Contextual Gaps in Machine Learning for Mental Illness Prediction: The Case of Diagnostic Disclosures

STEVIE CHANCELLOR, University of Minnesota, USA
JESSICA L. FEUSTON, University of Colorado Boulder, CO
JAYHYUN CHANG, Northwestern University, IL

Getting training data for machine learning (ML) prediction of mental illness on social media data is labor intensive. To work around this, ML teams will extrapolate proxy signals, or alternative signs from data to evaluate illness status and create training datasets. However, these signals' validity has not been determined, whether signals align with important contextual factors, and how proxy quality impacts downstream model integrity. We use ML and qualitative methods to evaluate whether a popular proxy signal, diagnostic self-disclosure, produces a conceptually sound ML model of mental illness. Our findings identify major conceptual errors only seen through a qualitative investigation – training data built from diagnostic disclosures encodes a narrow vision of diagnosis experiences that propagates into paradoxes in the downstream ML model. This gap is obscured by strong performance of the ML classifier (F1 = 0.91). We discuss the implications of conceptual gaps in creating training data for human-centered models, and make suggestions for improving research methods.

CCS Concepts: • **Human-centered computing** → **Collaborative and social computing design and evaluation methods**; **Empirical studies in collaborative and social computing**; • **Computing methodologies** → *Machine learning approaches.*

Additional Key Words and Phrases: social media, error analysis, Reddit, mental health, validity

ACM Reference Format:
Stevie Chancellor, Jessica L. Feuston, and Jayhyun Chang. 2023. Contextual Gaps in Machine Learning for Mental Illness Prediction: The Case of Diagnostic Disclosures. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW2, Article 332 (October 2023), 27 pages. https://doi.org/10.1145/3610181

1 INTRODUCTION

The use of machine learning (ML) to detect mental illness on social networks is a challenging and high-stakes research area. Significant improvements in this area have been made in the last decade of research [90], indicating that language signals can be used to predict experiences of mental illness, including depression [34], post-traumatic stress disorder [26], and other conditions and symptoms [18, 90]. Real systems use ML to predict aspects of mental health status – Facebook deploys ML models that predict when someone may discuss suicidal ideation on their platform [31]. Many companies use account holder information to target mental health advertisements [29]. If widely and accurately deployed, these models could facilitate early detection of mental illness and mitigate barriers to in-person clinical care [17, 90], hopefully reducing the average 11-year gap between the onset of mental health symptoms and treatment in the US [91].

Authors' addresses: Stevie Chancellor, University of Minnesota, Minneapolis, USA, steviec@umn.edu; Jessica L. Feuston, University of Colorado Boulder, Boulder, CO, jesfeuston@meta.com; Jayhyun Chang, Northwestern University, Evanston, IL, jayhyunchang2022@u.northwestern.edu.

2573-0142/2023/10-ART332 $15.00
https://doi.org/10.1145/3610181

A major challenge in this research is obtaining high-quality training data. Social media data is rich with information about peoples' experiences and thoughts; however, social media does not have explicit metadata or information about mental health that can be used for models. Therefore, research teams design *proxy signals* for mental illness status, or alternative/implicit indicators of mental illness in the absence of clinical evaluation [18, 39]. For example, researchers will use hashtags like #anorexic to infer that someone has anorexia. Many proxy signals can be automatically extracted from social media posts with minimal resources, requiring less capital and labor than recruiting participants.

If ML systems are pushed to solve long-standing gaps in mental illness care, diagnosis, and treatment, it is crucial to evaluate whether proxy signals lead to models that align with correct conceptualizations of the community and mental health. Most proxy signals are not externally validated against psychological theory, against the human-focused experiences of the account holders they predict, or against the contexts in which they are deployed [18]. Computational research has challenged the validity of proxy signals as representative experiences of mental illness [1, 2, 39, 43] as well as their use in emotional recognition tasks [30, 62]. Incorrect assumptions by researchers about proxy signals harm the conceptual integrity of modeling because proxy signals may rely on unsubstantiated "shortcuts" [50] to create machine representations. Furthermore, poor proxy signals risk creating *contextual* errors when labels and downstream models do not capture the experiences of people with mental illnesses [43, 45]. To correctly and compassionately build models that assess well-being, we must deeply interrogate the models and their components than blindly trust prior work.

In this paper, we conduct an evaluation of the quality of a popular proxy signal used to build training datasets – diagnostic self-disclosure – and whether it produces a conceptually sound model of mental illness in an online eating disorder (ED) community. Diagnostic self-disclosures are personal statements of clinical diagnoses on social media (*e.g.* "I was diagnosed with anxiety") [24, 26]. This signal is one of the most popular proxies, adopted in prediction tasks for various mental illnesses and symptoms (e.g. [10, 24, 65]) because people believe that they are honest about self-disclosing their health on social media sites [40, 68]. We study a social media eating disorder (ED) community as a case because these communities are often targeted by ML platform interventions [21]. Therefore, ED communities are uniquely at risk for errors in ML algorithms leading to adverse outcomes, like banning and exclusion from social support [21, 45, 51]. In short, diagnostic disclosures of EDs are an excellent case study of how proxy signals are leveraged for ML modeling given their closeness to "diagnostic assessment" and platform interventions.

To conduct this investigation, we use a mixed-methods approach to assess diagnostic disclosures for training dataset construction and their impacts on models – first to evaluate classic error measures of task design and performance, then qualitatively assess the contextual quality of the model. We do this evaluation building on and replicating seminal work in the space, both in the design of regular expressions for diagnostic disclosure detection [26] as well as standard ML practice in the area [18, 90]. We evaluate the conceptual quality of the training data and the resulting ML model with three methods: 1) a performance evaluation of the regular expression in denotative meanings; 2) an ML experiment and evaluation; and 3) a qualitative error analysis technique that identifies contextual gaps called *contextual error analysis*.

Our results identify fundamental *contextual gaps* in models built on diagnostic disclosure data for ED identification. We define a contextual gap as where the ML modeling process does not capture sufficient context [8] to make a model that is valid and sensitive to its deployment details. Only 1% of active account holders (285 total) disclosed a diagnosis of an ED in our community. Our contextual error analysis shows that training data encodes language around past clinical history as highly relevant to *future* diagnosis, leaking paradoxical context to the model and compromising its validity.

Said metaphorically, if a clinician operated similarly to the ML model, they would diagnose people with an ED only after someone mentioned that they had previously seen a doctor and received a diagnosis. Our contextual error analysis confirms Feuston and colleagues' prior findings [43] that the training data is contextually compromised with assumed temporal rigidity of diagnosis and ignoring clinically-grounded signs of distress. However, in our model, these contextual gaps are obscured by strong performance of traditional error analysis in the regular expression task (F1 = 0.85) and machine learning classifiers (F1 = 0.91).

This paper contributes several tools and analytical perspectives to HCI and CSCW to facilitate audit work of proxy signals for applied ML modeling. Empirically, this model's quantitative "success" obscures a mismatch between the model's intentions and actual predictive capabilities, raising questions about the credibility and internal consistency of prior and future work based on the definitions we operationalize in our study. Our contextual error analysis technique is a novel method to evaluate the success and trade-offs of proxy signals in the design of ML applications. Joining complementary methods work [1, 2, 39, 54], we question the generalizability and representativeness of proxies for the development of training data. Connecting with STS and informatics scholarship [36, 45, 83], we critique how the classification of disorder is enacted by rigid computing practices [13, 43]. Researchers and designers must prioritize training data's integrity and models' conceptual foundations before deployment. Finally, we provide pragmatic alternatives to creating situated participatory datasets and more human-centered models.

2 RELATED WORK

2.1 Gathering Training Data from Social Media Data and its Criticisms

Gathering high-quality labels is both a pervasive need and a difficult challenge for developing ML models. In social media data, exogenous labels of mental illness status do not exist and, therefore, must be generated for use in supervised ML contexts. High-quality, or "gold standard" data acquisition in this domain typically comes from prior clinical history or clinical evaluation. Gathering clinical labels is tricky because of time, monetary costs, and the perceived barriers of ethics board approval to work with health information.

Instead of clinical labels for social media data, study teams have turned to proxy diagnostic signals as alternative signs of illness, which we call proxy signals in the remainder of this paper [18, 39]. Ernala et al. [39] define proxy signals as "binary indicators of the presence or absence of these social media behaviors that might correspond to their clinical mental health state." Several proxy labeling strategies exist, including affiliation and community participation markers (such as hashtags on posts), collaboration with clinical partners to identify well-being characteristics, or hybrid approaches such as human annotation on smaller data sets or crowdworker annotation [18]. The choice of proxy label depends on platform affordances, the amount of data needed, and the problem task.

Despite the prevalence of proxy labels in this space, most signals have never been validated for their validity [1, 2, 18, 39, 54]. Most proxy signals are not externally validated against psychological theory or with clinical information. These validity issues have downstream impacts on the representativeness of the models built from them [1, 2, 54], causing demographic biases in mental health prediction samples [43, 74, 75], and gaps in modeling between patients and users who self-disclose [39, 54]. Additionally, these proxy signals are not cross-validated against individual experiences or preferences [43, 44]. In this work, we used a mixed methods approach to study the validity of proxy signals and their downstream impacts in the ML lifecycle.

