings of human research participants in a new interdisciplinary space. I focus on language as operative in explicating these representations. Language is a driving force in the ways we conceptualize problems, include (or exclude) individuals from analysis, and encourage others to advocate for social change [289, 290]. The discursive representations of personhood can have significant ramifications for whether and how those persons are justly and equitably treated [291, 292, 286, 293].

To conduct this investigation, I use the 55 papers from my SLR and apply thematic discourse analysis [289] to examine how the human as data provider and beneficiary is described within these papers. My results indicate that, perhaps unsurprisingly, the dominant paradigm to refer to the human is the "user." However, I identify a total of five discourses that frame the human in varying, sometimes conflicting ways. Crucially, many of these framings result in a translation [294, 295], constructing the human as a data point for machine training and optimization rather than as a person who should be justly, equitably, and sensitively treated. Further, a single paper will often invoke different discourses, leading to confusions and depersonalization. This paper is under submission at CSCW 2019.

7.3.2 Analysis Technique - Discourse Analysis

To understand how this research conceptualizes responsibilities to humans, I study how the community describes them in publications, or the *discourse*. Foucault famously described discourse as an action of language "that systematically form[s] the objects of which they speak." [290] Discourse frames, shapes, and changes our formations of social and political structures, and how power may be conferred to individuals and groups – with the ultimate goal of making such structures apparent for critique and change [290, 289, 296, 297].

Discourse has been a useful lens to understand language focused on the adoption and use of technology. In HCI, focuses on language and representation have been used to explore lay narratives around robots [298] and smartphones [299]. Hoffmann has explored the pitfalls of anti-discrimination and anti-bias discussions in research and practice [297], and how Facebook's CEO, Mark Zuckerberg, changed his conceptualization of the relationship between Facebook and its users during his tenure [300]. Discourse has been a useful frame for critically considering practices within HCI itself, such as intersectional identities of research participants [301] and the field's construction of sexuality [296].

Driven by these primary research question, I identified research sub-questions to better examine this. These included:

- Who is the human or subject of these predictions represented in the paper?
- How are these subject positions represented? [281]
- Are there notable proxies or substitutes for the notion of the human in the corpus?
- Who are the benefits/implications of this research offered to?

Using inductive coding [289], I conducted a close reading of the entire corpus, annotating the terms and phrases that conceptualized the human "research subject." I focused on the subject of the prediction task, the studies, as well as the purported beneficiaries. I coded at the sentence level for terms in written text, as disambiguating at the word level was too granular to draw larger conclusions. Statements could be simultaneously coded for the presence of multiple terms or concepts. I also wrote notes/memos from the insights gleaned from close readings and thematic corpus-wide observations about patterns in the dataset. I also tracked descriptive variables, such as the venue of publication and year. As the coding progressed, I also decided to not code Literature Reviews and Related Work, as these sections reported on other studies' representations, not the representations of the current authors of the corpus.[4] This analysis was done in Dedoose[5].

After the initial coding was complete, I then met with my collaborators and discussed the emergent themes from this research [250], which I identified as discourses governing

[4] I believe the transmission of these roles throughout literature reviews is ripe for future work, but was outside of the scope of the present study.

[5] https://www.dedoose.com/

the representation of the human within the corpus [289]. I also noticed the ways that rights, responsibilities, and roles were cultivated through the discursive practices of the documents. These observations and understandings form the basis of the analysis presented below.

7.3.3 Results - Discursive Representations of the Human

I identified 164 novel terms that describe who the human is across the corpus. I then grouped terms based on the ways they were used in the documents, which produced five discourses: Disorder/Patient, Social Media, Scientific, Data/Machine Learning, and Person. I clustered terms used in more than one document, as I felt this was more emblematic of patterns in the corpus. I present an overview of these discourses in Table 7.6. In the following sections, I unpack the representations of who the human is in these five discourses.

Human as Patient/Disorder

To begin, I find many conceptualizations of the human as a clinical subject, with an emphasis on their relationships with disorders, doctors, or clinical researchers. This was the most complex category in the dataset.

One of the most common patterns of language use was referring to the human as a "patient." Using measures such as self-reported clinical status, researchers refer to individuals as if they were in an active clinical care relationship. For instance, Nakamura *et al.* analyzed 200 authors of blogs tagged "depression" from a Japanese health blog portal. Throughout the paper, they refer to individuals who write the blogs as "long-term patients" [72] – the title of the paper "Defining patients with depressive disorder by using textual information" reflects this decision. I see this pattern across many documents.

The term "patient" may not accurately reflect the relationships these individuals have with clinical care providers. A "patient" is someone who is actively participating in a health care relationship — no studies in the corpus actively recruited participants through clinical

Table 7.6: High-level thematic categories by the analysis. I included words used in more than 1 paper. Ordered by appearance of the word in unique documents

Discourse	*Terms (number of documents/papers)*
Disorder/Patient	patient (17), depression (10), depressed (9), sufferer (9), behavior (7), condition (4), distressed (4), PTSD (4), neurotypicals (3), non-depressed (3), suicide (3), normal (3), victim (3), clinical (2), anxiety (2), bipolar (2), mentally ill (2), non-stressed (2), pro-anorexic (2), stressed (2), suicidal ideation (2), score (2), standard (2), state (2)
Social Media	user (55), post (25), tweets (16), content (15), account (14), author (14), community (10), microblog (7), text (7), document (6), member (6), activity (4), followers (4), message (3), poster (3), tweeter (3), corpus (3), blog (2), item (2), networks (2), publisher (2), profiles (2), lexicon (2)
Scientific	population (29), control (21), participant (16), subject (10), cohort (8), candidate (6), respondents (6), observation (2), pool (2)
Data/Machine Learning	data (31), sample (25), dataset (18), class (16), example (8), subset (8), test set (5), category (4), positive/negative (3), task (3), data point (2), model (2), prediction (2)
Person	people/person (47), individual (40), she/he (11), woman (7), one's (5), man (5), youth (5), student (5), mother (4), worker (4), crowdworker (4), female (3), someone (3), peers (3), friends (2), others (2), they (2), adolescents (2)
Not Grouped	group (35), case (9), counterpart (2), life/lives (2)

practices or formally verify a relationship with a health care professional around a mental disorder. A few studies verify clinical diagnosis or date of treatment [16, 59], though these studies do not call individuals "patients."

In addition to confusing clinical implications, I also noticed diverse language describing disorder status and the individuals grouped under it. A common pattern is to use the language of the disorder as shorthand for the positively identified group. For example, authors will "assert" that individuals identified through proxy measures actually suffer from that mental disorder, and use that language in the remainder of the paper. In one document, Shen and Rudzicz use participation in r/Anxiety (a subreddit for anxiety in general) to identify the "Anxiety group:"

> "we also find lexicons relating to feelings and first person...represented in the

> *Anxiety* group" [77]

Crucially, anxiety is an overloaded term; it can mean an emotional state or short-term experience, a symptom of other disorders, or the category of anxiety disorders. Therefore, defining participants as the "Anxiety group" may be misguided.

I identified similar patterns in individuals with presumed absence of the mental disorder of interest. Many papers adopted the language of "non-disordered," to contrast a group of "disordered" individuals, like the "non-stressed user." [83]

I also saw discourse framing the individual who did not have a mental disorder as "normal:"

> "We perform an empirical study...of potentially depressed users against a differential control group of *normal users*." [255]

or "neurotypical:"

> "Users who attempt to take their life generate tweet sat a level higher than *neurotypicals*" [195]

Using language like "normalcy" or "neurotypicality" to describe a lack of a mental disorder stigmatizes those who have mental disorders by "othering" them and their experiences. This stigmatizing language can be harmful to individuals [283, 221], and several guidelines written by both journalists and mental health advocacy groups suggest avoiding language that paints the individual as just a mental disorder [221, 283].

Overall, this discourse casts the human as active participants within clinical relationships, implying engagement with clinical partners ("patients,""the depressed") or with language that can be stigmatizing for individuals who suffer from mental disorders.

Human as Social Media User

The 2nd discourse I found was the social media as the mediating actor in these relationships. I begin by looking at the term "user," present in *all documents* in the dataset. It most commonly refers to the "user" in relationship to a social media platform like Twitter: "We extract several features from the activity histories of *Twitter users*." [75].

I saw similar patterns around the more generic term "poster:" "...this distribution was skewed by a smaller number of frequent *posters*" [260], or platform-specific language like "tweeter:"

> "this paper proposes to leverage details of social interactions between *tweeters* and their following friends" [267]

In these contexts, the human was portrayed as an active curator of their social media profiles who generated data or interacted with others on the platform.

In contrast, I also saw representations framing more passive engagement. Many documents described the entire collection of social media data as the object of prediction, or the individual units, using language like "account," "profiles," or "posts" and "content:"

> "All potential control Twitter *accounts* were also manually curated." [275]

Many documents use a single positive identification of mental health status on these passive sources of "post" or the "Tweet," then scale it to the human behind the account. In one paper, the authors describe how they can detect mental disorders in people and in populations, though they use a single post of an account in several mental illness Reddit communities to draw that conclusion about an individual [88]. A single episode of a behavior or symptom, measured through a "post," may not be not enough information to comprehensively identify mental health status. This post-to-human proxy transformation is subtly implied throughout the writing, thought rarely explicitly stated.

To summarize, I found the discourse around social media use to both promote active and passive engagement of the humans. Social media is one insight into well-being and cannot comprehensively represent individuals' thoughts and behaviors; thus, examining it requires a necessary compression of fidelity. However, I found that many papers overcompressed the representation of mental health status of individuals to a single behavior, message, or post on social media.

Human as Scientific Subject

I move to the third discourse, drawing on perspectives of the human as a scientific subject. To begin, one popular representation was the human as “participant.”

In some studies, individuals provide researcher access to their social media, whereas others used apps to measure activity that they linked to social media data [259, 70]. For instance, Guan *et al*. recruited over 900 participants through recruitment messages on Sina Weibo: “All *participants* interested in this survey were asked to log on to the Internet survey system by their Sina Weibo account.” [82]

Scientific language for “participation” has evolved around active consent into research through ethics boards protections [217]. However, “participant” and the closely related “subject” were not always used precisely to refer to human research subjects; in fact, several studies used this language to denote individuals passively gathered from public social media data:

> “...[This] evaluation demonstrates that our system can effectively identify potential*subjects* who are suffering depression but are unaware of it...” [251]

I saw similar confusion with scientific terms like “control,” referring to a group of individuals juxtaposed against the positively identified group.

> “Data...distinguish[es] users with schizophrenia from *healthy controls*” [78]

However, “control” is always juxtaposed against an implied “treatment” position in this corpus — in experimental setups, control groups are verified to not have the effect under observation. A very common practice (seen in more than 15 papers) would draw a random sample of users from the rest of the site and use this as their “control.” Although some studies use screeners with consenting participants to evaluate this lack of treatment [263, 16, 265], most do not validate their control group.

This mathematically guarantees that the "control" is not a true control group, as it will possess those who have the mental disorder of interest, given the occurrence rates of mental disorder in the general population. Coppersmith *et al*. are one of the few who both discuss the problems of contamination and explain their language choice for "control," saying, "...we draw an age- and gender-matched *control* from a large pool of random English users." [58]

I see similar dilution of terms like "population,", which can have scientific meaning as well as statistical relevance. Often, documents would refer to "population," as the whole group of individuals involved in a study, as Saha *et al*. do for studying all content from a campus community: "we proposed computational techniques to assess how the psychological stress of a campus *population* changes following an incident of gun violence." [259]

However, language around "population" would confuse whether the authors had the whole universe of people of interest to the research (all people who participate in a depression community) or a subpopulation of that group (500 people sampled from a depression community). Chancellor *et al*. make this confusion with "user population" to refer to a subpopulation of users identified to evaluate trajectories of anorexia recovery, "after 45.6 months, 50% of the *user population* have not recovered." [31]

I found that scientific discourse is incorporated into these studies in imprecise ways. Many studies will borrow terms from the experimental or human subjects literature ("participant," "subject," "control"), implying experimental rigor and human subjects protections that are not realized through the actual experimental design.

Human as Data Object

The fourth discourse I identified is the human as data object, translating the person into a part of an algorithm or machine learning pipeline.

I begin with the word "example." Here, Benton *et al.* use "example" to references the mathematical number of data points passed to their algorithm, as that is related to their primary contribution, "...we show how to model tasks with a large number of positive *examples* to improve the prediction accuracy of tasks with a small number of positive examples." [257]

Other data terms define the transformation of human to data object, such as "positive"/"negative":

> "*Positive*: the tweet content indicates the presence of one of the studied diseases/states in the person who has written the tweet." [269]

When describing the methods or results, often mathematically precise language carefully identify what data is incorporated into machine learning algorithms. Some studies are aware of this distinction and draw attention to it: "Because we did not interact with our subjects and *the data* is public, we did not seek institutional review board approval." [29], but most do not.

Data discourse is often used ambiguously to describe the contributions of the paper without describing how the data is ascribed to a person. Some examples refer to changes in groups of people, or "classes," despite predicting on individuals:

> "for the depression *class*, we observe considerable decrease in user engagement measures" [16]

These contextualizations can be useful for understanding group behavior, but must be kept in context to the operative research question of predicting on individuals.

In this example, the authors provide research contributions without referring to the individuals:

> "...data mining of online blogs has the potential to detect meaningful *data* for depression studies. The result highlights the potential applicability of machine learning to psychiatric practice and research." [268]

Finally, the situating language in the sentence highlights this division of the human to the data:

> "our work aims to make timely depression detection via harvesting social media *data*" [261]

Rather than collect or gather data, the researchers describe their process as "harvesting."

Extreme care must be taken around proxy language that converts individuals' social media content into data for machine learning observations, as it risks dehumanizing individuals in these documents. These discursive choices can make invisible individual experiences or imply that the research is not actually interested in serving the individuals it argues it helps.

Human as Person

The final category of discourse was focused on language around characteristics and roles of people. One common term from this category of terms was "individual," framing the contributions for them, "A technique to identify symptoms of depression in *individuals* from objective information would hasten recognition of depression..." [75]

I also occasionally see the use of "individual" in the prediction analysis,"...using the same feature set can build a classifier to classify depressed *individuals*..." [255]

I observed similar translational challenges to the social media language with the post-to-individual proxy, where presence of behaviors in posts are implied to affect the individual. In this document, the authors build a decision tree to predict suicidal ideation on posts, that then they imply transfers to the individual:

> "The tree first splits on the "achieve" category of LIWC, such that if an *individual's* usage rate of achievement-related words exceeds 1.46, that *individual* is labeled as nonsuicidal." [263]

Other documents make similar proxy assumptions, arguing that they have correctly identified "people" who are depressed without clinically verified proxy signals and explaining the findings in light of that, "depressed *people* sometimes suffer occupational function impairment, which leads to different mental conditions or behaviors between workday and weekend." [251]

I also saw reference to the roles and demographics of the individuals of interest in the study. Several studies examined the social media behaviors of "students," "teenagers," or "adolescents," and used these terms throughout:

> "36 *high school students* (15 males and 21 females, aged between 15 and 17) in Shaanxi Province, China, participated in the user study." [267]

Another demographic focus was on gender. One instance I saw was use of the term "mother" to refer exclusively to birth mothers who have postpartum mood changes:

> "the total timespan of our dataset is between March 2011 and July 2012, with a total of 36,948 posts from the 376 *mothers* during the prenatal period..." [59]

However, gendered demographic terms were also used imprecisely ("she," "her"), especially in discussions of eating disorders, "If *she* does not recover in 3 years, the probability of remaining anorexic for another 3 years is 0.39/0.56 = 69.6%" [31]. In this case, "she" was the only gendered language used in the paper.

In another document, the author collects a dataset of "females" who do not self-report to have eating disorders as the negative dataset:

> "As ED develop predominantly in young *females*, the effects of demographics can be further controlled in comparing ED and Younger user..." [230]

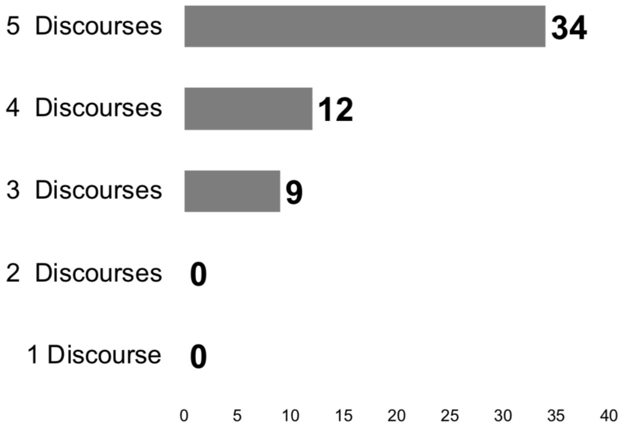
This language choice is concerning because the use of gendered or other demographic language coupled with mental disorders can reproduce stereotypes about who has certain conditions. About a third of those with eating disorders are male[6], and describing eating disorder sufferers as exclusively female is alienating to men with eating disorders. This could be applied to other mental health statuses that may perpetuate stereotypes on who suffers from a mental disorder.

In summary, I believe that the use of person-centered discourse in this paper is complex and may not always be directly tied to the person, involving proxy transformations of data to the person involved or overgeneralizing who suffers from a specific mental health status.

Relationships Between Discourses

Previously, I examined each discourse as an independent unit of analysis. Now I explore the interactions of these patterns within the dataset, as the interplay of these themes reveal larger trends in the way the representation of the human is constructed.

In Figure 7.3.3, I show the number of discourses present in papers. All papers have at least 3 discourses present, and the majority (46/55) have 4 or 5 present in the writing. This indicates that these discourses are at play in the majority of the documents.



To explore these relationships, I identified documents that had high discursive coherence, where one or two discourses were dominant and others were used sparingly. Few papers had very strong discursive coherence.

One of the documents with high discursive coherence was De Choudhury *et al.* [59], who investigated extreme emotional and behavioral changes to indicate risk for postpartum depression. The preferred term for the human was "mother," even

[6] https://www.nationaleatingdisorders.org/learn/general-information/research-on-males

in the data analysis, results, and methods:

> “...for the volume measure, *mothers* in the extreme change class (C1), exhibit median change of -0.88 postpartum, indicating an 88% drop in posts per day...” [59]

Another example was Reece *et al.* , consistently referring to the humans with Scientific and Person-Centered discourse [260]. They used “participants,” “individuals,” or “observations” throughout:

> “For the depression study, we analyzed 74,990 daily observations (23,541 depressed) from 204 individuals (105 depressed).” [260]

However, for the majority of the corpus (46/55), most documents use 4 or 5 of the five discourses throughout the document. In one example, the authors moves between Clinical, Social Media, Data, and Scientific language in a single sentence, describing the impacts of their work:

> “While we used all *users* posting on mental health subreddits, only a *subset* of *authors* appears in the *control dataset* (around 9% of the *users*; 32,280 appear in the *non mental health* subreddits and 348,040 appear in the mental health subreddits)” [88]

Many documents displayed increasingly complex and confusing representations of the human within sentence-level, “Data from disclosures deemed true were used to build a classifier aiming to distinguish *users with schizophrenia* from *healthy controls*” [78]

Here, the individuals identified to have schizophrenia are represented as the generic “users,” though those that do not are considered “healthy controls;” the reason for these distinctions of the individuals is unclear. These differences make it difficult to follow the sources of data within the paper, and could imply a difference in the sampling strategies for different groups.

More commonly, however, is a distinction between the framing of the individuals involved in the data process and the intended beneficiaries of the research. These documents frame beneficiaries in the Introduction and Discussion sections as “sufferers” or “individuals,” indicating broader societal impacts with more humanizing language.

As an extended example of these patterns, the authors of [230] describe how their approach may identify individuals who suffer from eating disorders, and improving monitoring and detection. In this section, they use person-centric language:

> “We sample *individuals* who self-identify as ED-ed in their profile descriptions on Twitter...thereby providing guidance to develop effective interventions not only for *individuals* but for large groups.” [230] (Introduction)

However, the paper shifts to describing the data, methods, and results oriented around the most dominantly used term, “users” and data-oriented language (“sample,” “dataset”):

> “We first present three types of measures to characterize differences between ED-ed and non-ED-ed *users* on Twitter” [230] (User Characterization)

These findings reveal the importance of discursive framing for building the representations of the human within prediction mental health status from social media. Using a close reading and techniques from discourse analysis, I identified five discourses used to identify the human: Patient/disorder, social media, scientific, data, and person, as well as novel interactions between these discourses in constructing this representation. These discourses reveal numerous gaps, inconsistencies, and potential harms that I explain in the next section.

7.3.4 Discussion

In my findings, I examined the ways that the human research subject is framed in an interdisciplinary area of CS – that which predicts mental health status using social media data.

I turn from presenting the findings to discuss what they mean within the context of other predictive work.

What Are the Harms?

This analysis reveals that this body of research risks the dehumanization of the humans involved in the research process. For mental health, this tension becomes attenuated because the person and the mental disorder – literally tied to their physical body – is the self-stated interest of the researchers. I worry that depersonalizing this data may harm an already vulnerable group of individuals that we are trying to help.

These inconsistent representations of the human have practical consequences for both the research and the humans involved. Just as discourse constructs notions of agency and power [290, 289] and influence lay opinions with downstream impacts on policy [302], these representations imply responsibilities, ethical decisions, privacy protections, and other obligations researchers have to the humans represented within their datasets. Below, I discuss potential risks from these shifting, inconsistent, and sometimes inaccurate descriptions.

Scientific and Collaborative Harms. Describing the representations of the human with shifting and poorly explicated terms can make it challenging to understand key questions for study design, such as inclusions and omissions of data. When four or five discourses come into play within a single section, let alone a single sentence, simply tracking the outcomes is difficult.

Disciplinary knowledge about terms and epistomologies manifest in narrative framings and specialized language through dissemination processes, like publications [303, 304]. When different disciplines are intertwined in collaborative work, interdisciplinarity causes problems with framing and language when it is not clearly established when to adopt certain concepts. Take the example of patient: doctors have different expectations for a "patient," as that has a certain meaning within the field of medicine. When used in a medical science paper, “patient” serves as a boundary maintenance mechanism [305]. However, similar translations of “patient” into this research (when the concerned individuals do not meet the medical definition of a patient) can make the work confusing and otherwise difficult for outsiders to understand. These shifting and imprecise uses of discourse impact interdisciplinary collaboration and understanding, which is critical to this case study on mental health prediction.

These practices also jeopardize reproducibility with these decisions, a core value of much empirical research. Describing the individuals as “patient” can imply to future researchers that actual patients were recruited for participation through undisclosed methods decisions. This relates to recent concerns around construct validity and proxy signals for mental health from social media seen at CHI [249] – are we, through our reporting practices, ensuring that future work can measure what they are trying to measure, and reproduce our own described methods?

Finally, complications can arise when “users” are a stand-in for a broader group of individuals, such as those suffering from depression. When algorithms detect depression in people who use Twitter (hence “user”), the approach may only be applicable to this group, not to non-users of Twitter [281]. Imprecise language describing users to be *all* individuals with depression can harm reproducibility, especially when these approaches imply transferability across disciplines and into practice (*e.g.*, algorithms built on Twitter users deployed among clinical patients at large). I envision this issue emerging in other areas of data-driven analysis, where stakeholders may anticipate generalizability of “solutions” to practice that are not correctly described in the papers.

Harms to the Individuals Involved. Discourse and framing can diminish as well as promote the validity of identities of vulnerable populations [306, 291, 293], and can even reproduce sexist gender roles [307]. Drawing on these insights, I highlight a number of harms directed to the individuals that may be caused as a result of the representations of the human.

Increasing Stigma: Discourse is a major factor in *stigma*, an attribute that makes an individual undesirable, tainted, or socially unworthy [308, 309]. Stigma has very dangerous consequences for those who suffer from mental disorders, such as causing delays in, non-compliance with, or unwillingness to participate in treatment [221], diminished social support [310], and decreased self-esteem and self-efficacy [310]. For other stigmatizing identities other than mental health, real harms exist [283, 311].

I noticed the use of highly stigmatizing language choices. Some discourse compared those who suffer from mental disorders to "normal" or "regular" people, or those who were "neurotypical." I was also concerned by discourse around sufferers, as language like this erases a person's complex identity (a person who suffers from schizophrenia) and replaces it with their mental disorder as the operative portion of their being (a "schizophrenic") [79]. These examples run counter to numerous research and journalistic/reporting guidelines on how to discuss mental disorders without promoting stigma [312, 313]. In particular, mental health advocates have moved away from using the term "neurotypical," which indicates that an individual is "normal," due to the reason that no human is normal after all and people should be acknowledged to be who they are [314]. I worry that discourse likely stigmatizes these identities, and risks causing the harms to the individuals involved in our datasets that we intend to help.

Dehumanizing and Harming the People We Intend to Help: The implications of research are decidedly human-centered – many documents celebrate impacts for monitoring, intervention efforts, and fundamental shifts in how we diagnose and treat mental disorders. Yet, my discourse analysis points to other forms of engagement with people as discursive objects. At the extreme, humans become the literal objects in social media: "accounts" or "blogs," and the data objects themselves,"positive/negative" and "samples".

This illuminates a key tension in this dataset: between the subjective, complex human subject of human-centered research and the neutral and depersonalized data object. Machine learning draws fields like statistics, optimization research, and computer science, who view representations of “the topic of study as an object, whereas to the social scientist the topic of study is the subject.” [315]. In these cases, “data as object” is often assumed to be neural and objective, and many of the translations in this work add objectivity, neutrality, and mathematical detachment from the humans involved [294, 295, 287]. Data is extracted, pre-processed, and provided to a machine learning algorithm for assessment of mental health status; the notion of “harvesting,” “exploiting,” or “extracting” information from a human, however, is incredibly depersonalized. These representations construct the human as a data point for machine translation and optimization, juxtaposed against the social scientist’s interest in the human as subject themselves [315].

What are the harms of depersonalization and dehumanization? As D’Ignazio and Klein contend, “Without the ability of individuals and communities to shape the terms of their own data collection, their bodies can be mined and their data can be extracted far too easily – and done so by powerful institutions who rarely have their best interests at heart.” [287] Dehumanizing data can lead to researcher negligence, ignoring risks baked into algorithmic design and practice. These algorithmic harms have been recognized in other areas [286], and I am concerned that similar harms may already be playing out, in part due to this research.

More broadly, dehumanization clouds the responsibilities and ethical priorities of researchers. It is not only in this case that such concerns have been articulated—I see such tensions emerge in critical data studies, which has challenged conceptualizing who the human research subject is in broader social media scenarios [204, 200] and how researchers interpret their own responsibilities [316]. These risks could include revealing personal or private data, failing to deanonymize quotes [235], releasing datasets that were unethically gathered [317], and conducting experiments on individuals without their consent [222].

Outside of research harms, I worry that such mixed and confusing standards of respecting the human, powerful actors with conflicting interests could cause harm to individuals, or that algorithms will reproduce discriminatory outcomes that perpetuate societal injustice towards those least able to counter their harmful effects [284, 285]. Fundamentally, this means that by depersonalizing and dehumanizing the individuals in the dataset, we will develop systems that do not meet people's needs and desires, and may reproduce socially unjust and undesirable outcomes.

The Complexities of Interdisciplinary Work

From the 164 distinct terms in the analysis, five dominant discourses appeared: social media, disorder/patient, data, scientific, and person-centered. These terms and meanings frequently meshed together, even at the sentence level, and discourses competed throughout the documents. What might be causing these shifting and inconsistent paradigms?

In the corpus, I saw 30 unique venues across an interdisciplinary range of fields. Examples include natural language processing (CLPsych), data mining (WSDM, WWW, KDD), HCI (CHI, CSCW), and health informatics (JMIR). I hypothesize that venue fragmentation within this area may lead to inconsistent scientific standards. Authors have an incentive to match the styles and practices in a venue in efforts to get published, and may not have been trained in how to conduct interdisciplinary work. Reviewers may not know how to navigate such dramatically different topic areas; there are few reviewers who are experts in mental health, social media, and machine learning at the same time, and also can review for all the conferences and journals of interest.

I argue that these papers/documents act as "boundary object," [318, 319] negotiating the tensions between the "social worlds" of disciplines that participate in this research. Star and Griesemer argue that actors must "reconcile these meanings if they wish to cooperate" [318] (p. 388) – I see such shifting and inconsistent discourse as evidence of this. These competing discourses result in translations [294, 295] that construct the human as non-human actor: as a "data point," as an "account." These translations – human to user, human to patient, human to data – may represent the necessary ontological abstractions needed to transform nuanced behavior into rigid computational structures that can be represented by databases, regression models, and neural networks [320].

These reconciliations are at play within these documents, and that there is no shared paradigm between the disciplines in this interdisciplinary example of research. The fluidity of discourses within this space illustrates how scientific frameworks fracture at the intersection of nascent interdisciplinary space, especially while such paradigms are shifting at a field's intellectual beginnings [35].

Guidelines for More Rigorous Work

In light of these findings, one may assume that there is a "correct" discourse to discuss these representations, or that we intend to discourage the use of certain terms in an effort to "police" the use of language within data-centric work. In some ways, this proposition is alluring, since it provides a simple and elegant solution to avoid the hazards and harms I identify above.

The intention of this work is not to direct deliberate language use, nor to advocate that certain terms may never be used within research about people. Directing deliberate language use is reductionist and dangerous for advocating meaningful engagement with these issues. It may make adoption more difficult in other disciplines than HCI, as such a set of rules may feel as if other disciplines are left out of its selection [315].

In the next chapter (Discussion, Ch. 8), I discuss higher-level guidelines that focus on field-level changes in the dissertation; but for now, here are several high-level guidelines to managing language use and conducting this work in more rigorous and ethical ways. My intention is to begin a discussion around these complex issues, and how the reporting processes of research influence the long-term impacts and adoption of research like the prediction of mental health status from social media data.

1. **Avoiding Stigmatizing Language.** Stigmatizing language harms people and communities [283, 221]. Journalistic outlets and non-profit organizations [313, 312] have created detailed standards on language that avoids reproducing or increasing stigma.

I envision best practices in discursive practices complement other topical examples of vulnerable or historically marginalized populations.

2. **Focus on Reproducibility through Language Choice.** I encourage researchers and practitioners to be vigilant in their reporting practices within scientific papers. Good proxy representations make explicit for researchers inside and outside the domain of publication, as well as to non-academic professionals interested in applying these findings. This involves make it abundantly clear what proxy language is chosen and how it is operationalized within a scientific article. There is not necessarily a problem with certain language choices, especially as an abstraction of the human experience is necessary for statistical analyses and processing. However, explicating proxy language may also make clear how the transformations of data produced by individuals is being incorporated into the analysis. Here is a good example of proxy language that makes this transformation clear: "each post thus gave rise to a vector consisting of 93 input attributes and 1 label, or output attribute. The collection of all of these 459 vectors makes up the training data for our machine learning approach" [86]

3. **Creating Shared and Precise Vocabularies.** Considering the above example of "patient" as an example, researchers from various areas are using terms in this new interdisciplinary space in imprecise ways. In addition to making language choice explicit through scientific publication, researchers may also consider working with other fields that are used in the research and coming to consensus around shared practices and vocabularies. To resolve this, I encourage researchers to partner with domain experts through project collaborations. For this example, key domain experts may be in-practice clinicians, medical doctors and researchers, or social workers. For other questions, this may involve a participatory approach early in the research process to understand the preferred terms and language of a specific community.

7.4 Summary

In this section, I discussed my empirical research that focuses on analyzing the methods, practices, and ethics within the larger field of using computational methods for mental health research. First, I identify a dataset of 55 interdisciplinary papers that predict mental health status on social media data. I do so through the methods of identifying candidate papers through a systematic literature review.

In the first empirical study, I examine the practices and methods within this field. I reported on patterns of data annotation and collection, data bias management, pre-processing and feature selection, and model selection and validation. I found that reporting strategies were varied, and often times had issues with the mapping of precise identification of mental health status to the signals available through social media data.