2.2 Eating Disorder Communities in Social Media

Eating disorders (EDs) are psychological disorders characterized by abnormal/disturbed eating and/or exercise habits. This includes conditions such as anorexia nervosa, bulimia nervosa, and binge eating disorder [38]. This paper acknowledges that EDs are more than their classification as a disorder. For people living with EDs, EDs encompass many non-normative and, at times, risky experiences with how people relate to and maintain their bodies, including eating, exercising, making posts on social media about their experiences, and seeking advice to maintain or prolong behaviors.

Many online communities host discussions about EDs. Previous work focuses on the composition and dynamics of discourse, support, and communication in ED communities. The research of [44] has highlighted the importance of support and care that these communities provide without clinical or emotional support from other parts of peoples' lives [44]. For example, Pater et al. [75] has studied visual expressions in these communities [74]. Recent work has sought to fill gaps in framing these communities as identity spaces and negotiated acceptance. For example, Pater et al. [75] have studied men and ED communities and Feuston et al. [42] have examined how trans people with EDs experience and navigate marginality in online ED spaces. ED communities are sometimes labeled as pro-ED (or promoting EDs) or pro-ana (promoting anorexia), which at their extremes advocate for EDs as a lifestyle choice [20, 88]. However, not all posts on EDs violate social media platform guidelines, and many communities may be mistakenly labeled dangerous when, in fact, they share advice about support and recovery [43, 44, 74]. Additionally, recent work highlights how people with EDs may conceptualize pro-ED differently than academic researchers – participants described pro-ED as pro-people with eating disorders rather than as promoting EDs [45].

Posts on social media are often used to predict account holders who may have an eating disorder [9, 19, 25, 32, 92]. This has been done across social platforms such as Twitter [25, 92], Tumblr [20, 33], and Reddit [19]. Across these studies, researchers use proxy signals to assess the status of mental illness, including hashtag use [20], self-disclosure [25], and community participation [19, 33]. The self-stated motivations of this research are to reach vulnerable populations *before* they are diagnosed to decrease the time to treatment and provide support resources. However, ED communities are uniquely at risk for ML technologies to impact them because of intense scrutiny of their behaviors, the focus on content removal, and banning efforts across platforms [21, 45, 51]. We join prior work in evaluating training data's impacts on ML classifiers' results and, subsequently, on our understanding of ED communities.

2.3 Identifying Error in Machine Learning

Error in machine learning is the deviation of predicted values produced by the ML model from their actual states. These deviations can be measured and evaluated – in classification, this is typically done by studying performance metrics like accuracy, F1, or AUC. However, researchers and practitioners like Bellotti and Edwards have realized that technology cannot capture all context and, therefore, humans must be involved [8]. This section focuses on how related areas of ML and HCI conceptualize errors and how to solve them.

Interfaces for Interactive ML Debugging. HCI has long known that interaction is crucial to finding context, and toolkits/interfaces have become a popular source of research to assist people in finding errors [4, 41]. For example, Wu et al. [93] developed Errudite, a hypothesis-driven error analysis toolkit that allows counterfactual testing of errors across the entire dataset (rather than a small part). Amershi et al. [5] developed ModelTracker, an iterative visualization tool to assist in performance analysis. Similarly, Ren et al. [78] proposed Squares, which helps with multi-class classification debugging and performance improvements. Finally, Yuan et al. [96] proposed iSEA,

which allows for similarity searching and clustering to identify subpopulations of error-prone groups in text documents.

Empirical Studies on How People and Organizations Find Error. In addition to tooling, HCI is interested in understanding how people and organizations make sense of models and resulting errors. In an early work on the area, Patel et al. [73] conducted interviews and did field studies on the difficulties software engineers face in building and evaluating the quality of models. More recent work in HCI and related areas of human-centered ML has studied how practitioners find (or do not find) errors in models around values like fairness([56] and interpretability [63].

Audits and Sociotechnical Critiques. Finally, error in classification systems has been a long-standing area of interest in STS and informatics scholarship [13]. Some are more conceptual, like developing taxonomies to conceptualize where error can happen in an ML model [77] and digital trace data itself [72, 86]. Closest to our work is the group of research that self-describes as audits or post-hoc evaluations of ML systems and the errors they produce. In their famous work, Buolamwini and Gebru [15] audit facial recognition algorithms and find that dark-skinned females are the most misclassified group, aligning with past feminist scholarship on intersectionality. Similarly, Scheuerman has led several audits of commercially deployed ML systems about gender [84] and computer vision datasets [83]. Blackwell et al. [11] found similar findings regarding the consequences of harassment classification. Our work builds on this prior work, specifically on understanding the inner workings of ML models and an audit of the sociotechnical outcomes of said models. We use a mixed-method approach of ML evaluation and qualitative contextual error analysis to study issues with proxy signals in mental health prediction. This mixed approach can address successes/failures, tradeoffs, and limitations of training data before systems have been deployed at scale.

3 DATA

Our data comes from Reddit, an online social media platform where people can post, vote, and discuss links and multimedia content. We identified an active ED subreddit, which we call `r/communED`. Per the community's self-stated preferences in their rules, we anonymize the name of this community and its distinguishing features in our narrative. Unlike previous work focused on recovery from EDs or promoting EDs [21], this community aims to be a neutral space for people with EDs to talk about their experiences in a socially supportive environment. We used the pushshift.io API to gather all posts and comments from `r/communED` for its lifespan. After removing posts deleted by posters or removed by moderators, we had approximately 70,000 posts and 415,000 comments. Summary statistics are in Table 1.

Unique posters	16000	Unique commenters	25000
Avg post len	106	Avg comm len	40
Avg post sd	116	Avg comm sd	50
Med. post len	73	Med. comm len	25
Avg post/user	4.8	Avg comm/user	16.5
Med. posts/user	2	Med. comm/user	3
Total posts	**70000**	**Total comments**	**415000**

Table 1. Summary statistics of our dataset. Statistics are lightly edited to similar standards of [22], per `r/communED`'s request for anonymity in papers.

4 METHODS

Building on similar work in triangulation and replication studies in applied AI systems [30, 39, 54, 63], we detail our replication strategies and methods decisions in this section.

4.1 Identifying Diagnostic Disclosures: History and Operationalization

Diagnostic disclosures are personal statements made on social media that indicate a person has been diagnosed with a condition or disorder, *i.e.*, matching statements such as "I was diagnosed with depression in 2016." Diagnostic disclosures were one of the first proxy signals in predictive mental health, pioneered in 2014 by Coppersmith et al. [24, 26]. This approach was seminal in the field and has been adopted in many different mental illnesses and prediction tasks (*e.g.* [10, 25, 71, 76, 94]). Diagnostic disclosure identification also grounds many benchmarking datasets like CLPsych 2015 and eRisk[1]. Diagnostic disclosures are helpful in building training datasets because of their proximity to clinical evaluation and presumed honesty in disclosures. Therefore, these disclosures are considered "gold standard" disclosures of mental illness. In addition, disclosures can be algorithmically extracted from large-scale public datasets using pattern matching. This speeds up development time and reduces the capital and labor costs necessary for a high-quality training dataset.

Building on this work, we operationalize diagnostic disclosure through the following definition: *disclosure occurs if a user states a personal clinical diagnosis of an eating disorder in a post/comment.* We designed our diagnostic disclosures in alignment with seminal work on detecting disclosures [26], recent systematic reviews on state of the art in mental health prediction [18, 55, 90], and triangulation studies similar to ours [39]. For training data, positive labels ($y = 1$) mean that a user made at least one disclosure in their post/comment history, while negative labels ($y = 0$) do not constitute a disclosure of a clinically diagnosed mental illness. This allows for disambiguation between general disclosure for mental health ("I have bulimia") versus the clinically specific ("I was diagnosed with bulimia"). This distinction is important because generic disclosures may have self-diagnosis in them and identity signaling characteristics identified in previous work [52]. Because of our dataset from `r/communED`, this task is more challenging than generic classification tasks of a randomly sampled negative dataset from all of Reddit.

4.2 Operationalizing Disclosure Detection: Creating the D-RegEx

Building on the recommendations of [24, 26], we iteratively developed a diagnostic disclosure regular expression to identify instances of diagnostic disclosures in ED communities, what we call the D-RegEx for short. Regular expressions are computational sets of characters that identify patterns in text data [60]. This rule-based system is widely used in many programming languages to identify health information in text data [14] and to identify diagnostic disclosures in previous work [10, 24, 26].