Then, I conduct a discourse analysis to understand the practices of representation within this body of work. My results suggest that competing discourses interact throughout to conceptualize and give agency to the humans within this dataset. I demonstrate how these competing discourses cause harms to scientific reproducibility as well as more urgent harms to the humans who contribute the data as well as are the beneficiaries of this work.

Taken together, these work suggest new opportunities and improvements to be made in the field of predicting mental health status via social media.

CHAPTER 8
FUTURE DIRECTIONS

In my work, I have outlined research in two areas that contribute to our understanding of human-centered algorithms to understand deviant mental health behaviors in online communities. In this section, I describe new impacts of this research agenda that are now possible due to my research.

8.1 Social Computing Implications for Community Management

My work has demonstrated that pro-ED communities have interesting interactions, both within themselves as well as with other communities and platforms. I have developed computational techniques for identifying pro-ED behaviors from other closely related communities around mental health, compared the support and behavior change that they promote, and examined the interactions of these communities with social networks.

One of the most interesting implications of this research is potential impacts for online communities and management. Current strategies for managing or engaging with this content is blunt and not tailored. These strategies involve outright banning of posts, communities, and users (this also includes shadowbans); placing interstitials before access to the content (as Instagram does before engaging with pro-ED content); and quarantining content. Although these strategies attempt to manage this content, these strategies may not be effective in engaging with the deep complexities of mental disorders, their latent impacts on the rest of the community, or in improving behavior more broadly (removing content does not imply that there will be behavior change of the target of these removals). In fact, in my prior work, such as #thyghgapp [28], I demonstrate that Instagram's moderation practices were ineffective at curbing content on their platform. What is missing is engagement with the motivations for behavior, and a more compassionate approach to thinking about how to promote healthier online communities.

How could we design more thoughtful social computing tools and techniques to promote better behavior? I envision these tools at different levels of engagement, tailored to key stakeholders in pro-ED communities.

For Individuals. Some platforms already have simplistic intervention systems to bring help to vulnerable individuals, through recommendations to contact trusted friends or call a hotline for emotional support. In my work, my algorithms show that moderated content shows high vulnerability [30, 28]. These rely on reports from others or a moderator to identify these behaviors, and it is unclear the effectiveness that these interventions have on behavior.

My computational approach can identify moments of high vulnerability and threats to personal safety, making these classifier a good start to identifying moments for just-in-time interventions for vulnerable individuals [321]. Just-in-time interventions are principled, personalized, and timely provisions that bring help to individuals looking for support through deviant mental health behaviors, but before they are potentially alienated from the community with removal of a post that characterizes their vulnerability. These interventions can be triggered immediately, depending on the context and needs of the platform, using the computational approaches I have developed.

After an intervention point is determined, there are several techniques that could be used for that intervention. Here are four examples of potential strategies to deploy just-in-time interventions:

1. A prompt could direct the user to privately message a trusted friend or family member on the platform to discuss their feelings.

2. Resources and click-to-call hotline to an eating disorder management group, or more general talk therapy like 7 Cups of Tea.

3. Ask the user to opt-in to uplifting and positive content being prioritized on their social

media feeds

4. If the user has location enabled on their device and with consent of the user, social networks could display the nearest locations of social support centers.

For less urgent situations, my work can also assist in building reflective tools that promote behavior change and management of unhealthy emotions and behaviors like pro-ED. I envision diary-centric tools built on top of social media platforms which would allow an individual to log artifacts (*e.g.*, posts, images, links) they share on Tumblr.

For Community Managers. These techniques be used by community managers and designers to build better online communities and mechanisms that channel timely support from the greater community. I imagine tools that could appear in the dashboard or sidebars of online communities to appropriate support from other individuals. For example, if one of my algorithmic approaches predicts someone may be moving towards "relapsing" into anorexia recovery through their posting behaviors, recommendations could appear in the Tumblr dashboard of community members to target support to those posts. The system could also match up those in anorexia recovery so that they can receive and provide peer motivation, or connect them to new communities that may be supportive of healthier behaviors.

Another use case of computational approaches is by building systems to visualize and understand community health more broadly. It often can be challenging to understand both the manifestations of physical health as well as the healthiness of the community as a group of individuals. Using the computational linguistic and computer vision tools I have developed, I imagine that interfaces could be designed that quantify shifts in discussions and engagement over time. This could allow community managers to identify trends in engagement and make informed decisions.

For Moderators and Platforms. Finally, these methods could be used in human-machine moderation systems, where it could improve moderation scale, efficiency, and skill development. First, these methods could be a broad "first pass" for human moderators to scale up their tasks. By surfacing posts that likely break guidelines, this method could prune the search space of posts that need intervention. Next, many platforms rely on the report/flag functionality to alert moderators of deviant content. Moderators could also focus on violations that are *not* reported by the platform's users, *e.g.*, , due to the clandestine nature of the pro-ED community [28], or are reported once they reach many users.

These methods could also be used in an online learning-based moderation system [322]. Moderators could set the decision threshold of the classifier per context and need, and adjust for the desired balance between false positives and false negatives. This is an important consideration for moderation of sensitive content like pro-ED—moderators of different platforms may want more or less stringency to reflect what their community guidelines consider harmful behaviors. The moderators could provide feedback to an algorithm so it can learn from images that it misclassifies by retuning the algorithm.

Finally, moderation of sensitive content like pro-ED needs a unique set of skills to assess community harm or guideline violations compared to other kinds of content (*e.g.*, salacious content). This skill set is often gained over time, potentially making new moderator training a time-consuming and emotionally intensive process [323]. With this approach, there is an opportunity to design novel moderation training systems to help new moderators recognize objectively codified rules from prior deviant content.

8.2 Recommendations for Conducting Ethical and Rigorous Work

Research in this area will continue to grow, with new algorithms, data collection means, and new implications for practical use of these algorithms. So, how do we resolve these tensions in predicting mental health status from social media data?

One common assumption or conclusion is that there is one "correct" way of conducting this research, engaging with stakeholders, or referring to the individuals in these datasets. This proposition is alluring, since it provides a simple and elegant solution to avoid harms. However, directing decision-making is reductionist and dangerous for advocating meaningful engagement with these issues. It reduces a complex research process to a checklist—this approach has been eschewed by ethicists who encourage researchers to adopt "ethics as a value." [324] Rigid rules can segment and "bureaucratize" knowledge to only certain stakeholders [325], running counter to a human-centered agenda with involvement of stakeholders. Third, it may make adoption more difficult in other disciplines outside of HCI and social computing, as such a set of rules may feel as if other disciplines are left out [315].

Rather than prescribe a set of strict rules, in this section I sketch out a set of guiding principles. I intend that these guidelines start a conversation around these issues, rather than a prescribed one-size-fits-all solution to solve these challenging issues in research and practice.

Be Mindful of Harmful Practices Within Research. I encourage researchers in this space to be mindful of practices within their work that may be harmful to the humans they claim to prioritize. Connected directly to my findings about discourse, stigmatizing language harms people and communities [283, 221], especially those that suffer from mental disorders. Journalistic outlets and non-profit organizations [313, 312] have created detailed standards on language that avoids reproducing or increasing stigma.

Another latent impact is the abstraction of complex life experiences to quantifiable bits of information that can readily be passed to a classifier. For pro-ED and mental health, I worry about the oversimplification of severity and type of mental illness to a binary classification task of positive or negative presence caused by simplification, which recent research has confirmed these theories [326]. Thinking more broadly, legitimate harms may happen through other kinds of methods decisions and abstractions, like the use of automatic gender recognition technologies that erase non-binary identities [286] or poor performance of language analysis on women and racial/ethnic minorities [327]. Researchers must be conscious of the points of abstractions and potential for harm within a given population.

Thinking more broadly, these harms often happen through the abstraction steps necessary for data analysis – for mental health, I worry about the oversimplification of severity and type of mental illness to a binary classification task of positive or negative presence. These harms may happen through other kinds of methods decisions in other ways, like the use of automatic gender recognition technologies that erase non-binary identities [286], or poor performance of language analysis on women and minorities [327]. I worry about data collection and reporting practices in papers that may expose information about people around sensitive and stigmatizing life experiences [235]. These questions of harm to the individuals involved ought be adopted and navigated throughout the whole research process.

Committing to Reproducibility in Scientific Publications. Given the challenges of interdisciplinarity and the risks to scientific practice I identify above, researchers and practitioners must be vigilant in their reporting practices within scientific papers. My research on methods practices demonstrates that there are large gaps in the precision of ground truth of mental health status, and that abstract language choices make understanding human participants in research challenging.

In published work, researchers may consider disclosing study design and methods decisions to promote reproducibility. Make explicit the decisions made in methods, study design, recruitment procedures, and algorithms selections. Additionally, good proxy representations make explicit for researchers inside and outside the domain of publication, as well as to non-academic professionals interested in applying these findings. These practices and standards for reproducibility are not just the responsibility of the authors – reviewers and disciplinary communities can advocate for higher reproducibility standards and methods reporting within its community.

Shared Vocabularies, and Collaboration. I also advocate for interdisciplinary workshops and shared spaces to develop community-wide shared understandings for this field. The academic community is already responding to these issues through cross-disciplinary seminars, symposia, and conferences, offering collaborative atmospheres for people to work through these problems. Examples of these venues include the recurrent Computational Linguistics and Clinical Psychology (CLPsych) workshop in NLP; the recurrent Computing and Mental Health symposium at CHI; ML4Health at NIPS in 2017; and FAT*. These meetings emphasize that interdisciplinary efforts in collegial environments can produce meaningful solutions.

In particular, Bracken *et al.* argues that "interdisciplinary projects must allocate time to the development of shared vocabularies and understandings" [315] (p. 371). For this kind of work, I encourage researchers to include key stakeholders in the research process, especially clinicians who have key domain knowledge and insights from their experiences. In addition to such partnerships, other stakeholders, like ethicists, designers, and social media platform owners, may be included as well, as they both offer their own perspective and incorporate such algorithms into their systems [190]. Incorporating the knowledge of fields like psychology, privacy, and design, we can work to craft careful algorithmic solutions that may help to mitigate emergent issues of bias, fairness, and discrimination, and execute thoughtful and novel intervention strategies.

Participatory Algorithm Design. I also believe that the individuals who are the target of predictions should be considered a primary stakeholder when developing these systems.

Yet, one underexplored research approach is through direct engagement with individuals as stakeholders to better understand their needs, opinions, and interest in this research. As they are both the providers and the recipients of the algorithmic assessments of mental health status, researchers have an obligation to involve them in these decision-making processes.

Involving individuals as partners can take many forms and may draw on methods from the broader realm of HCI and other areas. This could include participatory design of intervention systems before research begins, interview studies within communities, needs analysis for system design, or speculative design exercises. After computational analyses are designed, user studies of algorithmic systems, longitudinal adoption studies, walk-throughs, and other methods. Outside of the algorithm design process, I also envision continued work in critical algorithms to understand perceptions of research on social media data [212].

Depending on the needs of these communities and the questions involved, research teams interested in computational approaches should consider these participatory approaches in their work. In some cases, it may be undesirable or unethical to directly engage with participants; I encourage research teams to be mindful of these considerations.

Move Beyond Ethics Boards for Guidance. The combinations of benign streams of public data into high-accuracy predictions of mental health status creates complex intersections of research outcomes and stakeholders. Fundamentally, this research is human-centered in that the predictions we make are *on people's data*, not on data as an abstracted notion. When conducting work with direct ties to individuals, we cannot ignore considering implications of this research, even those that extend beyond the purview of ethics boards and oversight committees. Consider and discuss the implications of this research, outside of the normal considerations of ethics committees. Incorporate ethics as a key value in the research process from the beginning.

8.3 Negative Implications of this Research

In a provocative position article published by the ACM's Future of Computing Academy [328], the authors argue that computer scientists should be transparent about the full range of implications of their research, both good and bad. In interest of this goal, I would like to discuss some of the more ambiguous and potentially negative ramifications of using machine learning and data science to understand mental health behaviors in social media data.

One risk of this research is the misuse of algorithmic output for purposes other than those directly beneficial to participants in these communities. Benevolent actors intending to assist sufferers can inadvertently cause harm, as was seen in the case of Samaritan's Radar app [193]. The app scanned Twitter for key phrases, then informed users when their Twitter contacts were potentially in need of emotional support. Although the charity had the right intentions – instrumenting social media activity for suicide prevention – critics identified multiple issues with deployment, ranging from privacy and consent concerns to enabling stalkers and bullies to target victims when they were most vulnerable [193]. Similar concerns may emerge from more recent examples, such as social networks identifying individuals at risk of suicide.

Additional risks surface when these algorithms are used for more ambiguous or nefarious purposes. Mental disorders are complex, and stigma causes consequences for getting treatment. I worry that health care companies could use this research to identify disorders in those they insure – this could be used to give lower premiums for individuals who have "healthy" Twitter feeds, as has been done for adopting healthy behaviors for premium discounts. I also worry that this same data and these methods could be used to deny coverage for treatment for individuals.

There are also actors whose intentions are more complex. Police departments and law enforcement agencies could develop monitoring systems to surveill individuals, and government agencies could use this data in unprecedented ways.

I encourage data scientists and quantitative researchers to consider risks to participants and other stakeholders when examining new areas of sensitive research.

8.4 Towards a Human-Centered Paradigm for Machine Learning and Artificial Intelligence

Thinking more broadly, the use of machine learning and artificial intelligence to provide insights into new problem domains has had complex adoption and public reception. The public has voiced growing concerns around privacy and data use over AIs related to mental health and suicide [190, 193]. In other areas, such as justice, systems that predict the risks of recidivism of criminals are concerned with discriminatory and racist replications of the status quo [329]. Similar concerns have been echoed in the critical data studies space as well [330, 277].

These worrisome incidents imply a disconnect between data science and its application to broader societal questions. Although my research aims towards solving this problem from a technical perspective, there is a lingering question as to what is causing these frictions – why are computational approaches and the individuals, groups, and companies that apply them causing these problems?

One hypothesis suggests that there may a problem in the *paradigms* that govern this work, and appropriate methods and practices around its use. Thomas Kuhn identifies a scientific paradigm as a guiding set of theories, practices, values, and methods that govern appropriate and rigorous scientific research [35]. Numerous scholars have argued that the paradigm for data science revolves around notions of fairness, objectivity, and bias [331, 297, 332, 330], drawing in part from its positivist historical traditions within statistics, mathematics, and computer science. However, these scholars also argue that these paradigms are in constant tension with solving problems they claim, these paradigms and values falling short as they do not necessitate the contextual sensitivity required to engage with larger societal problems [332, 330, 297]. In particular, D'Ignazio and Klein illuminates these value tensions between two competing paradigms of knowing [287]

My area of interest – mental health and social media research – illustrates these tensions in shifting paradigms. The field itself is interdisciplinary, and as discussed in Chapter 7, is in flux with the paradigms and disciplines it considers important – machine learning, human subjects research, digital psychiatry, and mental health, to name a few. This is also illustrated in constant tensions between communities around ethics and appropriate implications of this work, and the growth of ethics as a core concern in natural language processing and machine learning. Fundamentally, the rapid emergence of numerous methods and ethics tensions, privacy issues, and consumer scandals of misuse of digital data implies that there is instability in the paradigms that govern this research.

One potential solution to this paradigm instability has been a *human-centered approach* to machine learning and data science. This has risen in prominence through numerous calls from HCI to reconsider its methods, and most popularly through Stanford's "Human-Centered Artificial Intelligence Institute"(https://hai.stanford.edu/). Many of these approaches have argued that the solutions to these problems is through technical innovation, through anti-bias frameworks, and transparent algorithms. However, this approach risks the values clash that scholars have already argued within – these approaches meaningfully do not engage with the causes of a paradigm problem.

A human-centered paradigm for this work ought be more than just an approach to AI, ML, or other computationally-oriented disciplines that brings "humans in the loop" or solely offer methods innovations (though I certainly do not discourage innovation in technical solutions to these problems). Human-centeredness is a deliberate refocus on the needs of humans, communities, and society and identifying the appropriate tools to solve problems. Commitment to this process is difficult – it involves collaborations with stakeholders and domain experts, an investment in the needs and goals of a particular problem domain, and then (potentially) engineering a solution that brings these people along as equitable partners. In fact, by following these approaches, these processes may reveal that a solution is not appropriate or desirable for a problem domain [333, 330].