The D-RegEx operates such that if the post/comment p contained the given regular expression R, it would be labeled as a diagnostic disclosure ($y = 1$) and, therefore, used in positive training data. The D-RegEx contains three parts:

(1) use of first-person personal pronouns for self-reference (*e.g.* "I" and "my")
(2) variants of the term `diagnosis`
(3) A list of clinically recognized eating disorders from the Diagnostic and Statistical Manual of Mental Disorders (DSM-5), as well as common abbreviations and slang derived from [21]

Additionally, we include spacing between the regular expression units to allow for more organic communication (*e.g.* "I got diagnosed last week with anorexia" would not be a match without extra

[1]For a more expansive review of this method, please see [18] or [39]

space characters). We include a shortened python implementation of the regular expression in Listing 1.

The D-RegEx was developed over six iterations, refining its precision for our definition of diagnostic disclosure while minimizing false positives and negatives. For each iteration, the authors of this paper would randomly select 25-50 posts/comments to evaluate, slightly oversampling for D-RegEx matches. The authors of this paper and an undergraduate research assistant familiar with social media and EDs verified the performance of the D-RegEx and recommended improvements to refine the D-RegEx. We stopped iteration when the team perceived that future changes to the D-RegEx did not impact its quality and performance.

```
SELF_REFERENCE = [ "I", "i", "my", "i\'m" ]
DIAGNOSIS_WORD = "diagnos(?:e|ed|es|is|ing)"
DISORDER_NAMES = [
   "anorexi(?:c|a)", "ana", "an", #anorexia
   "bulimi(?:a|c)", "bn", "mia", #bulimia
   "bed", "binge eating disorder", #binge eating disorder
   "ed","osfed", "eating disorder"] #generic eating disorder terms

def build_regex(): #Returns a compiled regex for all pieces
   regex_str = ''.join([
        SELF_REGEX,
        '.{0,20}', #20 characters of space
        DIAGNOSIS_REGEX,
        '.{0,30}', #30 characters of space
        DISORDER_REGEX ])
```

Listing 1. Shortened python implementation of the regular expression

4.3 Two Methods for Evaluating the D-RegEx

Next, we describe our methods for evaluating the D-RegEx. Recall that our motivation is to consider the gap between standard quality evaluation methods and contextual gaps. Therefore, we devised a mixed-methods approach to evaluate quality.

4.3.1 "Gold Standard" Prescriptive Human Rating Task. To evaluate the D-RegEx, we used humans to assess the "gold standard" data for its definitional quality. In this task, human raters consider whether the results of the D-RegEx match the definition of diagnostic disclosure we outline above, which is a common practice for verifying the quality of labels [18, 90]. After finalizing the D-RegEx, we randomly selected 100 posts/comments from the dataset, 50 posts/comments the D-RegEx labeled $y = 1$ (diagnostic disclosure present) and 50 labeled $y = 0$ (diagnostic disclosure not present). The posts/comments were randomized and the D-RegEx's decision was blinded to the two raters. This paper's first and last author served as raters on this task. They both hold Ph.D.s in Computer Science and Information Science and have experience across HCI and social computing. They are experts in mental health, social media, and ED communities. They independently rated the 100 posts/comments on whether the post/comment aligned with our diagnostic disclosure definition, using the same binary (1/0) labels as before.

Rater quality and consistency is commonly evaluated in labeling tasks and definition evaluations [69]. Therefore, we use Cohen's kappa to evaluate the interrater reliability (IRR) between the two raters. Cohen's k is a statistic of IRR that indicates if two annotators had similar rating patterns for a given annotation schema. Between the two raters for the definitional evaluation of the D-RegEx, Cohen's k was k=0.78 between the two raters, indicating a high and substantial

agreement. The first and last author collaboratively resolved their disagreements to create a "gold standard" rated dataset of posts/comments containing diagnostic disclosures (and those that did not). This "gold standard" dataset was used to evaluate the performance of the D-RegEx, using precision, recall, and F1 measures.

4.4 Machine Learning Experiments

We design a user-level prediction task based on past ML work with diagnostic disclosures [10, 25, 71, 76]: given a user's post/comment history on `r/communED` before disclosing a diagnosis, does this person have an ED even if they do not disclose it? The setup for this task is based on diagnostic disclosure's intended use – before disclosing (and receiving a diagnosis), could ML models identify *similar others* that behave analogous to those with diagnosis (and therefore may be a good candidate for some sort of intervention)? We made methods decisions in alignment with standard practices in applied ML for mental health prediction in social media data [18, 90].

Preprocessing. Using python `gensim`'s built-in functions, we lowercase all data, remove non-alphanumeric characters, remove punctuation, remove links and usernames, and stem the text. We removed stop words by combining `gensim` and `scikit-learn`'s built-in stopword lists.

Additionally, we removed all terms used to construct the D-RegEx from the data, including diagnosis words, first-person subject words, and clinically recognized eating disorders. All users in the positive dataset will have necessarily used these words at some point; therefore, including these words could potentially bias the model to look for those words rather than latent distress signals, artificially boosting performance. We then set up a binary classification task with the following classes:

Positive training data. ($y = 1$): given an account holder with a D-RegEx match in at least one post/comment, we mark this account holder as "diagnosed". We represent the set of all posts and comments *prior* to the diagnostic disclosure for that user as a single document for positive training data. This approach models the common task in ML and mental health prediction of using all prior data before the disclosure as useful for prediction [10, 26, 76]. We take data before disclosure to minimize data leakage of post-diagnosis behavior on our prediction task, such as taking medication or going to therapy. If a user has multiple diagnostic disclosures, we take the earliest one as our date of interest (n=292)

Negative training data. ($y = 0$): given all users who do not have posts or comments matching the D-RegEx in their user history, we randomly sample users to create a balanced classification task. We then represent the set of all posts/comments for that user as a single document for negative training data. Our downsampling procedures and creation of a balanced classification task for this scenario align with prior work [18, 55, 90] (n=292)

Features and Model Selection. As we are interested in exploring the model, we chose classical statistical models and feature representations for our analysis. Statistical models are a common choice in the ML and mental health prediction space given the need for interpretability and easy-to-intuit feature importance [54]. These models are commonly deployed in the domain for both prediction tasks and triangulation work and critique of models [2, 39, 55]. We choose these explicitly over more complex feature and ML architecture decisions (*e.g.* BERT, XLNet, transformers, and other architectures) because these are harder to scrutinize for data quality concerns. Previous work on deep approaches also cautions that more complex models can amplify spurious correlations [80]. Given the limitations of our dataset size and these concerns, using classical models and straightforward feature architectures ensures "apples-to-apples" comparison between prior work (which is mostly using classical models[18]) and our auditing results.

For features, we use term-frequency inverse-document frequency (TF-IDF). This feature set is very useful in creating highly interpretable models and is a common baseline model for auditing mental health prediction tasks [39, 54].

We tested several statistical models suitable for our task, including SVM with linear kernels, logistic regression, decision trees, and Random Forest. We use the implementations of these algorithms in `scikit-learn`. We treat other model parameters, feature dimensionality, and unigrams vs. bigrams as hyperparameters to produce the most performant model based on F1 using 5-fold cross-validation. Our quantitative results present the average performance across our heldout data.

4.5 Contextual Error Analysis

During the development of the definitional labeling task, the authors collectively noticed complex errors in the D-RegEx that were not captured by standard practice in gold standard labeling tasks. Most tasks for evaluating the quality of labels take a *denotative* perspective, where human labelers examine whether the task definition has been satisfied with the label [79]. However, we noticed that the evaluation of diagnostic disclosure, traditionally conceptualized as a definitional rating task, had connotative or contextual complexity. Spurred by these thoughts, we formalized a method we call *contextual error analysis* to study these contextual findings, which we describe as errors.

Contextual error analysis is an inductive qualitative approach to examining contextual errors in ML models. This qualitative method builds on the tradition of error analysis in ML evaluation of test-set accuracy, where misclassifications are reviewed to improve subsequent modeling [5]. Rather than relying on metrics or counterfactual analysis, contextual error analysis introduces qualitative coding and thematic analysis into error evaluation to study crucial contextual details that can be missed in definitional or "prescriptive" annotation tasks [79]. The formalization of this process is based on the participatory ML evaluation of Wikipedia's ORES system [53, 87] and prior work in error analysis for mental health annotation [20]. As an inductive approach to qualitative work, contextual error analysis involves open coding for errors and grouping codes into categories (*e.g.*, types of errors) [81].

The authors began our contextual error analysis by reexamining the 100 posts/comments labeled in the gold standard rating task. The first and last author again served as the coders for this task. The coders independently read and qualitatively annotated the posts/comments, memoing contextual errors they observed [81]. The first author then reviewed the memos, then discussed the codes (and the errors they represented) with the team. Using these conversations and coding artifacts, the first author developed a shared error codebook, synthesizing a "shared language" for future reference in identifying contextual errors. The codebook consisted of categories of errors, with descriptions and examples of each type.

Working with the codebook, the first author asked all authors to analyze a fresh set of new data. The first author sampled another 200 random posts/comments, 100 labeled $y = 1$ by the D-RegEx and 100 labeled $y = 0$. The authors then coded these 200 posts/comments, using the codebook to assist in their work. The team reflected on two further questions (spurred by the earlier analysis): "What classification errors related to disclosure were made by the D-RegEx?" and "What problems with the formulation of the research task, if any, are evident in how the D-RegEx has classified texts?". Asking these questions allowed us to refine the conceptual development of our work around error, apply previously developed codes, and identify any additional error categories we found.

As they went, the codebook was refined and expanded through this pass with new examples. The authors provided annotations, notes, and insights to the first author, who then synthesized everyone's insights into a single set of findings. We reached theoretical saturation after annotating 200 posts/comments (*i.e.* no new codes came up towards the end of our annotation processes). We report on our joint findings in this paper.

4.6 Reflexivity and Ethical Considerations

We are mindful of how our identities, expectations, and values influence our study. As a team, we are cautiously optimistic that ML can be improved to better support people with EDs. However, we are also critical optimists in that we believe scrutiny of these models is of utmost importance to identify, address, and, ideally, avoid harm. Our personal experiences with eating disorders, other mental illnesses, and experiences with friends and colleagues with eating disorders shape our work and perspectives.

We also believe that we are obligated to the community that we study, and therefore, we made several decisions related to the ethics and privacy of our research [17]. Because this community and associated dataset are public and the researchers had no interactions with the community for this study, our ethics review board (IRB) did not consider this human subjects research. However, we still feel that we must take precautions to avoid harm caused in whole or part by our research. Based on the community's self-stated preferences in their rules and public posts, we anonymize the name of `r/communED` and all participant names and posts. Like other research in this domain [19, 45], quotes in this paper have been modified to prevent direct reidentification of community participants [6] and because they did not consent to be directly quoted in our paper [46].

5 RESULTS

In this section, we describe our Findings from our quantitative and qualitative analyses. First, we applied the D-RegEx to the dataset of 70,000 posts and 415,000 comments from `r/communED` to identify content with diagnostic disclosure and evaluate its validity. Next, we evaluate the quality of the D-RegEx to match our definition of diagnostic disclosures via the definitional labeling task. Then, we present our evaluation of the training data with our contextual error analysis. Finally, we present our ML results using our assembled training data.

5.1 Application of the D-RegEx and Face Validity Checks

We begin with a descriptive overview of the dataset created by the D-RegEx. Then, we examine the face validity of these results against prior work in diagnostic disclosures.

Diagnostic disclosure is a rare event in `r/communED`. 178 posts and 159 comments contained a diagnostic disclosure – we note that disclosure does not seem to be a social norm or a practice necessary for community participation or membership. Taken as a percentage of the overall dataset, approximately 0.002% of all posts (178/71,000) and 0.0003% of all comments (159/416,000) matched the D-RegEx. Given our interest in individual-level diagnosis patterns, we found the set of all unique commenters' and posters' usernames to determine how many individual account holders had disclosed a diagnosis. 292 unique usernames disclosed a diagnosis. This is about 1% of the unique participants in the whole dataset.

To verify that our D-RegEx is robust to prior computational work, we compare our dataset size created by the D-RegEx to the prior work on diagnostic disclosures, taken from the literature reviews of Ernala et al. [39] and Chancellor and De Choudhury [18]. In Table 2, we present the data set sizes of other papers for comparison. We compare against a few highly cited related papers, papers we could find about Reddit, and papers about EDs specifically in Table 2.

Our analysis of prior work indicates that our dataset of positive disclosures is within the expected ranges of this technique's ability to generate positive training data. Most papers study Twitter for health disclosures, though they do not provide the base size of the datasets they work with – meaning it is not feasible to calculate percentages of disclosures found in these datasets. For EDs specifically, three papers looked at self-disclosures in general Twitter datasets [9, 25, 76]. Prieto et al. [76] had about 800 users in their site-wide analysis (10 million Tweets) of Spanish and Portuguese

Prior Work	Condition	Data Source	User Count
De Choudhury et al. [33]	Post-partum depression	Facebook (consented data)	165
Mitchell et al. [71]	Schizophrenia	Twitter (broad)	174
Coppersmith et al. [25]	Eating disorders	Twitter (broad)	239
Our study	**Eating disorders**	**Reddit (one subreddit)**	**292**
Birnbaum et al. [10]	Schizophrenia	Twitter (broad)	671
Benton et al. [9]	Eating disorders	Twitter (broad)	749
Prieto et al. [76]	Eating disorders and obesity	Twitter (broad)	800
Yates et al. [94]	Depression	Reddit (all, 2006-2016)	9210

Table 2. An overview of user count and size for similar studies that use diagnostic disclosure. The vast majority of studies that use diagnostic disclosure do not provide the base size of the datasets they work with (meaning that it is not feasible to calculate percentages of disclosures found in datasets).

Twitter on obesity and eating disorders [76], Benton et al. [9] have 749 in a large set of Twitter users compiled from multiple studies [9], and Coppersmith et al. [25] have 239 users. We conclude that our dataset aligns with the training data size in previous work.

5.2 Gold Standard Evaluation of the D-Regex

Next, we measure the definitional performance of the D-RegEx to the criteria we set out in the methods. We measure this using macro precision, recall, and F1 score. Recall that we sampled 50 random posts positively identified by the D-RegEx and 50 random unmatched posts, and used experts to annotate whether they conformed to our diagnostic disclosure definition. We decided this amount based on prior work [18].

True/Pred	0	1
0	45	5
1	9	41
Precision	0.891	
Recall	0.82	
F1	0.854	

Table 3. Confusion matrix. "Predicted" represents the evaluation of a diagnostic disclosure evaluated by the D-RegEx. "True" values are the gold-standard hand annotations by the research team.

In Table 3, we show the results of evaluating the definitional quality of our D-RegEx. Our D-RegEx shows high empirical performance in identifying diagnostic disclosure in `r/communED`, with precision at 0.8913, recall at 0.82, and F1 at 0.854. We note that the performance is especially good at balancing for recall because regular expressions typically show much higher precision with fewer false positives.

Given our interest in error, we also examined the sources of error for the regular expression related to definitional errors. These insights were made by the research team to evaluate common reasons for erroneous matches with the regular expression. Definitional errors were fairly rare in our dataset. When considering the denotative goals of the D-RegEx, false positives in our dataset were primarily from people looking for support for others ("i'm worried about my sister's diagnosis of bulimia"). False negatives were caused mostly by misspellings of words like "diagnosis" or disordered eating behaviors not included in the DSM-5, such as orthorexia. Orthorexia is an

obsession with only eating healthy or "clean" foods, but, clinically, is considered a subset of other eating disorder diagnoses and therefore was not included in our D-RegEx.

5.3 Contextual Error Analysis

In this section, we explore the two error categories we developed through our contextual error analysis: temporal precision and decontextualization of clinical language/symptoms. We use anonymized and modified text excerpts from our dataset to illustrate our findings, which are heavily disguised but emblematic of quotes from our data. We also provide the annotation from the D-RegEx for context.

5.3.1 Time: Diagnostic Disclosures are Temporally Rigid and Brittle. One major contextual error for the D-RegEx was the inability to manage time and its impact on diagnostic validity. We call this *temporal rigidity*, or the inflexibility of diagnostic disclosures to account for changes in diagnosis and people's perceptions of their illness journeys. Time mediates the validity and relevance of disclosures for assigning a static diagnostic status.

In prior work, the presence of diagnostic disclosures is taken at face value to assign a person to a positive or negative training dataset. Time is not considered unless needed for a date-specific mental health event (such as a suicide attempt or date of hospitalization) [27]. As stated before, it is assumed that all posts of a person who discloses a genuine diagnosis are considered in the "treatment" or positive training group.

Using our contextual error analysis technique, we found that time mediated many diagnostic disclosures. In `r/communED`, time is often used to talk about diagnosis, contextualize story details, or request advice. This assumption resonates with and extends the assumption of an objective record described by [43]. The assumption of an objective record of diagnosis highlights how algorithmic approaches to mental illness prediction "do not support understanding how the meaning of recorded content may change over time or differ depending on the timeline of posting or viewing" [43].

For instance, diagnostic disclosures range from being very recent, as people talk about the process of being diagnosed in the last few days and how they were feeling about it:

> ...yesterday I went to the doctor, and bam! I got a diagnosis of bulimia... (D-RegEx =1)

In another example, a person talks about the recent impacts of their diagnosis on treatment and eating habits from a few weeks ago:

> On my 17th birthday a couple weeks ago i was diagnosed with anorexia...since then my food has been monitored (D-RegEx = 1)

In both examples, the research team agreed that the D-RegEx did identify people disclosing recent and relevant diagnoses. The diagnosis is temporally salient and valid for very specific posts like these two examples. If the specificity of diagnostic dates was needed, researchers and engineers could reasonably assess these dates and carefully build training data that deals with time.

However, more often than not, diagnostic disclosures and the D-RegEx assumed that diagnosis was rigid and unchanging for people in `r/communED`, much like Feuston and Piper's notion of the "objective record" [43]. This happened in two primary ways:

Temporally Ambiguous or Old Diagnoses: In our dataset, we found that most diagnosis disclosures did not explicitly disclose the date of diagnosis or include specific information to help determine a diagnosis date.