For mental health and social media research, a human-centered approach will direct better engagement with communities and can facilitate appropriate modes of representation within the research. Engaging with the guidelines for better research practices above, it might direct engagement with domain experts like clinicians, social media managers, and individuals who participate in these communities. It will likely direct engagement with methods and strategies that balance privacy protections with other values such as equity and justice. In extending to other research topics, I hope that human-centered approaches to problem solving are able to influence

It is also important to recognize that engagement with guidelines on better algorithm design is only one portion of a larger push towards a paradigm of human-centered research. Following human-centered guidelines or engagement with participants does not guarantee that a researcher or a project is “human-centered” and therefore appropriate to conduct, nor do they absolve the researcher of potential harm. In fact, human-centered AI may conclude that AI and ML is an inappropriate solution to these problems

Why must this approach be human-centered? Ultimately, AI and ML serve humans, in their abilities to augment and improve decision-making, bring forth new insights for understandings. Without a human-centered approach to AI and ML, we run the risk of scientific insights that take advantage of vulnerable groups, harm those involved in the studies, and ultimately compromise the integrity of this work. When we as researchers adopt such a brisk approach to research, we risk harming individuals, communities, and groups, and potentially causing more harm than value created through this work. Care must be taken in these circumstances, trying best to adhere to the principles of human-centeredness.

In sum, I hope the work in this thesis is deliberately *towards* a human-centered agenda for data-centric work.

8.5 Limitations

This work focuses on the case example of pro-ED, one example of a deviant mental health behavior. Other areas, like suicidality and self-harm, among others, could be very different in their behaviors. Future work will need to explore this area. I also focus primarily on text and behavioral signals in this work (with the exception of [32]) — pro-ED is a highly visual community, and there is likely much information encoded into the photos that are shared on social media. Future work can explore these new areas of photo sharing practices, as has already been done in recent qualitative work [6, 326].

Another limitation with social media research around stigmatizing topics like pro-ED and mental health are issues of self-presentation and engagement of the users of social media sites. This work focuses on publicly accessible data, and my methods do not learn from any private pro-ED content. There are likely population biases of those who disclose on public social media that they engage in these behaviors, and it is unclear how these compare to traditional clinical populations.

Related to this, it is possible that there are confounding issues of self-presentation of behaviors versus actual behavior within these communities [129]. Although I believe this is mitigated by the highly stigmatizing nature of some of these behaviors (and therefore lower incentive to positively portray one's self), this could impact the content discussed in these communities. Finally, I also expect positive survivorship biases in these datasets, oversampling those who continue to use social media platforms. It is difficult to know the reasons why people stop using online communities to discuss mental health, and drop-off of use may indicate many things.

Another challenge is the complex notion and framing of "deviance" within sociological thought and thinking. In this thesis, I specifically draw on the notion of deviance as part of a socio-ecological model of framing pro-ED behavior. This model focuses on the external, social factors that frame pro-ED as a bad or otherwise harmful behavior. However, deviance has its limitations as an explanatory concept, as it historically can be used to shame individuals with socially "undesirable" behaviors [326]. Future work could consider new frameworks of deviance or socially undesirable behavior that better accommodates these problems.

8.6 New Research Directions

In my future research, I am excited to begin thinking about new problem domains and how to better conduct human-centered research as a problem solving approach.

The pro-ED constitutes one group of deviant mental health behaviors on social media - future work could explore other areas of worrisome or dangerous behaviors in this domain area. This includes research into addiction and recovery, suicidality, and self-injury, to name a few areas that are of interest to me. In addition to new communities, I am also interested in incorporating new techniques into my analysis, such as unsupervised learning, more deep learning, and causal analyses.

I also envision extending this work outside online communities and online trace data - the techniques and approaches I have developed could be applied to other kinds of digital trace data, such as smartphone or app usage patterns, search engine queries, or email data, all gathered with the appropriate consent of parties involved.

Another area I am interested in exploring is more positive perspectives on promoting better behaviors and norms on social media. Much of my work so far has been on identifying the bad content on social media—what about communities that are excellent buffers against bad or dangerous behaviors? Future projects in this space could examine healthy communities, the characteristics that help them maintain their own community health, and how non-moderator actors work to prevent the spread of bad or dangerous behaviors through online spaces. I also interested in studying the ways that communities organically promote healthier discussion, and how we could adopt those findings to practices of community development and management.

Finally, I see a ripe area of future work related to the "full-stack" development of human-centered AI/ML. In this thesis, I focus on data science and ethics, only two pieces of a puzzle to bring a holistic, human-centered approach to these problems in data science. Important, future work can focus on incorporating more stakeholders into these analyses, participatory approaches to community engagement, and designing/deploying interventions based on the algorithmic approaches developed in this line of research. This work would need to be tempered by many of the ethical tensions my research identifies, and I am excited by the opportunities in the potential for human-centered solutions to challenging problems of societal importance.

8.7 Conclusion

My research focuses on designing human-centered algorithms to understand deviant mental health communities in social networks. I study these communities through an extensive examination of pro-ED, a kind of online community that promotes eating disorders as a lifestyle choice and encourages dangerous and life-threatening behaviors. I approach this question through four areas of interest, intersecting with the complexities of managing and moderating a deviant and dangerous mental health behavior in online platforms.

First, I discuss the new methods I bring to the mental health and online communities space, using machine learning, linguistic analysis, and statistical modeling to computationally understand mental health status and pro-ED. Second, I discuss the implications this kind of research (large scale social media analysis) may have on understanding normative behavior and deviance. Third, I cover the impacts that moderation and management can have on pro-ED communities. Finally, I take a step back and examine the methods, ethics, and practices within the field through an analytical analysis of the area and two empirical studies of a corpus of literature.

Collectively, my research makes several interdisciplinary contributions to the fields of HCI, social computing, applied computational social science, and health informatics. My long-term goal is that these approaches help social networks better understand how to develop stronger communities and facilitate better interventions to make a difference in those suffering from mental health challenges, like pro-ED. I hope that my approaches to this complex area of social computing and health inform better understandings of these communities and encourage better approaches to interactions, moderation, and interventions with these communities and more broadly helps contribute to healthier online communities and digital social behavior.

Appendices

APPENDIX A

EXPANDED RUBRIC FOR SYSTEMATIC LITERATURE REVIEW

Basic Information		
Paper Title	Authors	Social Media Site
Year	Venue	Mental health status of study
Overall performance	Summary of contributions	
Construct Validity & Ground Truth		
Source of ground truth (self-report, external validation, etc)	Validation of ground truth	Clinical Involvement
Manual vs automated labeling of ground truth	Diagnostic Level	
Participant Recruitment and Consent		
IRB/Ethics board mentioned?	Public or private data	Was consent obtained? IF not, was not obtaining consent mentioned?
Interaction with Subjects	Type of interaction with human subjects	Subject inclusion/exclusion criteria
Any deception and/or debriefing afterwards	Compensation for participation	Vulnerable population identified
Was community consent acquired	Did researchers share back results with the community	Data solicited or gathered via API
Active vs. passive consent declared?	Anticipated risk or harm to participants	
Data Gathering and Filtering		
Data gathering method (API, scraped, etc)	Language of data	Unit of analysis (post, user, etc)
Number of users	Number of posts	Source of negative data/control
Filtering criteria	Pre-processing steps	Data sampling strategy
Removal of adversarial accounts?	Data stream adjustments (from source)	Data omission/drop criteria
How does the study handle missing data?		
Feature Engineering		
Modality/Type of Data Analyzed	Number of Features	Data-driven or grounded feature engineering

List all linguistic features (tf-idf, LDA, BoW, etc)	Image features (color, tone, pixel information, etc)	Social Interaction Features (retweets, replies, interactions with others on the platform)
Clinical features (self-reported medication use)	Psycholinguistic features (LIWC, PA and NA, etc)	Demographic features
Other features	Dimensionality reduction or feature selection techniques	
Predictive Model/Algorithm Details		
Algorithm or method of choice	Type of prediction (categorical; binary; continuous	Algorithm exclusion criteria
Selection of performance metrics	Hyperparameter tuning	Baseline assumption
Validation step (k-fold cross val, heldout dataset, etc)	Performance (on chosen stat)	Error Analysis Conducted?
Any synthetic sampling used?		
Privacy and Data Protections		
Data de-identified during analysis	Data anonymized later	Data modified in publication
Were quotes used in a paper? If so, were they anonymized?	Data storage and access protections	Data sharing mentioned?
Secondary dataset use	cross-referencing multiple data sources	
Legal Issues and Research Environment		
Terms of service mentioned?	Other applicable laws mentioned?	First author affiliation
Last author affiliation	Author research environments (academic, industry, etc)	Country of residence (for all authors)
Funding sources	Other affiliations	
Implications		
Monitoring of Users	Interventions	Clinical Implications
False positives vs. false negatives	Social network implications	Other stakeholders mentioned (caregivers, family, etc)
Implications to individuals	Bad actors/data misuse	Other social benefits
Algorithmic harms	Other implications mentioned	Reflexivity/awareness of the author's position on the project
Limitations (list all)	Ethics disclosure	

Table A.1: This rubric contains many of the anticipated elements for my systematic literature review. Some of these are yes/no, categorical, or numeric values. Others may be qualitative/thematic.

A.1 List of Papers in Corpus

Authors Year, Citation	**Mental Illness Status**
Facebook	
De Choudhury et al 2014 [71]	Post-partum depression
Park et al 2013 [70]	Depression
Schwartz et al 2014 [73]	Degree of depression
Instagram	
Chancellor et al 2016 [29]	Mental illness severity
Reece and Danforth 2017 [85]	Depression
Zhou, Zhan, and Luo 2017 [258]	Depression; eating disorders
Sina Weibo	
Cheng et al 2017 [194]	5 risk factors for suicidality - suicide probability; Weibo suicide communication; depression; anxiety; stress levels
Guan et al 2015 [82]	High suicide risk
Huang et al 2015 [84]	Suicidal ideation
Huang et al 2014 [264]	Suicidal ideation
Lin et al 2014 [83]	Stressed
Lin et al 2016 [266]	Stressed; stress item (What is causing stress)
Wang et al 2013 [76]	Depression
Zhang et al 2015 [265]	Suicide risk score (SPS value)
Zhao, Jia, and Feng 2015 [267]	Stress
Reddit	
De Choudhury et al 2016‘[81]	Suicidal ideation
Gkotsis et al 2017 [88]	Bipolar disorder; borderline personality disorder; schizophrenia; anxiety; depression; self harm; suicide crisis

Saha and De Choudhury 2017 [259]	High or low stress
Shen and Rudzicz 2017 [77]	Anxiety
Tumblr	
Chancellor, Mitra, and De Choudhury 2016 [31]	Recovery from anorexia
De Choudhury 2015 [80]	Anorexia content; Anorexia versus in-recovery
Simms et al 2017 [86]	Cognitive distortions
Twitter	
Benton, Mitchell, and Hovy 2017 [213]	Non-neurotypical; anxiety; depression; suicide; eating disorder; panic attack; schizophrenia; bipolar disorder; post-traumatic stress disorder
Birnbaum et al 2017 [78]	Schizophrenia
Braithwaite et al 2016 [263]	Suicidal communication
Burnap, Colombo, and Scourfield 2015 [274]	Suidical vs 5 other classes about suicide-related communication
Coppersmith, Dredze, and Harman 2014 [58]	Bipolar disorder; depression; post-traumatic stress disorder; seasonal affective disorder
Coppersmith et al 2015 [232]	Anxiety; bipolar disorder; borderline personality disorder; depression; eating disorder; obssesive compulsive disorder; post-traumatic stress disorder; schizophrenia; seasonal affective disorder
Coppersmith, Harman, and Dredze 2014 [17]	Post-traumatic stress disorder
Coppersmith et al 2016 [195]	Suicide Attempts
De Choudhury, Counts, and Horvitz 2013 [59]	Post-partum changes
De Choudhury, Counts, and Horvitz 2013 [254]	Depression

De Choudhury et al 2013 [16]	Depression
Homan et al 2014 [272]	Distress (related to suicide)
Jamil et al 2017 [234]	Depression (both user and tweet level)
Loveys et al 2017 [256]	Anxiety; eating disorder; schizophrenia; suicide attempt; panic attacks
McManus et al 2015 [275]	Schizophrenia
Mitchell, Hollingshead, and Coppersmith 2015 [79]	Schizophrenia
O'Dea et al 2015 [273]	Suicide
Preotiuc-Pietro et al 2015 [270]	Depression; post-traumatic stress disorder
Prieto et al 2014 [269]	Depression; eating disorders
Reece et al 2017 [260]	Depression; post-traumatic stress disorder
Resnik et al 2015 [271]	Depression
Saha et al 2017 [87]	High or low mood instability
Saravia et al 2016 [262]	Bipolar disorder; borderline personality disorders
Shen et al 2017 [261]	Depressed
Tsugawa et al 2015 [75]	Depression
Tsugawa et al 2013 [253]	Depression score (Zung Self-rating)
Vedula and Parthasarathy 2017 [255]	Depression
Wang et al 2017 [230]	Eating disorders
Other SNS	
Nakamura et al 2014 [72]	Depressive symptoms [TOBYO Toshoshitsu]
Nguyen et al 2014 [268]	Depression [LiveJournal]
Wang et al 2017 [89]	Self-harm [Flickr]
Shen et al 2013 [251]	Depressed vs. sad [PTT (Taiwanese Bulletin Board System)]
Masuda, Kurahashi, and Onari 2013 [252]	Suicide Ideation [mixi (Japanese social network)]

REFERENCES

[1] G. Eysenbach, J. Powell, M. Englesakis, C. Rizo, and A. Stern, “Health related virtual communities and electronic support groups: Systematic review of the effects of online peer to peer interactions,” *Bmj*, vol. 328, no. 7449, p. 1166, 2004.

[2] M. De Choudhury and E. Kıcıman, “The language of social support in social media and its effect on suicidal ideation risk,” in *Proceedings of the... International AAAI Conference on Weblogs and Social Media. International AAAI Conference on Weblogs and Social Media*, vol. 2017, 2017, p. 32.

[3] Y. Hong, N. C. Pena-Purcell, and M. G. Ory, “Outcomes of online support and resources for cancer survivors: A systematic literature review,” *Patient education and counseling*, vol. 86, no. 3, pp. 288–296, 2012.

[4] D. L. Borzekowski, S. Schenk, J. L. Wilson, and R. Peebles, “E-ana and e-mia: A content analysis of pro–eating disorder web sites,” *American journal of public health*, vol. 100, no. 8, p. 1526, 2010.

[5] R. A. Fleming-May and L. E. Miller, “i’m scared to look. but i’m dying to knowi: Information seeking and sharing on pro-ana weblogs,” *Proceedings of the Association for Information Science and Technology*, vol. 47, no. 1, pp. 1–9, 2010.

[6] J. A. Pater, O. L. Haimston, N. Andalibi, and E. D. Mynatt, “‘‘hunger hurts but starving works”: Characterizing the presentation of eating disorders online,” in *Proceedings of the 19th ACM conference on Computer Supported Cooperative Work & Social Computing (CSCW)*, 2016.

[7] P. F. Sullivan, “Mortality in anorexia nervosa,” *The American journal of psychiatry*, vol. 152, no. 7, p. 1073, 1995.

[8] A. P. Association *et al.*, *Diagnostic and statistical manual of mental disorders (DSM-5®)*. American Psychiatric Pub, 2013.

[9] J. Pater and E. Mynatt, “Defining digital self-harm,” in *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, ACM, 2017, pp. 1501–1513.

[10] A. Laye-Gindhu and K. A. Schonert-Reichl, “Nonsuicidal self-harm among community adolescents: Understanding the whatsi and whysi of self-harm,” *Journal of youth and Adolescence*, vol. 34, no. 5, pp. 447–457, 2005.