For example, this poster describes being diagnosed with binge eating disorder last year but having symptoms for 6 years:

> I was diagnosed with BED last year but I've been struggling with it for 6 years. My doctor told me I needed inpatient, but my family couldn't afford it (D-RegEx =1)

In some cases, the diagnosis extended back months, years, or even a decade:

> I was diagnosed with anorexia when I was 16, I'm 28 (D-RegEx =1)

The research team struggled with annotating temporally imprecise diagnostic statements. Considering these statements from the clinical and critical perspective sheds light on their ambiguity. Clinically, it is hard to use diagnostic information from many years ago to assess a person's current state because the diagnosis may become less clinically relevant over time. For example, while a diagnosis may be relevant to a person's experiences or identity, a diagnosis from many years prior is not as clinically relevant as current behaviors and symptoms, making intervention recommendations hard. Critically, the D-RegEx makes diagnostic disclosure a label for someone no matter how old the diagnosis is. Diagnoses may change (*e.g.*, binge eating disorder to anorexia) and do not reflect how people heal and recover, where conceptualizations of "having" an eating disorder may shift from being diagnosed to being in recovery).

However, we want to recognize that for many people, a diagnosis of an eating disorder can be a lifelong diagnosis or part of the identity that they carry with them. This may have been the case for the poster diagnosed a decade ago. Nevertheless, past diagnoses may not accurately reflect the current personal state of a person – meaning that the data we have about them may not neatly map to the ideas of a clear-cut diagnosis in the present. This leads to ambiguity and reflection about how – and whether – a cutoff window can be developed without inadvertently invalidating the experiences and identities of people with eating disorders.

Changes in Diagnosis over Time: Another challenge of temporal rigidity involves changes in diagnosis and the inability of the D-RegEx to conceptualize illness journeys. There are many prior works that conceptualize illness as a journey that has multiple stages [48, 58, 64].

Some posters within our dataset mentioned their belief that their eating disorder diagnosis would change, were they to be rediagnosed.

> Little background, I was diagnosed with bulimia several months ago...but I think if I went to get diagnosed again I'd say I have binge eating disorder. (D-RegEx =1)

> I've been diagnosed with anorexia in the past, although now I would say things are closer to EDNOS/OSFED if I would even be diagnosed now. (D-RegEx =1)

Some disclosed as they described their success toward or in recovery, whereas others were worried about relapsing, calling into question the staying power and personal relevance of older diagnoses.

> I was diagnosed bulimic and went through recovery last year after like 9 years. (D-RegEx =1)

As these examples highlight, individuals with EDs do not necessarily think of diagnosis as a static artifact – and the ML task operationalizes this concept in a very fixed and static method. Our contextual error analysis highlights that people conceptualize their eating disorders differently, depending on their lived experiences and disordered eating practices [44]. However, the D-RegEx makes assumptions about what diagnosis looks like for people – as a static construct that supports machine learning rather than the lived experiences of people who may be sub-clinical, between diagnoses, or may not map neatly to a typology of illness [43].

We see a variety of temporal narratives related to diagnosis that illustrate the different ways people conceptualize and experience their eating disorders. A D-RegEx that does not consider these different temporal narratives fails to accurately classify people's experiences and, therefore, the status (related to diagnosis or not) they feel currently accounts for their experiences.

In summary, we argue that the D-RegEx statically operationalizes diagnosis versus diagnosis as a fluid and personal concept to individuals in `r/communED`. This affects how training data is built

from the dataset and can lead to data leakage problems. By referencing old, changing, or other diagnostic information from *past clinical encounters*, the positive training data overprioritizes prior clinical interaction and diagnosis as relevant for features related to diagnosis. There is a risk of a contextual gap, related closely to data leakage, and error propagating through the model because past diagnoses and clinical interventions can change people's behaviors.

5.3.2 Missing Symptoms, Recovery, and Interventions. Next, we examine our second category of contextual error – the inability to evaluate contextual details that convey the gravity of someone's ED and that someone does NOT have an ED even with this information.

The most common conceptual false negative was that the D-RegEx could not identify complex symptoms presentations without accompanying diagnostic language. Because the regular expression was tuned to precisely identify the diagnostic disclosure, the D-RegEx *also* says that anything nondiagnostic becomes a nonrelationship with an ED.

The primary indication was in language about symptoms and behaviors that a clinical outsider viewed as unequivocally connected to eating disorders. For example, this individual lists the symptoms they are experiencing and considers the consequences of diagnosis:

> I wonder if i should try to get diagnosed with BED and get help crazy binge episodes and wanting to purge (D-RegEx =0)

In the above post, the individual mentions several disordered eating practices, including binge episodes and desires to purge; however, the D-RegEx did not label this post as a diagnostic disclosure indicative of an ED. These posts had no mention of diagnosis (and, as such, were not diagnostic disclosures) and it is unclear whether the individual would meet the requirements for clinical diagnosis. Nevertheless, individuals who mention disordered practices (or behaviors we may clinically consider"symptoms") are legitimately experiencing some distress related to their relationship with eating, food, or their body – even if this concern is ultimately subclinical or on the threshold. By creating models that exclude constellations of "symptoms" or disordered practices, we risk erasing and invalidating the experiences of certain people, as well as developing classifiers that are completely unaware of symptoms and people – people who may be at risk or who may benefit from targeted interventions.

In another salient example, one account holder described the process of traveling overseas for bariatric surgery to maintain their low weight resulting from severe ED behaviors and prolonged restriction. The poster warned others that their severe ED had pushed them toward surgery. Although this example is extreme, the D-RegEx marked that it did not contain a disclosure and, therefore, would be classified as negative training data (indicative that the account holder did not have an ED).

The problems we see with symptoms extend to other ED experiences, such as recovery or treatment and intervention, which may also be inadvertently excluded by D-RegEx.

> So I just got news that the recommended level of care for my ED is a residential program (D-RegEx =0)

> I am going to an ed recovery place (D-RegEx =0)

Being admitted to a residential program can be a major step in someone's healing journey – and, in most cases, would certainly necessitate a clinical diagnosis. By incorrectly classifying posts that talk about residential programs or recovery places (as well as similar posts), the D-RegEx completely misses instances where diagnosis may be implicit. This gap in detection raises questions about other situations where the diagnosis may be implicit and, as such, not included or recognized. Although diagnosis is not all that matters for people with eating disorders (as we see above with discussions of symptoms and experiences), a D-RegEx intended to capture diagnostic disclosure

may effectively exclude people who are diagnosed but who do not use language that can be captured by a machine – which means that ML work may inadvertently disregard and exclude people simply because they did not say something as specific as "diagnosis". We reason more about community norms and why this is happening in the Discussion.

Finally, the classifier made errors in understanding peoples' experiences with recovery. For example, this poster discussed their movement through therapy to recovery:

> I was diagnosed with bulimia when I was 16, now I'm 20. It was an incredibly hard and long process to stop purging but it feels great to be purge free. (D-RegEx =1)

Though the above poster does not explicitly mention recovery, they describe themselves as currently purge-free. Combined with the past tense of their diagnosis and the temporal relationship between when they were diagnosed and their current age, one interpretation is that this individual is in recovery (or quasi-recovery) and their diagnosis does not accurately represent their current experience [44]. In this case, someone now described as being in recovery would be placed into a training data class of "disordered".

In sum, the D-RegEx ignored many common signals and symptoms for people with valid ED experiences. At their extremes, this meant missing people with clinical levels of EDs implied by the larger narrative of their content (e.g., residential programs). It also invalidated many subclinical and subdiagnostic experiences that people have with eating disorders and the ways people move into and through recovery and non-clinical healing processes. We demonstrate that this creates bias and poor representativeness in the next section.

5.4 Machine Learning Experiments

Recall that our positive training data comes from account holders who had a post or comment positively labeled by the D-RegEx as containing a diagnostic disclosure. Negative training data is randomly sampled from users who do not have a post or comment labeled by the D-RegEx. To avoid possibly training our classifier on words that naturally appear in the D-RegEx (and therefore cheating performance), we remove all string matches in the posts and comments to terms present in the regular expression (see Section on Preprocessing).

Through our investigation, the most performant model was $l-2$ regularized logistic regression, based on F1 score. Over 5-fold cross-validation, we found that the best-performing model had 270 unigram features and C = 1.

In Table 4, we present the confusion matrix and the results of the average model performance from our 5-fold cross-validation. The performance of the model is strong at distinguishing between the positive and negative datasets, with an average accuracy of 0.91 and a macro-averaged F-1 at 0.91. Our model has no trade-off between precision and recall, with both metrics at a macro-averaged 0.91. In summary, our model demonstrates strong empirical separation between the two classes.

5.5 Feature Analysis

In Table 5, we present the top 25 features and the relative importance of each feature in prediction through the β (beta) values associated with the best model from our cross-validation. For β values, the sign dictates the direction of the influence of the presence of the word on positive (y = 1) or negative (y= 0) prediction. A larger magnitude implies a stronger influence on the features. Recall that, in addition to removing standard stop words, we also removed any word stems that appeared in our D-RegEx for *all* posts. This avoided biasing the model towards our regular expression pattern as well as removing textual indicators of diagnosis and specific disorders from the dataset.

True/Pred	Class 0	Class 1	Total
	0	1	
0	54	4	58
1	6	50	56
Precision	0.90	0.93	0.91
Recall	0.93	0.89	0.91
F-1	0.92	0.91	0.91
Accuracy	0.91		

Table 4. Summary of model fit and performance of the classifier on the average cross-validation scores

feature	β	feature	β
mom	-0.59	eat	2.01
brain	-0.43	treatment	1.77
super	-0.42	know	1.73
enjoi	-0.40	recover	1.32
lmao	-0.37	time	1.26
comment	-0.33	doctor	1.26
worth	-0.32	year	1.22
look	-0.30	therapist	1.20
weird	-0.29	weight	1.19
pretti	-0.22	week	1.14
bui	-0.21	want	1.11
size	-0.21	normal	1.08
great	-0.18	month	1.07
sugar	-0.17	ago	1.03
mean	-0.16	bing	0.99
talk	-0.15	actual	0.98
couldn	-0.15	thing	0.96
watch	-0.15	fuck	0.95
plan	-0.12	need	0.95
face	-0.12	underweight	0.92
nice	-0.12	medic	0.92
freak	-0.11	purg	0.88
big	-0.07	point	0.83
felt	-0.07	absolut	0.83
allow	-0.06	control	0.82
happi	-0.06	health	0.80
isn	-0.05	depress	0.80
good	-0.05	new	0.78
yeah	-0.05	pound	0.77

Table 5. Top 25 features in the model with the largest positive/negative coefficients (β). Purple words indicate their connection to medical or clinical interactions, and pink words indicate time-related words.

Despite the strong empirical performance of the classifier, we noticed peculiar language patterns in what the classifier associated with a higher likelihood of appearing in the content of people

with diagnoses (positive β values). For example, we see medical terms like "treatment", "doctor", "therapist", "recover", and "medic-", highlighted in purple in the table. This implies that, prior to diagnostic disclosure, users who are more likely to disclose a diagnosis (and, therefore, labeled positive training data) are talking about potential or past clinical engagement.

This is a paradoxical condition and an example of data leakage if used to identify individuals with EDs who are not in our positive training dataset (or who may be in a real-world application of this model) – the model looks for language that indicated prior clinical/medical involvement as a positive signal to identify a future diagnosis. In other words, if a clinician operated similarly to D-RegEx, they would identify individuals with an ED only after someone mentioned previously seeing a doctor or therapist.

We see similar contextual gaps with discussions of time and self-reflection that occur in contextual error analysis as connected to content around diagnostic disclosure(positive β values). We also see the presence of temporal language as well, such as "time", "year", "week", "month", and "ago". In summary, even though training data is based on clinical diagnosis, communities are using the language around diagnoses for purposes beyond just describing the present-day clinical reality. As our qualitative results show, discussions of time connected with diagnosis may connect to other narrative purposes of diagnosis than just talking about what happened recently.

6 DISCUSSION

6.1 Contextualizing Our Results with Prior Computational Work

Our empirical results suggest that the D-RegEx replicates past methods in terms of the size and accuracy of datasets [18, 24]. Using the D-RegEx, we found that approximately 1% of all participants in the lifetime of `r/communED` have made an explicit ED diagnostic disclosure (292 unique usernames in total). Looking at individual interactions on the subreddit, we found that 0.002% of posts and 0.0003% of comments contain a diagnostic disclosure of an ED. The dataset size aligns with the amount of data collected in previous studies (see Section 5.1). Furthermore, the training data produced by the D-RegEx effectively distinguishes people with a valid diagnosis who have communicated this experience online (F1 = 0.854). The model built from this data also had strong empirical performance (F1 = 0.91, Acc = 0.91).

Our contextual error analysis identifies subtle but systematic problems that show that the model does not accomplish its intended goals – *contextual gaps* in the design and development of models made from this proxy signal [8]. Contextual gaps are similar to those found in Human-AI scholarship, such as the semantic gap [30] and methodological gaps [39]. Unique to our work, contextual gaps focus on details that can influence what proxy signals are appropriate to use in a given context. In our case, important contextual details are those about mental illness journeys [58], about EDs specifically [44], and normative/community considerations in `r/communED`. Our feature analysis shows that, even before the diagnostic disclosure, the ML classifier built using erroneous data looks for erroneous context, searching for people who have disclosed prior medical interactions as crucial to identifying people who would become diagnosed later. If an analogous clinician operated like the ML model, they would identify individuals with an ED only after someone mentioned previously seeing a doctor and receiving a diagnosis *before disclosure*. We see similar findings in the qualitative results as well. Stated in terms of gaps, the model has a contextual gap that fails to operationalize nuances of diagnosis in the intended task. However, the model does so in a very subtle way that is hard to detect through quantitative metrics alone. This aligns with recent work that critiques the generalizability and representativeness of work that predicts mental health in social media data by experts in the field [1, 2, 39, 54].

Joining previous scholarship in CSCW and HCI research on gaps in applied ML around well-being [30, 39, 62], we argue that **researchers and designers must prioritize the intellectual integrity and validity of training data and the conceptual foundations of models before any deployment.** CSCW and HCI scholars point to the risks and challenges of creating training data that are not considered in current methodological practices – we believe that what we have found is both a "data cascade" [82], or, in the most extreme, "garbage in, garbage out" [49]. In the past work on diagnostic disclosure (see Section 5.1), we found no evidence that these proxy signals were validated beyond a hand-labeled precision check that the regular expression worked as "intended" – meaning that they met the definitional criteria of diagnosis. Our team did not initially see the contextual gaps ourselves until we performed a qualitative analysis of the data for a related project, repeating the same contextual errors. Recall that the purpose of modeling in previous work [18, 90] is to improve detection *before* clinical interventions have happened, thus leveraging social media as an alternate source of information. A precisely implemented definition of diagnosis avoids the ambiguity of self-diagnosis as a sensemaking tool [52] – *e.g.* a post saying, "I have anorexia," is ambiguous about how the assessment was made and who made it. However, Chancellor and De Choudhury [18] identify that very few proxy signals are validated during the creation and training data phase, which risks downstream impacts on model quality [18].

These contextual gaps are relevant to more than just mental health prediction research. Our findings also suggest that even high-quality proxy signals are prone to problems with epistemological gaps that create what Geirhos et al. [50] describes as "shortcuts" in AI models. Geirhos et al. [50] define a shortcut as decision rules that perform well on face but fail on out-of-domain testing that highlights the "mismatches between intended and learned solutions" [50]. Our results show that, indeed, the model for predicting ED diagnosis shortcuts to a solution that produces high "accuracy" but has systematic problems in the way training data operationalizes the intended solution. In fact, in other medical AI domains, AI systems have been found to select shortcuts in chest radiographs that "detect" COVID-19 but do not generalize to new hospitals [35] and use spurious data about X-ray quality to determine COVID-19 status [66]. Problems with shortcuts cannot be solved just by increasing the size of datasets or increasing the complexity of feature or machine learning architectures. In fact, more complex models can make shortcutting and spurious correlations even worse [80].

6.2 Contextual Error Analysis as a Method for Error Analysis

We formalize the method of "contextual error analysis" to conduct more thoughtful error analysis during the model conceptualization and training data creation process. Contextual error analysis is based on the strengths of key qualitative methods [81] and calls for better assembly and annotation of training data [36, 49, 79]. We are not the first to perform analyses like these; this work is conceptually similar to evaluations in participatory Wikipedia ML design [53] and content moderation AI design [20]. And, in the work on triangulating signals, previous scholarship has used qualitative analysis to help triangulate issues in self-reported affect quality for well-being and emotion prediction [30, 62].

Contextual error analysis provides a nuanced qualitative tool for evaluating ML models in a few distinctive ways. First, contextual error analysis extends prior work by focusing on questions of validity and representativeness, raising the notion of error beyond just evaluating false positives or negatives. This method reconceptualizes the notion of "error" to more than just definitional debates [79] – annotations and their errors can be related to denotative (or strict) definitions that must be applied versus descriptive definitions that a labeler can subjectively interpret. Second, contextual error analysis is designed to occur during the model development rather than being a post-hoc activity like many error analyses. Contextual error analysis moves beyond checking

the definitional quality of proxy signals as documented by Chancellor and De Choudhury [18], because it can be sensitive to contextual factors, such as stakeholders' opinions and considerations of how community norms and practices shape the ways content is shared.