[11] C. S. Crandall, “Social contagion of binge eating.,” *Journal of personality and social psychology*, vol. 55, no. 4, p. 588, 1988.

[12] T Emmens and A Phippen, “Evaluating online safety programs,” *Harvard Berkman Center for Internet and Society.[23 July 2011]*, 2010.

[13] D. Giles, “Constructing identities in cyberspace: The case of eating disorders,” *British journal of social psychology*, vol. 45, no. 3, pp. 463–477, 2006.

[14] E. Yom-Tov, L. Fernandez-Luque, I. Weber, and S. P. Crain, “Pro-anorexia and pro-recovery photo sharing: A tale of two warring tribes,” *Journal of medical Internet research*, vol. 14, no. 6, 2012.

[15] A. Chen, *The laborers who keep dick pics and beheadings out of your facebook feed*, 2014.

[16] M. De Choudhury, M. Gamon, S. Counts, and E. Horvitz, “Predicting depression via social media.,” *ICWSM*, vol. 2, pp. 128–137, 2013.

[17] G. Coppersmith, C. Harman, and M. H. Dredze, “Measuring post traumatic stress disorder in Twitter,” in *ICWSM*, vol. 2, 2014, pp. 579–582.

[18] E. Chandrasekharan, M. Samory, A. Srinivasan, and E. Gilbert, “The bag of communities: Identifying abusive behavior online with preexisting internet data,” in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, ACM, 2017, pp. 3175–3187.

[19] R. L. Akers, “Deviant behavior: A social learning approach,” 1977.

[20] J. R. Suler and W. L. Phillips, “The bad boys of cyberspace: Deviant behavior in a multimedia chat community,” *CyberPsychology & Behavior*, vol. 1, no. 3, pp. 275–294, 1998.

[21] M. A. Hogg and S. A. Reid, “Social identity, self-categorization, and the communication of group norms,” *Communication theory*, vol. 16, no. 1, pp. 7–30, 2006.

[22] M. K. Nock, “Self-injury,” *Annual review of clinical psychology*, vol. 6, pp. 339–363, 2010.

[23] O. Knapton, “Pro-anorexia: Extensions of ingrained concepts,” *Discourse & Society*, vol. 24, no. 4, pp. 461–477, 2013.

[24] P. Maloney, “Quod me nutrit, me destruit: The pro-anorexia movement and religion,” in *annual meeting of the American Sociological Association, Boston.*, 2008.

[25] Tumblr, *"tumblr community guidelines"*, 2016.

[26] D. Restauri, *''tumblr to pinterest to instagram – the self-harm 'thinspo' community is house-hunting"*, 2012.

[27] A. Hess, *Let them blog: The panic over pro-anorexia websites and social media isnt healthy*, 2015.

[28] S. Chancellor, J. Pater, T. Clear, E. Gilbert, and M. De Choudhury, "#thyghgapp: Instagram content moderation and lexical variation in pro-eating disorder communities," in *Proceedings of the 2016 Conference on Computer Supported Cooperative Work & Social Computing(CSCW)*, ACM, 2016.

[29] S. Chancellor, Z. J. Lin, E. Goodman, S. Zerwas, and M. De Choudhury, "Quantifying and predicting mental illness severity in online pro-eating disorder communities," in *Proceedings of the 19th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW)*, ACM, 2016.

[30] S. Chancellor, Z. J. Lin, and M. De Choudhury, "This post will just get taken down: Characterizing removed pro-eating disorder social media content," in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, ACM, 2016, pp. 1157–1162.

[31] S. Chancellor, T. Mitra, and M. De Choudhury, "Recovery amid pro-anorexia: Analysis of recovery in social media," in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, ACM, 2016, pp. 2111–2123.

[32] S. Chancellor, Y. Kalantidis, J. A. Pater, M. De Choudhury, and D. A. Shamma, "Multimodal classification of moderated online pro-eating disorder content," in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, ACM, 2017, pp. 3213–3226.

[33] S. Chancellor, A. Hu, and M. De Choudhury, "Norms matter: Contrasting social support around behavior change in online weight loss communities," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, ACM, 2018.

[34] S. Chancellor, M. L. Birnbaum, E. D. Caine, V. Silenzio, and M. De Choudhury, "A taxonomy of ethical tensions in inferring mental health states from social media," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, ACM, 2019, pp. 79–88.

[35] T. S. Kuhn, *The structure of scientific revolutions*. University of Chicago press, 2012.

[36] J. Arcelus, A. J. Mitchell, J. Wales, and S. Nielsen, “Mortality rates in patients with anorexia nervosa and other eating disorders: A meta-analysis of 36 studies,” *Archives of general psychiatry*, vol. 68, no. 7, pp. 724–731, 2011.

[37] J. I. Hudson, E. Hiripi, H. G. Pope, and R. C. Kessler, “The prevalence and correlates of eating disorders in the national comorbidity survey replication,” *Biological psychiatry*, vol. 61, no. 3, pp. 348–358, 2007.

[38] N. E. D. Association, *Health consequences of eating disorders*, `https://www.nationaleatingdisorders.org/health-consequences-eating-disorders`, [Online; accessed 5-February-2018], 2018.

[39] A. A. Casilli, P. Tubaro, and P. Araya, “Ten years of ana: Lessons from a transdisciplinary body of literature on online pro-eating disorder websites,” *Social Science Information*, vol. 51, no. 1, pp. 120–139, 2012.

[40] L. R. Shade, “Weborexics: The ethical issues surrounding pro-ana websites,” 2003.

[41] S. R. Brotsky and D. Giles, “Inside the pro-anai community: A covert online participant observation,” *Eating disorders*, vol. 15, no. 2, pp. 93–109, 2007.

[42] C. F. Bates, “i am a waste of breath, of space, of timei metaphors of self in a pro-anorexia group,” *Qualitative health research*, vol. 25, no. 2, pp. 189–204, 2015.

[43] N. Boero and C. J. Pascoe, “Pro-anorexia communities and online interaction: Bringing the pro-ana body online,” *Body & Society*, vol. 18, no. 2, pp. 27–57, 2012.

[44] C. F. Bates, “‘‘i am a waste of breath, of space, of time” metaphors of self in a pro-anorexia group,” *Qualitative health research*, vol. 25, no. 2, pp. 189–204, 2015.

[45] C. R. Rouleau and K. M. Von Ranson, “Potential risks of pro-eating disorder websites,” *Clinical Psychology Review*, vol. 31, no. 4, pp. 525–531, 2011.

[46] S. P. Lewis and Y. Seko, “A double-edged sword: A review of benefits and risks of online nonsuicidal self-injury activities,” *Journal of clinical psychology*, vol. 72, no. 3, pp. 249–262, 2016.

[47] Y. Gerrard, “Beyond the hashtag: Circumventing content moderation on social media,” *New Media & Society*, vol. 20, no. 12, pp. 4492–4511, 2018.

[48] J. Callaghan, “Research in online spaces:’tumblr’and eating’disorder’,” 2013.

[49] M. De Choudhury, “Anorexia on tumblr: A characterization study,” in *Proceedings of the 5th International Conference on Digital Health 2015*, ACM, 2015, pp. 43–50.

[50] S. Syed-Abdul, L. Fernandez-Luque, W.-S. Jian, Y.-C. Li, S. Crain, M.-H. Hsu, Y.-C. Wang, D. Khandregzen, E. Chuluunbaatar, P. A. Nguyen, *et al.*, "Misleading health-related information promoted through video-based social media: Anorexia on youtube," *Journal of medical Internet research*, vol. 15, no. 2, 2013.

[51] L. Claes, K. Luyckx, P. Bijttebier, B. Turner, A. Ghandi, J. Smets, J. Norre, L. Van Assche, E. Verheyen, Y. Goris, *et al.*, "Non-suicidal self-injury in patients with eating disorder: Associations with identity formation above and beyond anxiety and depression," *European eating disorders review*, vol. 23, no. 2, pp. 119–125, 2015.

[52] Instagram, *Community guidelines*, `https://help.instagram.com/?helpref=faq_content`, [Online; accessed 5-February-2018], 2018.

[53] J. Preece and D. Maloney-Krichmar, "Online communities: Design, theory, and practice," *Journal of Computer-Mediated Communication*, vol. 10, no. 4, pp. 00–00, 2005.

[54] N. Andalibi and A. Forte, "Social computing researchers, vulnerability, and peer support," in *Ethical Encounters in HCI: Research in Sensitive and Complex Settings Workshop at the Conference on Human Factors in Computing Systems*, 2016.

[55] H. A. Schwartz, J. C. Eichstaedt, M. L. Kern, L. Dziurzynski, R. E. Lucas, M. Agrawal, G. J. Park, S. K. Lakshmikanth, S. Jha, M. E. Seligman, *et al.*, "Characterizing geographic variation in well-being using tweets.," in *ICWSM*, 2013, pp. 583–591.

[56] A. Culotta, "Estimating county health statistics with twitter," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, 2014, pp. 1335–1344.

[57] M. J. Paul and M. Dredze, "You are what you tweet: Analyzing twitter for public health.," in *ICWSM*, 2011.

[58] G. Coppersmith, M. Dredze, and C. Harman, "Quantifying mental health signals in twitter," in *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, 2014, pp. 51–60.

[59] M. De Choudhury, S. Counts, and E. Horvitz, "Predicting postpartum changes in emotion and behavior via social media," *CHI*, pp. 3267–3276, 2013.

[60] R. A. Calvo, D. N. Milne, M. S. Hussain, and H. Christensen, "Natural language processing in mental health applications using non-clinical texts," *Natural Language Engineering*, vol. 23, no. 5, pp. 649–685, 2017.

[61] J. W. Pennebaker, "Writing about emotional experiences as a therapeutic process," *Psychological science*, vol. 8, no. 3, pp. 162–166, 1997.

[62] J. Bollen, H. Mao, and A. Pepe, "Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena," in *Fifth International AAAI Conference on Weblogs and Social Media*, 2011.

[63] S. A. Golder and M. W. Macy, "Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures," *Science*, vol. 333, no. 6051, pp. 1878–1881, 2011.

[64] P. S. Dodds, K. D. Harris, I. M. Kloumann, C. A. Bliss, and C. M. Danforth, "Temporal patterns of happiness and information in a global social network: Hedonometrics and twitter," *PloS one*, vol. 6, no. 12, e26752, 2011.

[65] A. D. Kramer, "An unobtrusive behavioral model of gross national happiness," in *Proceedings of the SIGCHI conference on human factors in computing systems*, ACM, 2010, pp. 287–290.

[66] M. D. Back, J. M. Stopfer, S. Vazire, S. Gaddis, S. C. Schmukle, B. Egloff, and S. D. Gosling, "Facebook profiles reflect actual personality, not self-idealization," *Psychological science*, vol. 21, no. 3, pp. 372–374, 2010.

[67] G. Eysenbach, "Infodemiology and infoveillance: Framework for an emerging set of public health informatics methods to analyze search, communication and publication behavior on the internet," *Journal of medical Internet research*, vol. 11, no. 1, e11, 2009.

[68] J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant, "Detecting influenza epidemics using search engine query data," *Nature*, vol. 457, no. 7232, p. 1012, 2009.

[69] A. Sadilek, H. Kautz, and V. Silenzio, "Modeling spread of disease from social interactions," in *Sixth International AAAI Conference on Weblogs and Social Media*, 2012.

[70] S. Park, S. W. Lee, J. Kwak, M. Cha, and B. Jeong, "Activities on Facebook reveal the depressive state of users," *Journal of Medical Internet Research*, vol. 15, no. 10, pp. 1–15, 2013.

[71] M. De Choudhury, S. Counts, E. J. Horvitz, and A. Hoff, "Characterizing and Predicting Postpartum Depression from Shared Facebook Data," in *CSCW*, ser. CSCW '14, ACM, 2014, pp. 626–638.

[72] T Nakamura, K Kubo, Y Usuda, and E Aramaki, "Defining patients with depressive disorder by using textual information," *AAAI*, 2014.

[73] H. A. Schwartz, J. Eichstaedt, M. L. Kern, G. Park, M. Sap, D. Stillwell, M. Kosinski, and L. Ungar, "Towards assessing changes in degree of depression through facebook," in *CLPsych*, 2014, pp. 118–125.

[74] P. Resnik, A. Garron, and R. Resnik, "Using topic modeling to improve prediction of neuroticism and depression in college students," in *Proceedings of the 2013 conference on empirical methods in natural language processing*, 2013, pp. 1348–1353.

[75] S. Tsugawa, Y. Kikuchi, F. Kishino, K. Nakajima, Y. Itoh, and H. Ohsaki, "Recognizing Depression from Twitter Activity," ser. CHI '15, ACM, 2015, pp. 3187–3196.

[76] X. Wang, C. Zhang, Y. Ji, L. Sun, L. Wu, and Z. Bao, "A depression detection model based on sentiment analysis in micro-blog social network," in *PAKDD*, vol. 7867 LNAI, 2013, pp. 201–213.

[77] J. H. Shen and F. Rudzicz, "Detecting anxiety on Reddit," in *CLPsych*, 2017, pp. 58–65.

[78] M. L. Birnbaum, S. K. Ernala, A. F. Rizvi, M. De Choudhury, and J. M. Kane, "A Collaborative Approach to Identifying Social Media Markers of Schizophrenia by Employing Machine Learning and Clinical Appraisals," *JMIR*, vol. 19, no. 8, 2017.

[79] M. Mitchell, K. Hollingshead, and G. Coppersmith, "Quantifying the language of schizophrenia in social media," in *CLPsych*, 2015, pp. 11–20.

[80] M. De Choudhury, "Anorexia on Tumblr : A Characterization Study on Anorexia," in *Proceedings of DH'15: 5th ACM Digital Health Conference. DH'15.*, ser. DH '15, ACM, 2015, pp. 43–50.

[81] M. De Choudhury, E. Kiciman, M. Dredze, G. Coppersmith, and M. Kumar, "Discovering shifts to suicidal ideation from mental health content in social media," in *CHI*, ACM, 2016, pp. 2098–2110.

[82] L. Guan, B. Hao, Q. Cheng, P. S. F. Yip, and T. Zhu, "Identifying Chinese Microblog Users With High Suicide Probability Using Internet-Based Profile and Linguistic Features: Classification Model," *JMIR Mental Health*, vol. 2, no. 2, e17, 2015.

[83] H. Lin, J. Jia, Q. Guo, Y. Xue, Q. Li, J. Huang, L. Cai, and L. Feng, "User-level psychological stress detection from social media using deep neural network," in *MM*, New York, NY, USA: IEEE, 2014, pp. 507–516.

[84] X. Huang, X. Li, L. Zhang, T. Liu, D. Chiu, and T. Zhu, "Topic Model for Identifying Suicidal Ideation in Chinese Microblog," in *29th Pacific Asia Conference on Language, Information and Computation*, Proceeedings of the 29th Pacific Asia Conference on Language, 2015, pp. 553–562.

[85] A. G. Reece and C. M. Danforth, "Instagram photos reveal predictive markers of depression," *EPJ Data Science*, vol. 6, no. 1, p. 15, 2017.

[86] T Simms, C Ramstedt, M Rich, M Richards, T Martinez, and C. Giraud-Carrier, "Detecting Cognitive Distortions Through Machine Learning Text Analytics," *ICHI*, 2017.

[87] K. Saha, L. Chan, K. De Barbaro, G. D. Abowd, and M. De Choudhury, "Inferring mood instability on social media by leveraging ecological momentary assessments," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 1, no. 3, p. 95, 2017.

[88] G. Gkotsis, A. Oellrich, S. Velupillai, M. Liakata, T. J. P. Hubbard, R. J. B. Dobson, and R. Dutta, "Characterisation of mental health conditions in social media using Informed Deep Learning," *SCIENTIFIC REPORTS*, vol. 7, 2017.

[89] Y. Wang, J. Tang, J. Li, B. Li, Y. Wan, C. Mellina, N. O'Hare, and Y. Chang, "Understanding and Discovering Deliberate Self-harm Content in Social Media," in *WWW*, ser. WWW '17, International World Wide Web Conferences Steering Committee, 2017, pp. 93–102.

[90] L. Manikonda and M. De Choudhury, "Modeling and understanding visual attributes of mental health disclosures in social media," in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, ACM, 2017, pp. 170–181.

[91] D. Yang, Z. Yao, and R. E. Kraut, "Self-disclosure and channel difference in online health support groups.," in *ICWSM*, 2017, pp. 704–707.

[92] M. De Choudhury, S. S. Sharma, T. Logar, W. Eekhout, and R. C. Nielsen, "Gender and cross-cultural differences in social media disclosures of mental illness," in *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, ACM, 2017, pp. 353–369.

[93] M. De Choudhury, S. Counts, and M. Gamon, “Not all moods are created equal! exploring human emotional states in social media,” in *Sixth international AAAI conference on weblogs and social media*, 2012.

[94] Y.-C. Wang, R. Kraut, and J. M. Levine, “To stay or leave?: The relationship of emotional and informational support to commitment in online health support groups,” in *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, ACM, 2012, pp. 833–842.

[95] M. J. Paul and M. Dredze, “Discovering health topics in social media using topic models,” *PloS one*, vol. 9, no. 8, e103408, 2014.

[96] M. Kumar, M. Dredze, G. Coppersmith, and M. De Choudhury, “Detecting changes in suicide content manifested in social media following celebrity suicides,” in *Proceedings of the 26th ACM Conference on Hypertext & Social Media*, ACM, 2015, pp. 85–94.

[97] M. De Choudhury, M. Kumar, and I. Weber, “Computational approaches toward integrating quantified self sensing and social media,” in *CSCW: proceedings of the Conference on Computer-Supported Cooperative Work. Conference on Computer-Supported Cooperative Work*, NIH Public Access, vol. 2017, 2017, p. 1334.

[98] H. Ma, C. E. Smith, L. He, S. Narayanan, R. A. Giaquinto, R. Evans, L. Hanson, and S. Yarosh, “Write for Life : Persisting in Online Health Communities with Expressive Writing and Social Support,” *Proc. ACM Human-Computer Interaction. CSCW*, vol. 1, no. 2, 2017.

[99] A. R. Favazza, L. DeRosear, and K. Conterio, “Self-mutilation and eating disorders,” *Suicide and Life-Threatening Behavior*, vol. 19, no. 4, pp. 352–361, 1989.

[100] R. Peebles, J. L. Wilson, and J. D. Lock, “Self-injury in adolescents with eating disorders: Correlates and provider bias,” *Journal of Adolescent Health*, vol. 48, no. 3, pp. 310–313, 2011.

[101] K. Kostro, J. B. Lerman, and E. Attia, “The current status of suicide and self-injury in eating disorders: A narrative review,” *Journal of eating disorders*, vol. 2, no. 1, p. 19, 2014.

[102] F. R. Smink, D. Van Hoeken, and H. W. Hoek, “Epidemiology of eating disorders: Incidence, prevalence and mortality rates,” *Current psychiatry reports*, vol. 14, no. 4, pp. 406–414, 2012.

[103] S. K. Farber, C. C. Jackson, J. K. Tabin, and E. Bachar, “Death and annihilation anxieties in anorexia nervosa, bulimia, and self-mutilation.,” *Psychoanalytic Psychology*, vol. 24, no. 2, p. 289, 2007.

[104] A. Favaro and P. Santonastaso, "Impulsive and compulsive self-injurious behavior in bulimia nervosa: Prevalence and psychological correlates," *The Journal of Nervous and Mental Disease*, vol. 186, no. 3, pp. 157–165, 1998.

[105] S. A. S. Germain and J. M. Hooley, "Direct and indirect forms of non-suicidal self-injury: Evidence for a distinction," *Psychiatry research*, vol. 197, no. 1, pp. 78–84, 2012.

[106] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *the Journal of machine Learning research*, vol. 3, pp. 993–1022, 2003.

[107] M. Collins, R. E. Schapire, and Y. Singer, "Logistic regression, adaboost and bregman distances," *Machine Learning*, vol. 48, no. 1-3, pp. 253–285, 2002.

[108] T. Clark, M. Bradburn, S. Love, and D. Altman, "Survival analysis part i: Basic concepts and first analyses," *British journal of cancer*, vol. 89, no. 2, pp. 232–238, 2003.

[109] M. K. Parmar and D. Machin, *Survival analysis: a practical approach*. John Wiley & Sons Chichester, 1995.

[110] J. B. Willett and J. D. Singer, "Investigating onset, cessation, relapse, and recovery: Why you should, and how you can, use discrete-time survival analysis to examine event occurrence.," *Journal of consulting and clinical psychology*, vol. 61, no. 6, p. 952, 1993.

[111] E. L. Kaplan and P. Meier, "Nonparametric estimation from incomplete observations," *Journal of the American statistical association*, vol. 53, no. 282, pp. 457–481, 1958.

[112] D. R. Cox, "Regression models and life-tables," in *Breakthroughs in statistics*, Springer, 1992, pp. 527–541.

[113] C. Chung and J. W. Pennebaker, "The psychological functions of function words," *Social communication*, pp. 343–359, 2007.

[114] T. E. Oxman, S. D. Rosenberg, and G. J. Tucker, "The language of paranoia.," *The American Journal of Psychiatry*, 1982.

[115] S. Rude, E.-M. Gortner, and J. Pennebaker, "Language use of depressed and depression-vulnerable college students," *Cognition & Emotion*, vol. 18, no. 8, pp. 1121–1133, 2004.

[116] J. W. Pennebaker, T. J. Mayne, and M. E. Francis, “Linguistic predictors of adaptive bereavement.,” *Journal of personality and social psychology*, vol. 72, no. 4, p. 863, 1997.

[117] C. G. Fairburn, R. Shafran, and Z. Cooper, “A cognitive behavioural theory of anorexia nervosa,” *Behaviour Research and Therapy*, vol. 37, no. 1, pp. 1–13, 1999.

[118] J. Gavin, K. Rodham, and H. Poyer, “The presentation of “pro-anorexia in online group interactions,” *Qualitative Health Research*, vol. 18, no. 3, pp. 325–333, 2008.

[119] K. M. Pike, “Long-term course of anorexia nervosa: Response, relapse, remission, and recovery,” *Clinical psychology review*, vol. 18, no. 4, pp. 447–475, 1998.

[120] C. M. Bulik, N. D. Berkman, K. A. Brownley, J. A. Sedway, and K. N. Lohr, “Anorexia nervosa treatment: A systematic review of randomized controlled trials,” *International Journal of Eating Disorders*, vol. 40, no. 4, pp. 310–320, 2007.

[121] C. Hiruncharoenvate, Z. Lin, and E. Gilbert, “Algorithmically bypassing censorship on sina weibo with nondeterministic homophone substitutions,” in *Ninth International AAAI Conference on Web and Social Media*, 2015.

[122] L. Blackwell, J. Dimond, S. Schoenebeck, and C. Lampe, “Classification and its consequences for online harassment: Design insights from heartmob,” *PACM on Human-Computer Interaction*, vol. 1, no. CSCW, 2018.

[123] M. K. Lapinski and R. N. Rimal, “An explication of social norms,” *Communication theory*, vol. 15, no. 2, pp. 127–147, 2005.

[124] R. B. Cialdini, R. R. Reno, and C. a. Kallgren, “A focus theory of normative conduct: Recycling the concept of norms to reduce littering in public places.,” *Journal of Personality and Social Psychology*, vol. 58, no. 6, pp. 1015–1026, 1990.

[125] G. Burnett, M. Besant, and E. A. Chatman, “Small worlds : Normative behavior in virtual communities and feminist bookselling,” *Journal of the American Society for Information Science and Technology*, vol. 52, no. 7, 2001.

[126] T. Postmes, R. Spears, and M. Lea, “The formation of group norms in computer-mediated communication,” *Human communication research*, vol. 26, no. 3, pp. 341–371, 2000.

[127] G. Burnett and L. Bonnici, “Beyond the FAQ : Explicit and implicit norms in Usenet newsgroups,” *Library and Information Science Research*, vol. 25, pp. 333–351, 2003.

[128] J. S. Donath, “Identity and deception in the virtual community,” *Communities in cyberspace*, vol. 1996, pp. 29–59, 1999.

[129] E. Goffman, *The presentation of self in everyday life*. Harmondsworth, 1978.

[130] H. Giles, D. M. Taylor, and R. Bourhis, “Towards a theory of interpersonal accommodation through language: Some canadian data,” *Language in society*, vol. 2, no. 2, pp. 177–192, 1973.

[131] W. Labov, *Principles of linguistic change, cognitive and cultural factors*. John Wiley & Sons, 2011, vol. 3.

[132] W. S. E. Lam, “Language socialization in online communities,” in *Encyclopedia of language and education*, Springer, 2008, pp. 2859–2869.

[133] D. Nguyen and C. P. Rosé, “Language use as a reflection of socialization in online communities,” in *Proceedings of the Workshop on Languages in Social Media*, Association for Computational Linguistics, 2011, pp. 76–85.

[134] R. Hughes and J. Coakley, “Positive Deviance Among Athletes : The Implications of Overconformity to the Sport Ethic,” *Sociology of Sport Journal*, vol. 8, pp. 307–325, 1991.

[135] E. Durkheim, “Suicide: A study in sociology (ja spaulding & g. simpson, trans.),” *Glencoe, IL: Free Press.(Original work published 1897)*, 1951.

[136] R. K. Merton, “Anomie, anomia, and social interaction: Contexts of deviant behavior,” *Anomie and deviant behavior: A discussion and critique*, pp. 213–242, 1964.

[137] R. L. Burgess and R. L. Akers, “A Differential Association-Reinforcement Theory of Criminal Behavior,” *Social Problems*, vol. 14, no. 2, pp. 128–147, 1966.

[138] H. S. Becker, *Outsiders*. Simon and Schuster, 2008.

[139] K. Marx, “From capital,” in *Readings In The Economics Of The Division Of Labor: The Classical Tradition*, World Scientific, 2005, pp. 177–187.

[140] M. Millman, “She did it all for love: A feminist view of the sociology of deviance,” *Sociological Inquiry*, vol. 45, no. 2-3, pp. 251–279, 1975.

[141] A. J. Kim, *Community building on the web: Secret strategies for successful online communities*. Addison-Wesley Longman Publishing Co., Inc., 2000.

[142] A. Leavitt, “This is a throwaway account: Temporary technical identities and perceptions of anonymity in a massive online community,” in *Proceedings of the 18th*

ACM Conference on Computer Supported Cooperative Work & Social Computing, ACM, 2015, pp. 317–327.

[143] R. E. Kraut, P. Resnick, S. Kiesler, M. Burke, Y. Chen, N. Kittur, J. Konstan, Y. Ren, and J. Riedl, *Building successful online communities: Evidence-based social design*. Mit Press, 2012.

[144] S. L. Bryant, A. Forte, and A. Bruckman, "Becoming wikipedian: Transformation of participation in a collaborative online encyclopedia," in *Proceedings of the 2005 international ACM SIGGROUP conference on Supporting group work*, ACM, 2005, pp. 1–10.

[145] C. Fiesler, S. Morrison, and A. S. Bruckman, "An Archive of Their Own: A Case Study of Feminist HCI and Values in Design," in *CHI*, 2016, pp. 2574–2585, ISBN: 9781450333627.

[146] J. Arguello, B. S. Butler, E. Joyce, R. Kraut, K. S. Ling, C. Rosé, and X. Wang, "Talk to me: Foundations for successful individual-group interactions in online communities," in *Proceedings of the SIGCHI conference on Human Factors in computing systems*, ACM, 2006, pp. 959–968.

[147] Y. Ren, F. M. Harper, S. Drenner, L. Terveen, S. Kiesler, J. Riedl, and R. E. Kraut, "Building member attachment in online communities: Applying theories of group identity and interpersonal bonds," *MIS Quarterly*, vol. 36, no. 3, pp. 841–864, 2012.

[148] C. Lampe and P. Resnick, "Slash (dot) and burn: Distributed moderation in a large online conversation space," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, ACM, 2004, pp. 543–550.

[149] J. Fox and M. C. Rooney, "The dark triad and trait self-objectification as predictors of men's use and self-presentation behaviors on social networking sites," *Personality and Individual Differences*, vol. 76, pp. 161–165, 2015.

[150] J. Suler, "The online disinhibition effect," *Cyberpsychology & behavior*, vol. 7, no. 3, pp. 321–326, 2004.

[151] L. Hall and C. E. George, "Law and punishment in virtual communities," *Proceedings of Cybersociety*, 1999.

[152] A. Bruckman, C. Danis, C. Lampe, J. Sternberg, and C. Waldron, "Managing deviant behavior in online communities," in *CHI'06 extended abstracts on Human factors in computing systems*, ACM, 2006, pp. 21–24.

[153] J. Cheng, C. Danescu-Niculescu-Mizil, and J. Leskovec, “Antisocial behavior in online discussion communities,” in *International Conference on Weblogs and Social Media (ICWSM)*, AAAI, 2015.

[154] J. Cheng, M. Bernstein, C. Danescu-Niculescu-Mizel, and J. Leskovic, “Anyone can become a troll: Causes of trolling behavior in online discussions,” in *Proceedings of the ACM Conference on Computer Supported Cooperative Work & Social Computing*, 2017.

[155] M. Samory and E. Peserico, “Sizing up the troll: A quantitative characterization of moderator-identified trolling in an online forum,” in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, ACM, 2017, pp. 6943–6947.

[156] E. Chandrasekharan, U. Pavalanathan, A. Srinivasan, A. Glynn, J. Eisenstein, and E. Gilbert, “You can’t stay here: The efficacy of reddit’s 2015 ban examined through hate speech,” in *Proceedings of the 2018 ACM conference on Computer supported cooperative work and social computing*, ACM, 2018.

[157] R. S. Geiger and D. Ribes, “The work of sustaining order in wikipedia: The banning of a vandal,” in *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, ACM, 2010, pp. 117–126.

[158] J. A. Pater, Y. Nadji, E. Mynatt, and A. S. Bruckman, “Just Awful Enough: The Functional Dysfunction of the Something Awful Forums,” *CHI*, pp. 2407–2410, 2014.

[159] G. Navarro, “A guided tour to approximate string matching,” *ACM computing surveys (CSUR)*, vol. 33, no. 1, pp. 31–88, 2001.

[160] E. B. Hekler, G. Dubey, D. W. McDonald, E. S. Poole, V. Li, and E. Eikey, “Exploring the relationship between changes in weight and utterances in an online weight loss forum: A content and correlational analysis study,” *Journal of medical Internet research*, vol. 16, no. 12, 2014.

[161] J. Harvey-Berino, S. Pintauro, P. Buzzell, E. C. Gold, S. Pintauro, P. Buzzell, and E. C. Gold, “Effect of internet support on the long-term maintenance of weight loss.,” *Obesity Research*, vol. 12, no. 2, pp. 320–329, 2004.

[162] M. Neve, P. J. Morgan, P. R. Jones, and C. E. Collins, “Effectiveness of web-based interventions in achieving weight loss and weight loss maintenance in overweight and obese adults: A systematic review with meta-analysis,” *Obesity Reviews*, vol. 11, no. 4, pp. 306–321, 2010.

[163] S. Saperstein, N. Atkinson, and R. Gold, “The impact of internet use for weight loss,” *Obesity reviews*, vol. 8, no. 5, pp. 459–465, 2007.

[164] R. L. Subreddit, *R/loseit*, `http://www.reddit.com/r/loseit`, [Online; accessed 9-September-2017], 2010.

[165] R. P. Subreddit, *R/proed*, `http://www.reddit.com/r/proED`, [Online; accessed 9-September-2017], 2015.

[166] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.

[167] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *International Conference on Machine Learning*, 2013.

[168] J. Firth, *A synopsis of linguistic theory 1930-1955, volume 1952-59. the philological society*, 1957.

[169] A. Bruckman, P. Curtis, C. Figallo, and B. Laurel, "Approaches to managing deviant behavior in virtual communities," in *Conference Companion on Human Factors in Computing Systems*, ACM, 1994, pp. 183–184.