We envision several pragmatic ways contextual error analysis can be used in future applied ML research. As we demonstrate, contextual error analysis is a strong candidate for unearthing challenges with similar shortcuts in human-centered machine learning research. This method may provide ML designers with another tool in their toolkit of error analysis techniques, which predominantly rely on quantitative metrics for evaluation [5, 93]. We believe this applies to the domain of mental health prediction, but also to applied ML more broadly. Contextual error analysis focuses the qualitative work on these questions and resulting themes by design, providing a framework for the error analysis strategy. For example, contextual error analysis could be applied in subjective annotation of hateful language on social media, considering more than a rigid policy (the strict definition). Contextual error analysis would allow researchers to examine contextual factors about a community, its dynamics, who participates, and its history in evaluating when something may be hateful. Another application of contextual error analysis is to facilitate more robust auditing and triangulation research, which has gained popularity in the last few years as AI systems are more widely deployed. We imagine that contextual error analysis could identify some semantic or methodological gaps [30, 39] and facilitate better audits of training data [83, 84]. Third, conducting a contextual error analysis is deliberative and reflective– which we argue will help support more deliberative reasoning about applying ML models. Our team spent a fair amount of time discussing the findings, building consensus, and critiquing our methodological decisions in this process. We imagine similar processes carried out in teams would facilitate critical conversations about model development, when model work is often done in isolation [82]. Fourth and finally, contextual error analysis provides empirical evidence to help teams identify trade-offs in modeling decisions and through the discursive process, weigh trade-offs, benefits, and consequences. We are excited about the potential for contextual error analysis to be adopted as a formal process in applied ML research.

One challenge to the wider deployment of contextual error analysis is the scalability and guarantees provided by qualitative methods. Qualitative methods take time and substantial effort. Even an analysis of 1000 items (large for qualitative analysis) would likely not scale or guarantee that the analysis uncovers all errors. We are optimistic that qualitative error analysis is useful for teams as they conduct other kinds of error analyses or possibly in concert with these approaches. For example, contextual error analysis could sample based on the results of quantitative findings, or provide directed insights when prior model releases have uncovered errors in the past. Likewise, we are excited about crossover research combining qualitative methods with text analysis [7, 59], helping to better scale methods like ours to larger datasets that will require substantially more labor.

6.3 Building on Sociotechnical and Critical Approaches to Categorization and Classification

Classification and ML models like ours – both in the creation of training data and the prediction results of these models – is a form of imposing order along fixed and rigid dimensions of classifications and categories. As Bowker and Star write, classification systems "valorizes some point of view and silences another. This is not inherently bad —- indeed, it is inescapable. But it *is* an ethical choice" [13][p. 5].

Who is "valorized" and who is "silenced" when training data is assembled from proxy signals like diagnostic disclosure? As we see with `r/communED`, diagnostic self-disclosures of ED are exceedingly rare in this community. However, the experiences of a few people who have made ED diagnostic disclosures are valorized, or prioritized, over all other community members – the majority, no less –

by D-RegEx and by models that have been developed similarly. Researchers choose to valorize some voices over others to build robust datasets, who may inadvertently be making these decisions (e.g., to valorize, to silence) that we describe here. However, we firmly believe people who participate within online eating disorder communities, who describe ED experiences from their points of view and who may or may not have a clinical diagnosis, have valid EDs and experiences. With or without a clinical diagnosis, the experiences of online ED community members should be included in datasets to increase dataset diversity and to develop datasets that represent a broader swath of ED experiences and online accounts.

The need for computational precision mandated by the D-RegEx and the ML model ignored that diagnosis was discussed alongside a larger, everyday narrative of people's ED journeys [44]. Expanding Feuston and Piper's research, our qualitative findings illustrate how working with diagnostic self-disclosure as a proxy signal runs the risk of three assumptions that they find happen in ML systems: 1) downplaying the importance of context and assuming that the meaning can be generalized across large groups of people with diverse and deeply personal experiences, 2) assuming that content represents a static record with an objective and persistent meaning, and 3) assuming that content can be classified (when, in fact, many experiences and the ways they are represented online occupy thresholds and liminal spaces in which classification may oversimplify and misinterpret) [43]. In trying to find good training data, the D-RegEx "reduced" the rich signal of the community to narratives of clinical interaction and care.

Valorization of certain voices and perspectives cannot occur without silencing others (even inadvertently). When experiences with ED diagnosis are valorized, those representing other types of first-hand accounts and tellings may be "silenced", in the sense of Bowker and Star [13], or invalidated (e.g., by researchers, by the ML models we build when not supported by diagnostic definitions). For example, in analyzing influential features, the ML model relies on language related to the findings of the contextual error analysis. This leads to the systematic inclusion of individuals who made posts or comments about *past* ED diagnoses in the sample subsequently used to train models to detect users who may *currently* have an ED. Although effective in returning some true positives, this approach systematically produces false negatives for individuals with potential ED symptoms (e.g., binge episodes, restriction), experiences that indicate distress, and (in some cases) clinical intervention that does not include a retelling of diagnosis and/or does not match a definition of diagnosis. This means that for whatever downstream purpose a model has, posts and comments mentioning diagnosis become prioritized in model development. We argue that this privileging of certain experiences can have mixed and somewhat surprising impacts on people in these online ED communities. For example, if AI systems were used to predict who should receive an intervention with support resources, people without visible indications of medical care may not receive those resources because of the paradoxical setup of the ML model. If used for content moderation, this overly narrow definition could mean that people with diagnoses are seen as more "severe" and "dangerous", leading to outsize scrutiny of their posts, while ignoring other risky behaviors in our dataset.

By creating these hierarchies embedded in AI systems, ML research like this risks inadvertently perpetuating stereotypes about people with mental illness and exclusion. Many disciplines, including critical HCI, critical psychiatry, and disability studies, have reflected on long-standing issues with how mental illness is constructed (through diagnosis), treated, and how people living with mental illness have been fundamentally excluded from care decisions [23, 44, 57, 89, 95]. Although a powerful tool for providing support and treatment, mental illness diagnosis has an unfortunate history of being used to classify and control people who do not meet societal norms and individuals at odds with those in power [85]. At its worst, the abuse of the power of diagnosis and misunderstanding of people with mental illness have contributed to discrimination through sanism and

ableism and violence against people with mental illness (*e.g.*, forced institutionalization) [28, 47]. While we do not expect AI developers to intentionally want to cause these harms, valorizing diagnosis can have many downstream consequences. Additionally, given the criteria associated with diagnosis (which are important for identifying and treating diseases), many people do not meet the requirements for an ED diagnosis (though they may still experience certain ED behaviors and practices), including those in r/communED– meaning that their experiences with eating disorders are sub-clinical and potentially excluded from ML work that uses clinical definitions.

6.4 Recommendations for Future ML Work

Studies like ours seek to "invert", in the Bowker and Star sense, the opacity and inscrutability of datasets and ML models by making visible decisions, processes, and consequences of ML models [12, 36]. Through this inversion, we join prior scholarship that mathematically and conceptually critiques ML systems [1, 2, 39] and, through social critique, seeks opportunities to improve them going forward.

6.4.1 Composited and Representative Datasets. The rarity of diagnostic disclosures in certain communities, in both r/communED and in prior work (see Table 2, Results), should give pause to researchers intent on using diagnostic disclosure as a standalone proxy signal. A 1% user sample of r/communED or for other conditions is not representative and introduces the risks of homogeneity and systematic biases in the dataset. By not considering subclinical or non-clinical experiences in the ML lifecycle, this process focuses on a narrow subset of valid ED experiences that do not represent the community.

In line with recommendations from Ernala et al. [39], developing more robust and representative datasets could improve the quality of training data. More representative datasets should address the diversity of mental illness by including people with different experiences (*e.g.*, of the same mental illnesses), identities (*e.g.*, age, class, gender, race), and points of view (*e.g.*, EDs as socially constructed and as lived experiences, EDs as clinical diagnoses). We envision combining multiple proxy signals for mental illnesses to form a more robust dataset [18, 39]. Contextual error analysis may help here by identifying gaps in current applications of proxy signals and pointing to promising others that may fill those gaps. By including additional proxy signals, datasets could highlight the specificity of experience without potentially overgeneralizing and obscuring diversity [39, 43, 45]. Scaling this dataset would need to be balanced against concerns about psychological validity as discussed by Harrigian et al. [54], as well as the risks of classification systems applied at scale [11].

6.4.2 Humans-in-the-Loop for Training Data and Incentivizing Hard Data Work. Human annotation supplements automated proxy signals like the D-RegEx [18] – What if we reenvisioned people's role in building hard datasets? More human interaction with data can help identify contextual errors, an idea promoted by Bellotti and Edwards [8] more than 20 years ago. However, data work is emotionally difficult and time-consuming and is often perceived as less valuable than model development [82]. To incentivize this labor, researchers need to be mindful of what they are asking from annotators, especially on crowdwork or distributed work platforms. These external observers and annotators may be at risk of assuming that content has a specific objective record, where their interpretation from an external, supposedly objective point of view does not align with – and, therefore, misrepresents and even conceals – the lived experience being expressed through content [43].