[170] J. P. Davis, S. Farnham, and C. Jensen, "Decreasing online 'bad' behavior," in *CHI'02 Extended Abstracts on Human Factors in Computing Systems*, ACM, 2002, pp. 718–719.

[171] J. Sternberg, *Misbehavior in cyber places: The regulation of online conduct in virtual communities on the Internet*. Rowman & Littlefield, 2012.

[172] M. S. Bernstein, A. Monroy-Hernández, D. Harry, P. André, K. Panovich, and G. G. Vargas, "4chan and/b: An analysis of anonymity and ephemerality in a large online community.," 2011.

[173] E. Gilbert, "Widespread underprovision on reddit," in *Proceedings of the 2013 conference on Computer supported cooperative work*, ACM, 2013, pp. 803–808.

[174] A. Centivany and B. Glushko, "'popcorn tastes good': Participatory policymaking and reddit's 'amageddon''," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2016.

[175] D. A. Shamma, L. Kennedy, J. Li, B. Thomee, H. Jin, and J. Yuan, "Finding weather photos: Community-supervised methods for editorial curation of online sources," in *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, ACM, 2016, pp. 86–96.

[176] E. Agichtein, C. Castillo, D. Donato, A. Gionis, and G. Mishne, “Finding high-quality content in social media,” in *Proceedings of the 2008 international conference on web search and data mining*, ACM, 2008, pp. 183–194.

[177] N. A. Diakopoulos, “The editor’s eye: Curation and comment relevance on the new york times,” in *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, ACM, 2015, pp. 1153–1157.

[178] D. Park, S. Sachar, N. Diakopoulos, and N. Elmqvist, “Supporting comment moderators in identifying high quality online news comments,” in *Proc. Conference on Human Factors in Computing Systems (CHI)*, 2016.

[179] E. Wulczyn, N. Thain, and L. Dixon, “Ex machina: Personal attacks seen at scale,” in *Proceedings of the 26th International Conference on World Wide Web*, International World Wide Web Conferences Steering Committee, 2017, pp. 1391–1399.

[180] D. Soni and V. K. Singh, “See no evil, hear no evil: Audio-visual-textual cyberbullying detection,” *Proceedings of the ACM on Human-Computer Interaction*, vol. 2, no. CSCW, p. 164, 2018.

[181] N. Djuric, J. Zhou, R. Morris, M. Grbovic, V. Radosavljevic, and N. Bhamidipati, “Hate speech detection with comment embeddings,” in *Proceedings of the 24th international conference on world wide web*, ACM, 2015, pp. 29–30.

[182] M. ElSherief, V. Kulkarni, D. Nguyen, W. Y. Wang, and E. Belding, “Hate lingo: A target-based linguistic analysis of hate speech in social media,” in *Twelfth International AAAI Conference on Web and Social Media*, 2018.

[183] T. Davidson, D. Warmsley, M. Macy, and I. Weber, “Automated hate speech detection and the problem of offensive language,” in *Eleventh International AAAI Conference on Web and Social Media*, 2017.

[184] P. Garrigues, S. Farfade, H. Izadinia, K. Boakye, and Y. Kalantidis, “Tag prediction at flickr: A view from the darkroom,” *arXiv preprint arXiv:1612.01922*, 2016.

[185] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *NIPS*, 2012.

[186] J. Duchi, E. Hazan, and Y. Singer, “Adaptive subgradient methods for online learning and stochastic optimization,” *Journal of Machine Learning Research*, vol. 12, no. Jul, pp. 2121–2159, 2011.

[187] J. B. Martin, “The development of ideal body image perceptions in the united states,” *Nutrition Today*, vol. 45, no. 3, pp. 98–110, 2010.

[188] R. Kang, L. Dabbish, and K. Sutton, “Strangers on your phone: Why people use anonymous communication applications,” in *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, ACM, 2016, pp. 359–370.

[189] G. Rosen, *Getting our community help in real time*, 2017.

[190] D. Muriello, L. Donahue, D. Ben-David, U. Ozertem, and R. Shilon, *Under the hood: Suicide prevention tools powered by ai*, 2018.

[191] J. Vincent, *Facebook is using ai to spot users with suicidal thoughts and send them help*, 2017.

[192] D. Lee, *Samaritans pulls ’suicide watch’ radar app*, 2014.

[193] J. Barrie, *People are freaking out over this new anti-suicide twitter app*, 2014.

[194] Q. Cheng, T. M. H. Li, C.-L. L. Kwok, T. Zhu, and P. S. F. Yip, “Assessing suicide risk and emotional distress in Chinese social media: A text mining and machine learning study,” *Journal of Medical Internet Research*, vol. 19, no. 7, pp. 1–10, 2017.

[195] G. Coppersmith, K. Ngo, R. Leary, and A. Wood, “Exploratory analysis of social media prior to a suicide attempt,” in *CLPsych*, 2016, pp. 106–117.

[196] W. H. Organization, “Depression and other common mental disorders: Global health estimates,” 2017.

[197] J. Kleinberg, H. Lakkaraju, J. Leskovec, J. Ludwig, and S. Mullainathan, “Human decisions and machine predictions,” *The quarterly journal of economics*, vol. 133, no. 1, pp. 237–293, 2017.

[198] A. Datta, J Makagon, D. Mulligan, and M. Tschantz, “Discrimination in online personalization: A multidisciplinary inquiry.,” 2018.

[199] d. boyd and K. Crawford, “Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon,” *Information, communication & society*, vol. 15, no. 5, pp. 662–679, 2012.

[200] J. Metcalf and K. Crawford, “Where are human subjects in Big Data research? The emerging ethics divide,” *Big Data & Society*, vol. 3, no. 1, p. 205 395 171 665 021, 2016.

[201] T. Gillespie and N. Seaver, “Critical algorithm studies: A reading list,” *Social Media Collective*, 2016.

[202] J. M. Hudson and A. Bruckman, “”go away”: Participant objections to being studied and the ethics of chatroom research,” *The Information Society*, vol. 20, no. 2, pp. 127–139, 2004.

[203] E. Hargittai, “Is Bigger Always Better? Potential Biases of Big Data Derived from Social Network Sites,” *The ANNALS of the American Academy of Political and Social Science*, no. May, pp. 63–76, 2015.

[204] M. Zimmer, “but the data is already publici: On the ethics of research in facebook,” *Ethics and information technology*, vol. 12, no. 4, pp. 313–325, 2010.

[205] M. Zimmer and N. J. Proferes, “A topology of twitter research: Disciplines, methods, and ethics,” *Aslib Journal of Information Management*, vol. 66, no. 3, pp. 250–261, 2014.

[206] A. Olteanu, C. Castillo, F. Diaz, and E. Kiciman, “Social data: Biases, methodological pitfalls, and ethical boundaries,” 2016.

[207] M. J. Paul and M. Dredze, “Social monitoring for public health,” *Synthesis Lectures on Information Concepts, Retrieval, and Services*, vol. 9, no. 5, pp. 1–183, 2017.

[208] M. Conway and D. OConnor, “Social media, big data, and mental health: Current advances and ethical implications,” *Current opinion in psychology*, vol. 9, pp. 77–82, 2016.

[209] “Ethical Challenges of Big Data in Public Health,” *PLoS Computational Biology*, vol. 11, no. 2, pp. 1–7, 2015.

[210] E. Horvitz and D. Mulligan, “Data, privacy, and the greater good,” *Science*, vol. 349, no. 6245, pp. 253–255, 2015.

[211] C. Norval and T. Henderson, “Contextual consent: Ethical mining of social media for health research,” in *Proceedings of the WSDM 2017 Workshop on Mining Online Health Reports*, 2017.

[212] J. Mikal, S. Hurst, and M. Conway, “Ethical issues in using twitter for population-level depression monitoring: A qualitative study,” *BMC medical ethics*, vol. 17, no. 1, p. 22, 2016.

[213] A. Benton, G. Coppersmith, and M. Dredze, “Ethical research protocols for social media health research,” *EACL 2017*, p. 94, 2017.

[214] D. Lazer, R. Kennedy, G. King, and A. Vespignani, “The Parable of Google Flu: Traps in Big Data Analysis,” *Science*, vol. 343, no. 6167, pp. 1203–1205, 2014.

[215] B. F. Welles, “On minorities and outliers: The case for making big data small,” *Big Data & Society*, vol. 1, no. 1, p. 2 053 951 714 540 613, 2014.

[216] M. Conway, “Ethical issues in using twitter for public health surveillance and research: Developing a taxonomy of ethical concepts from the research literature,” *Journal of medical Internet research*, vol. 16, no. 12, 2014.

[217] N. C. for the Protection of Human Subjects of Biomedicaland Behavioral Research, *The Belmont report: Ethical principles and guidelines for the protection of human subjects of research*. Superintendent of Documents, 1978.

[218] *Summary of the hipaa security rule*, 2018.

[219] *International compilation of human research standards*, 2018.

[220] E. J. Emanuel, D. Wendler, and C. Grady, “What makes clinical research ethical?” *Jama*, vol. 283, no. 20, pp. 2701–2711, 2000.

[221] P. Corrigan, “How stigma interferes with mental health care.,” *American psychologist*, vol. 59, no. 7, p. 614, 2004.

[222] A. D. Kramer, J. E. Guillory, and J. T. Hancock, “Experimental evidence of massive-scale emotional contagion through social networks,” *Proceedings of the National Academy of Sciences*, vol. 111, no. 24, pp. 8788–8790, 2014.

[223] C. Fiesler and N. Proferes, “”participant” perceptions of twitter research ethics,” *Social Media+ Society*, vol. 4, no. 1, 2018.

[224] D. of Health and H. Services, *Vulnerable populations*, 2018.

[225] N. I. of Mental Health, *Eating disorders*, 2018.

[226] H. A. Schwartz, J. C. Eichstaedt, M. L. Kern, L. Dziurzynski, S. M. Ramones, M. Agrawal, A. Shah, M. Kosinski, D. Stillwell, M. E. Seligman, *et al.*, “Personality, gender, and age in the language of social media: The open-vocabulary approach,” *PloS one*, vol. 8, no. 9, e73791, 2013.

[227] H. Nissenbaum, “Privacy as contextual integrity,” *Wash. L. Rev.*, vol. 79, p. 119, 2004.

[228] M. Zimmer, “Addressing Conceptual Gaps in Big Data Research Ethics: An Application of Contextual Integrity,” *Social Media + Society*, vol. 4, no. 2, 2018.

[229] A. Bruckman, “Studying the amateur artist: A perspective on disguising data collected in human subjects research on the Internet,” *Ethics and Information Technology*, vol. 4, no. 3, p. 217, 2002.

[230] T. Wang, M. Brede, A. Ianni, and E. Mentzakis, “Detecting and Characterizing Eating-Disorder Communities on Social Media,” in *WSDM*, ser. WSDM ’17, ACM, 2017, pp. 91–100.

[231] J. M. Hofman, A. Sharma, and D. J. Watts, “Prediction and explanation in social systems,” *Science*, vol. 355, no. 6324, pp. 486–488, 2017.

[232] G. Coppersmith, M. Dredze, C. Harman, Holli, and K. Hollingshead, “From ADHD to SAD: Analyzing the language of mental health on Twitter through self-reported diagnoses,” in *CLPsych*, 2015, pp. 1–10.

[233] J. Burrell, “How the machine ’thinks’: Understanding opacity in machine learning algorithms,” *Big Data & Society*, vol. 3, no. 1, p. 2 053 951 715 622 512, 2016.

[234] Z. Jamil, D. Inkpen, P. Buddhitha, and K. White, “Monitoring Tweets for Depression to Detect At-risk Users,” in *CLPsych*, 2017, pp. 32–40.

[235] J. W. Ayers, T. L. Caputi, C. Nebeker, and M. Dredze, “Don’t quote me: reverse identification of research participants in social media studies,” *npj Digital Medicine*, vol. 1, no. 1, p. 30, 2018.

[236] W. Moncur, “The emotional wellbeing of researchers: Considerations for practice,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, 2013, pp. 1883–1890.

[237] D. Hovy and S. L. Spruit, “The social impact of natural language processing,” in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, vol. 2, 2016, pp. 591–598.

[238] A. Wongkoblap, M. A. Vadillo, and V. Curcin, “Researching mental health disorders in the era of social media: Systematic review,” *Journal of medical Internet research*, vol. 19, no. 6, 2017.

[239] C. DiSalvo, P. Sengers, and H. Brynjarsdóttir, “Mapping the landscape of sustainable hci,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, 2010, pp. 1975–1984.

[240] A. Kittur, J. V. Nickerson, M. Bernstein, E. Gerber, A. Shaw, J. Zimmerman, M. Lease, and J. Horton, “The future of crowd work,” in *Proceedings of the 2013 conference on Computer supported cooperative work*, ACM, 2013, pp. 1301–1318.

[241] T. R. Dillahunt, X. Wang, E. Wheeler, H. F. Cheng, B. J. Hecht, and H. Zhu, "The sharing economy in computing: A systematic literature review.," *PACMHCI*, vol. 1, no. CSCW, pp. 38–1, 2017.

[242] E. P. S. Baumer, "Reflective Informatics: Conceptual Dimensions for Designing Technologies of Reflection," in *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*, Seoul: ACM, 2015, pp. 585–594, ISBN: 978-1-4503-3145-6.

[243] G. Bell and P. Dourish, "Yesterdays tomorrows: Notes on ubiquitous computings dominant vision," *Personal and ubiquitous computing*, vol. 11, no. 2, pp. 133–143, 2007.

[244] A. Liberati, D. G. Altman, J. Tetzlaff, C. Mulrow, P. C. Gøtzsche, J. P. Ioannidis, M. Clarke, P. J. Devereaux, J. Kleijnen, and D. Moher, "The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: Explanation and elaboration," *PLoS Medicine*, vol. 6, no. 7, 2009.

[245] C. DiSalvo, P. Sengers, and H. Brynjarsdóttir, "Mapping the landscape of sustainable HCI," in *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*, Atlanta, GA: ACM, 2010, pp. 1975–1984.

[246] d. boyd and N. B. Ellison, "Social network sites: Definition, history, and scholarship," *Journal of Computer-Mediated Communication*, vol. 13, no. 1, pp. 210–230, 2007.

[247] E. M. Seabrook, M. L. Kern, and N. S. Rickard, "Social networking sites, depression, and anxiety: A systematic review," *JMIR mental health*, vol. 3, no. 4, e50, 2016.

[248] A.-W. Harzing *et al.*, "Publish or perish," 2007.

[249] S. K. Ernala, M. L. Birnbaum, K. A. Candan, A. F. Rizvi, W. A. Sterling, J. M. Kane, and M. De Choudhury, "Methodological Gaps in Predicting Mental Health States from Social Media : Triangulating Diagnostic Signals," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019, ISBN: 9781450359702.

[250] V. Braun and V. Clarke, "Using thematic analysis in psychology," *Qualitative research in psychology*, vol. 3, no. 2, pp. 77–101, 2006.

[251] Y.-c. Shen, T.-t. Kuo, I.-n. Yeh, T.-t. Chen, and S.-d. Lin, "Exploiting Temporal Information in a Two-Stage Classification Framework for Content-Based Depression," *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 276–288, 2013.

[252] N. Masuda, I. Kurahashi, and H. Onari, “Suicide Ideation of Individuals in Online Social Networks,” *PLOS ONE*, vol. 8, no. 4, 2013.

[253] F. Tsugawa, S., Mogi, Y., Kikuchi, Y., Kishino, F. and H. K., Itoh, Y., and Ohsaki, “On estimating depressive tendency of twitter users from their tweet data,” *Proceedings of the 2nd International Workshop on Ambient Information Technologies (AMBIT’12)*, vol. 2, pp. 29–32, 2013.

[254] M. De Choudhury, S. Counts, and E. Horvitz, “Social Media As a Measurement Tool of Depression in Populations,” in *WebSci*, ser. WebSci ’13, ACM, 2013, pp. 47–56.

[255] N. Vedula and S. Parthasarathy, “Emotional and Linguistic Cues of Depression from Social Media,” in *DH*, ACM, 2017, pp. 127–136.

[256] K. Loveys, P. Crutchley, E. Wyatt, and G. Coppersmith, “Small but Mighty: Affective Micropatterns for Quantifying Mental Health from Social Media Language,” in *CLPsych*, 2017, pp. 85–95.

[257] A. Benton, M. Mitchell, and D. Hovy, “Multitask learning for mental health conditions with limited social media data,” *EACL*, 2017.

[258] Y Zhou, J Zhan, and J Luo, “Predicting Multiple Risky Behaviors via Multimedia Content,” *International Conference on Social Informatics*, 2017.

[259] K. Saha and M. De Choudhury, “Modeling Stress with Social Media Around Incidents of Gun Violence on College Campuses,” *Proc. ACM Hum.-Comput. Interact.*, vol. 1, no. CSCW, 92:1–92:27, 2017.

[260] A. G. Reece, A. J. Reagan, K. L. M. Lix, P. S. Dodds, C. M. Danforth, and E. J. Langer, “Forecasting the onset and course of mental illness with Twitter data,” *SCIENTIFIC REPORTS*, vol. 7, no. 1, 2017.

[261] G Shen, J Jia, L Nie, F Feng, and ..., “Depression detection via harvesting social media: A multimodal dictionary learning solution,” *IJCAI*, 2017.

[262] E. Saravia, C. H. Chang, R. J. De Lorenzo, and Y. S. Chen, “MIDAS: Mental illness detection and analysis via social media,” in *ASONAM*, ser. ASONAM ’16, IEEE Press, 2016, pp. 1418–1421.

[263] S. R. Braithwaite, C. Giraud-Carrier, J. West, M. D. Barnes, and C. L. Hanson, “Validating Machine Learning Algorithms for Twitter Data Against Established Measures of Suicidality,” *JMIR Mental Health*, vol. 3, no. 2, e21, 2016.

[264] X. Huang, L. Zhang, D. Chiu, T. Liu, X. Li, and T. Zhu, "Detecting Suicidal Ideation in Chinese Microblogs with Psychological Lexicons," *2014 IEEE International Conference on Autonomic and Trusted Computing, 2014 IEEE International Conference on Scalable Computing and Communications and Associated Symposia/Workshops, UIC-ATC-ScalCom 2014*, vol. 2014, pp. 844–849, 2014.

[265] L. Zhang, X. Huang, T. Liu, A. Li, Z. Chen, and T. Zhu, "Using Linguistic Features to Estimate Suicide Probability of Chinese Microblog Users," *ICHI*, vol. 8944, pp. 549–559, 2015.

[266] H Lin, J Jia, L Nie, G Shen, and T. S. Chua, "What Does Social Media Say about Your Stress?.," *IJCAI*, 2016.

[267] L. Zhao, J. Jia, and L. Feng, "Teenagers' stress detection based on time-sensitive micro-blog com- ment/response actions," in *IFIP International Conference on Artificial Intelligence in Theory and Practice*, 2015, pp. 26–36.

[268] T. Nguyen, D. Phung, B. Dao, S. Venkatesh, and M. Berk, "Affective and content analysis of online depression communities," *Ieee Transactions on Affective Computing*, vol. 5, no. 3, pp. 217–226, 2014.

[269] V. M. Prieto, S. Matos, M. Alvarez, F. Cacheda, and J. L. Oliveira, "Twitter: A Good Place to Detect Health Conditions," *PLOS One*, vol. 9, no. 1, 2014.

[270] D. Preotiuc-Pietro, J. Eichstaedt, G. Park, M. Sap, L. Smith, V. Tobolsky, H. A. Schwartz, and L. Ungar, "The Role of Personality , Age and Gender in Tweeting about Mental Illnesses," *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, no. January, pp. 21–30, 2015.

[271] P. Resnik, W. Armstrong, L. Claudino, T. Nguyen, V.-a. Nguyen, and J. Boyd-Graber, "Beyond LDA: exploring supervised topic modeling for depression-related language in Twitter," in *CLPsych*, vol. 1, 2015, pp. 99–107.

[272] C. M. Homan, R. Johar, T. Liu, M. Lytle, V. Silenzio, and C. O. Alm, "Toward Macro-Insights for Suicide Prevention: Analyzing Fine-Grained Distress at Scale," in *CLPsych*, 2014, p. 107, ISBN: 9781941643167.

[273] B. O'Dea, S. Wan, P. J. Batterham, A. L. Calear, C. Paris, and H. Christensen, "Detecting suicidality on twitter," *Internet Interventions*, vol. 2, no. 2, pp. 183–188, 2015.

[274] P. Burnap, W. Colombo, and J. Scourfield, "Machine Classification and Analysis of Suicide-Related Communication on Twitter," in *HT*, ser. HT '15, ACM, 2015, pp. 75–84.

[275] K McManus, E. K. Mallory, R. L. Goldfeder, W. A. Haynes, and J. D. Tatum, "Mining Twitter Data to Improve Detection of Schizophrenia," *AMIA*, vol. 2015, pp. 122–126, 2015.

[276] W. Luo, D. Phung, T. Tran, S. Gupta, S. Rana, C. Karmakar, A. Shilton, J. Yearwood, N. Dimitrova, T. B. Ho, *et al.*, "Guidelines for developing and reporting machine learning predictive models in biomedical research: A multidisciplinary view," *Journal of medical Internet research*, vol. 18, no. 12, e323, 2016.

[277] d. boyd and K. Crawford, "Critical Questions for Big Data," *Information, Communication & Society*, vol. 15, no. 5, pp. 662–679, 2012.

[278] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*, Ieee, 2009, pp. 248–255.

[279] D. Lazer, A. Pentland, L. Adamic, S. Aral, A.-L. Barabási, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, *et al.*, "Computational social science," *Science*, vol. 323, no. 5915, pp. 721–723, 2009.

[280] E. P. S. Baumer, "Usees," in *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*, Seoul: ACM Press, 2015, pp. 3295–3298, ISBN: 978-1-4503-3145-6.

[281] E. P. S. Baumer and J. R. Brubaker, "Post-userism," in *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*, Denver, CO: ACM, 2017, pp. 6291–6303, ISBN: 978-1-4503-4655-9.

[282] Z. Tufekci, "Engineering the public: Big data, surveillance and computational politics," *First Monday*, vol. 19, no. 7, 2014.

[283] B. G. Link and J. C. Phelan, "Stigma and its public health implications," *The Lancet*, vol. 367, no. 9509, pp. 528–529, 2006.

[284] S. Barocas and A. D. Selbst, "Big data's disparate impact," *Calif. L. Rev.*, vol. 104, p. 671, 2016.

[285] S. U. Noble, *Algorithms of oppression: How search engines reinforce racism*. NYU Press, 2018.

[286] O. Keyes, "The misgendering machines: Trans/hci implications of automatic gender recognition," *Proceedings of the ACM on Human-Computer Interaction*, vol. 2, no. CSCW, p. 88, 2018.

[287] C. D'Ignazio and L. Klein, "Chapter One: Bring Back the Bodies," *MIT Press Open*, Jan. 14, 2019, https://bookbook.pubpub.org/pub/zrlj0jqb.

[288] G. Bell and P. Dourish, "Yesterday's tomorrows: Notes on ubiquitous computing's dominant vision," *Personal and Ubiquitous Computing*, vol. 11, no. 2, pp. 133–143, Jan. 2007.

[289] J. P. Gee, *An introduction to discourse analysis: Theory and method*. Routledge, 2011.

[290] M. Foucault, *The archaeology of knowledge: Translated from the french by AM Sheridan Smith*. Pantheon Books, 1972.

[291] V. Namaste, *Invisible lives: The erasure of transsexual and transgendered people*. University of Chicago Press, 2000.

[292] L. C. Irani and M. S. Silberman, "Stories We Tell About Labor: Turkopticon and the Trouble with "Design"," in *CHI 2016*, 2016, pp. 4573–4586, ISBN: 9781450333627.

[293] J. Butler, "Doing justice to someone: Sex reassignment and allegories of transsexuality," *GLQ: A Journal of Lesbian and Gay Studies*, vol. 7, no. 4, pp. 621–636, 2001.

[294] M. Callon, "Some Elements of a Sociology of Translation: Domestication of the Scallops and the Fishermen of Saint Brieuc Bay," in *Power, Action and Belief: A New Sociology of Knowledge?* Ser. Sociological Review Monograph 32, J. Law, Ed., London: Routledge, 1986, pp. 196–223.

[295] B. Latour, "Ethnography of a high-tech case," *Technological Choices: transformation in material cultures since the neolithic*, pp. 372–98, 1993.

[296] G. Kannabiran, J. Bardzell, and S. Bardzell, "How hci talks about sexuality: Discursive strategies, blind spots, and opportunities for future research," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, 2011, pp. 695–704.

[297] A. L. Hoffmann, "Where fairness fails: Data, algorithms, and the limits of antidiscrimination discourse," *Information, Communication, and Society (forthcoming)*, 201p.

[298] N. M. Su, L. S. Liu, and A. Lazar, "Mundanely miraculous: The robot in healthcare," in *Proceedings of the 8th Nordic Conference on Human-Computer Interaction: Fun, Fast, Foundational*, ACM, 2014, pp. 391–400.

[299] E. Harmon and M. Mazmanian, "Stories of the smartphone in everyday discourse: Conflict, tension & instability," in *Proceedings of the SIGCHI conference on human factors in computing systems*, ACM, 2013, pp. 1051–1060.

[300] A. L. Hoffmann, N. Proferes, and M. Zimmer, ""Making the world more open and connected": Mark Zuckerberg and the discursive construction of Facebook and its users," *New Media and Society*, vol. 20, no. 1, pp. 199–218, 2018.

[301] A. Schlesinger, W. K. Edwards, and R. E. Grinter, "Intersectional hci: Engaging identity through gender, race, and class," in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, ACM, 2017, pp. 5412–5427.

[302] M. C. Nisbet and C. Mooney, "Framing science," *Science*, vol. 316, no. 5821, pp. 56–56, 2007.

[303] H. H. Bauer, "Barriers against interdisciplinarity: Implications for studies of science, technology, and society (sts," *Science, Technology, & Human Values*, vol. 15, no. 1, pp. 105–119, 1990.

[304] A. Barry, G. Born, and G. Weszkalnys, "Logics of interdisciplinarity," *Economy and Society*, vol. 37, no. 1, pp. 20–49, 2008.

[305] T. Van Leeuwen, "Three models of interdisciplinarity," *A new agenda in (critical) discourse analysis: Theory, methodology and interdisciplinarity*, pp. 3–18, 2005.

[306] J. Drescher, P. Cohen-Kettenis, and S. Winter, "Minding the body: Situating gender identity diagnoses in the icd-11," *International Review of Psychiatry*, vol. 24, no. 6, pp. 568–577, 2012.

[307] E. Martin, "The egg and the sperm: How science has constructed a romance based on stereotypical male-female roles," *Signs: Journal of Women in Culture and Society*, vol. 16, no. 3, pp. 485–501, 1991.

[308] E. Goffman, *Stigma: Notes on the management of spoiled identity*. Simon and Schuster, 2009.

[309] J. Arboleda-Flórez and H. Stuart, "From sin to science: Fighting the stigmatization of mental illnesses," *Canadian Journal of Psychiatry*, vol. 57, no. 8, pp. 457–463, 2012.

[310] N. Rüsch, M. C. Angermeyer, and P. W. Corrigan, "Mental illness stigma: Concepts, consequences, and initiatives to reduce stigma," *European psychiatry*, vol. 20, no. 8, pp. 529–539, 2005.

[311] M. L. Hatzenbuehler, A. Bellatorre, Y. Lee, B. K. Finch, P. Muennig, and K. Fiscella, “Structural stigma and all-cause mortality in sexual minority populations,” *Social Science & Medicine*, vol. 103, pp. 33–41, 2014.

[312] L. Greenstein, *Why suicide reporting guidelines matter*.

[313] K. Savchuk, *5 tips for journalists covering mental and behavioral health*.

[314] J. A. Clausen, “Stigma and mental disorder: Phenomena and terminology,” *Psychiatry*, vol. 44, no. 4, pp. 287–296, 1981.

[315] L. J. Bracken and E. A. Oughton, “what do you mean?the importance of language in developing interdisciplinary research,” *Transactions of the Institute of British Geographers*, vol. 31, no. 3, pp. 371–382, 2006.

[316] J. Vitak, K. Shilton, and Z. Ashktorab, “Beyond the belmont principles: Ethical challenges, practices, and beliefs in the online data research community,” in *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, ACM, 2016, pp. 941–953.

[317] B. Resnick, “Researchers just released profile data on 70,000 okcupid users without permission,” *Vox. Retreived from: https://www. vox. com/2016/5/12/11666116/70000-okcupid-users-data-release*, 2016.

[318] S. L. Star and J. R. Griesemer, “Institutional ecology,translations’ and boundary objects: Amateurs and professionals in berkeley’s museum of vertebrate zoology, 1907-39,” *Social studies of science*, vol. 19, no. 3, pp. 387–420, 1989.

[319] B. Hammarfelt, F. Åström, and J. Hansson, “Scientific publications as boundary objects: Theorising the intersection of classification and research evaluation,” in *Information research*, vol. 22, 2017.

[320] J. R. Brubaker and G. R. Hayes, “Select* from user: Infrastructure and socio-technical representation,” in *Proceedings of the ACM 2011 conference on Computer supported cooperative work*, ACM, 2011, pp. 369–378.

[321] P. Klasnja and W. Pratt, “Healthcare in the pocket: Mapping the space of mobile-phone health interventions,” *Journal of biomedical informatics*, vol. 45, no. 1, pp. 184–198, 2012.

[322] J. Ma, L. K. Saul, S. Savage, and G. M. Voelker, “Identifying suspicious urls: An application of large-scale online learning,” in *Proceedings of the 26th annual international conference on machine learning*, ACM, 2009, pp. 681–688.

[323] N. Andalibi, P. Ozturk, and A. Forte, “Sensitive self-disclosures, responses, and social support on instagram: The case of #depression,” in *Proceedings of the ACM Conference on Computer Supported Cooperative Work & Social Computing*, Forthcoming, 2017.

[324] K. Carpenter and D. Dittrich, “Bridging the distance: Removing the technology buffer and seeking consistent ethical analysis in computer security research,” in *1st International Digital Ethics Symposium. Loyola University Chicago Center for Digital Ethics and Policy*, 2011.

[325] R. J. Lawrence and C. Després, “Futures of transdisciplinarity,” *Futures*, vol. 4, no. 36, pp. 397–405, 2004.

[326] J. L. Feuston and A. M. Piper, “Everyday experiences: Small stories and mental illness on instagram,” in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, ser. CHI ’19, New York, NY, USA: ACM, 2019, 265:1–265:14, ISBN: 978-1-4503-5970-2.

[327] S. L. Blodgett and B. O’Connor, “Racial disparity in natural language processing: A case study of social media african-american english,” *arXiv preprint arXiv:1707.00061*, 2017.

[328] B. Hecht, L. Wilcox, J. Bigham, J. Schoning, E. Hoque, J. Ernst, Y. Bisk, L. De Russis, L. Yarosh, B. Anjum, D. Contractor, and C. Wu, *It’s time to do something: Mitigating the negative impacts of computing through a change to the peer review process*, Mar. 2018.

[329] K. Hao, *‘‘ai is sending people to jailand getting it wrong”*, 2019.

[330] A. D. Selbst, S. Friedler, S. Venkatasubramanian, J. Vertesi, *et al.*, “Fairness and abstraction in sociotechnical systems,” in *ACM Conference on Fairness, Accountability, and Transparency (FAT*)*, 2019.

[331] D. McQuillan, “Data Science as Machinic Neoplatonism,” *Philosophy and Technology*, vol. 31, no. 2, pp. 253–272, 2018.

[332] B. Green and L. Hu, “The myth in the methodology: Towards a recontextualization of fairness in machine learning,” Proceedings of the Machine Learning: The Debates Workshop, 2018.

[333] E. P. Baumer and M Silberman, “When the implication is not to design (technology),” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, 2011, pp. 2271–2274.