6.4.3 Encouraging Action from the ML and Research Community. Another avenue for improvement is looking towards the ML and research communities to improve accountability for proxy signals in research papers. In alignment with prior work [18, 39], we hypothesize that the use of proxy

signals like diagnostic disclosure is what Ajmani et al. [3] describe as normatively "sticky" – commonly used but with limited scrutiny about their validity for a given ML task [3]. Methodological stickiness can be challenging to overcome purely with *implicit* practice and can discount the harms published papers can have on scientific accuracy, applicability, and benchmarking [17]. Therefore, we encourage ML researchers and paper reviewers to adopt explicit scrutiny of proxy signals in research work. First, reviewers could ask authors to empirically validate the proxy signal's ability to measure concepts as a necessary condition of paper acceptance. This may be through methods similar to those that expose data leakage, empirical studies of the signal's validity, qualitative analyses like ours, or careful work with domain experts. Systematic reviews can also critique practices in the field and provide guidance for future work.

6.4.4 Encouraging Deliberation and Reflexivity About ML Models. Tools to support more human-centered goals around fairness and interpretability are often misapplied and misunderstood by machine learning experts and data scientists. For example, in interpretability, Kaur et al. [63] found that empirical tools, such as visualizations and fairness metrics, lead to overtrusting system results and misuse of metrics that need much deeper contextualization and domain awareness [63]. Our contextual error analysis method assists with what Kaur and colleagues call "deliberative reasoning" about contextual bias and data leakage [61]. Our formalization of the contextual error analysis is a step in this direction — qualitative analysis can encourage teams to consider errors beyond what tooling and metrics report.

Contextual error analysis also requires a reflexive way of viewing and interpreting data. By acknowledging and grappling with their positions, labelers can reflect on how their own experiences color their interpretation and annotation of data [16]. Researchers and labelers can address the assumption of an objective record [43] and maintain space for other interpretations to live alongside (and compliment) the interpretation that researchers put forward in their dataset annotations and model building.

6.4.5 Community-Centered and Participatory ML. Rather than relying on ML engineers who may not have mental illness domain expertise or may have different lived experiences, we strongly advocate for working with community members throughout the research process. This involves more than bringing people to serve as temporary experts, but developing deep and multi-year-long partnerships [55]. This community-centered approach has similarities to human-centered and value-sensitive approaches to machine learning [97], as well participatory design for marginalized and vulnerable populations [67]. We believe involving community members will be mutually beneficial for ML model development – stakeholders can identify task formulation and training data problems before costly and time-intensive development and deployment takes place. Likewise, communities benefit because their norms and views are more justly represented in system design decisions, and they may be able to contest unreasonable expectations or assumptions. Communities may find these models more meaningful and useful, ultimately increasing their acceptability for community deployment.

6.4.6 New Paradigms for ML Tasks. The use of diagnostic definitions to study online spaces is one interpretation of how people experience their ED or other mental illnesses [43]. Diagnosis has always been – and continues to be – a contested and changing classification system in medicine [13, 37, 70]. What would it be like to consider new perspectives on what illness experiences mean to people and our ML systems? One way this could be operationalized is by reconsidering the intended goal of an ML system – must it use diagnosis to find people who are in distress and need support? Another idea is reconceptualizing what other indicators of experiences of mental illness are considered in our training data definitions – is it worth including people indicating distress or support seeking in

online communities? We encourage ML researchers and practitioners to evolve problem statements and definitions to alleviate these problems.

6.5 Limitations and Generalizability

Our work is not without limitations. Our study focuses on r/communED, a single community on Reddit. Therefore, our findings around clinical paradoxes and temporal rigidity may not apply to other communities or platforms and cannot be immediately assumed to hold in those circumstances. Likewise, our findings are focused on eating disorders, and future work will need to be conducted that examines the use of diagnostic disclosure as a proxy signal for other conditions on other platforms. Our work also cannot make claims about the quality of all proxy signals in this community, such as hashtag use, clinical appraisal of social media text, or community participation. In short, this work does not intend to criticize all uses of proxy signals for all mental illness detection tasks.

Instead, we call for closer consideration of validity as a core value in the applied ML process. Through an evaluation of an end-to-end ML pipeline, our work shows that validity errors are insidious, can be difficult to diagnose, and lurk behind strong model performance. We also demonstrate that other studies may have this problem, given a lack of introspection and methodological validation of proxy signal quality [2, 18, 39]. Our study was based on similar premises for prediction before we recognized validity errors. We sincerely hope that computational researchers and practitioners take our approach and the contextual error analysis method we proposed to validate proxy signals before being incorporated into ML lifecycles.

7 CONCLUSION

Diagnostic disclosure has been touted as a strong proxy signal for identifying high-precision training examples in social media data. Using standards from state-of-the-art research, we show that diagnostic disclosure causes complications in operationalizing the theoretical task to training data and resulting ML model. Like previous examples of error analysis, many of these problems are obscured by strong empirical performance. We hope that this research inspires new research to evaluate the validity of training data and new methods to address these concerns.

REFERENCES

[1] Carlos Aguirre and Mark Dredze. 2021. Qualitative Analysis of Depression Models by Demographics. In *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*. 169–180.

[2] Carlos Aguirre, Keith Harrigian, and Mark Dredze. 2021. Gender and Racial Fairness in Depression Research using Social Media. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. 2932–2949.

[3] Leah Ajmani, Stevie Chancellor, Bijal Mehta, Casey Fiesler, Michael Zimmer, and Munmun De Choudhury. 2023. A Systematic Review of Ethics Disclosures in Predictive Mental Health Research. In *Forthcoming at FAccT 2023*. 1–15.

[4] Saleema Amershi, Maya Cakmak, William Bradley Knox, and Todd Kulesza. 2014. Power to the people: The role of humans in interactive machine learning. *Ai Magazine* 35, 4 (2014), 105–120.

[5] Saleema Amershi, Max Chickering, Steven M Drucker, Bongshin Lee, Patrice Simard, and Jina Suh. 2015. Modeltracker: Redesigning performance analysis tools for machine learning. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 337–346.

[6] John W Ayers, Theodore L Caputi, Camille Nebeker, and Mark Dredze. 2018. Don't quote me: reverse identification of research participants in social media studies. *NPJ digital medicine* 1, 1 (2018), 1–2.

[7] Eric PS Baumer, David Mimno, Shion Guha, Emily Quan, and Geri K Gay. 2017. Comparing grounded theory and topic modeling: Extreme divergence or unlikely convergence? *Journal of the Association for Information Science and Technology* 68, 6 (2017), 1397–1410.

[8] Victoria Bellotti and Keith Edwards. 2001. Intelligibility and accountability: human considerations in context-aware systems. *Human–Computer Interaction* 16, 2-4 (2001), 193–212.

[9] Adrian Benton, Margaret Mitchell, and Dirk Hovy. 2017. Multi-task learning for mental health using social media text. *arXiv preprint arXiv:1712.03538* (2017).

[10] Michael L Birnbaum, Sindhu Kiranmai Ernala, Asra F Rizvi, Munmun De Choudhury, and John M Kane. 2017. A collaborative approach to identifying social media markers of schizophrenia by employing machine learning and clinical appraisals. *Journal of medical Internet research* 19, 8 (2017), e7956.

[11] Lindsay Blackwell, Jill Dimond, Sarita Schoenebeck, and Cliff Lampe. 2017. Classification and its consequences for online harassment: Design insights from heartmob. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW (2017), 1–19.

[12] Geoffrey C Bowker, Karen Baker, Florence Millerand, and David Ribes. 2009. Toward information infrastructure studies: Ways of knowing in a networked environment. In *International handbook of internet research*. Springer, 97–117.

[13] Geoffrey C Bowker and Susan Leigh Star. 2000. *Sorting things out: Classification and its consequences*. MIT press.

[14] Duy Duc An Bui and Qing Zeng-Treitler. 2014. Learning regular expressions for clinical text classification. *Journal of the American Medical Informatics Association* 21, 5 (2014), 850–857.

[15] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*. PMLR, 77–91.

[16] Stevie Chancellor. 2022. Towards Practices for Human-Centered Machine Learning. *arXiv preprint arXiv:2203.00432* (2022).

[17] Stevie Chancellor, Michael L Birnbaum, Eric D Caine, Vincent MB Silenzio, and Munmun De Choudhury. 2019. A taxonomy of ethical tensions in inferring mental health states from social media. In *Proceedings of the conference on fairness, accountability, and transparency*. 79–88.

[18] Stevie Chancellor and Munmun De Choudhury. 2020. Methods in predictive techniques for mental health status on social media: a critical review. *NPJ digital medicine* 3, 1 (2020), 1–11.

[19] Stevie Chancellor, Andrea Hu, and Munmun De Choudhury. 2018. 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*. 1–14.

[20] Stevie Chancellor, Yannis Kalantidis, Jessica A Pater, Munmun De Choudhury, and David A Shamma. 2017. Multimodal classification of moderated online pro-eating disorder content. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 3213–3226.

[21] Stevie Chancellor, Jessica Annette Pater, Trustin Clear, Eric Gilbert, and Munmun De Choudhury. 2016. # thyghgapp: Instagram content moderation and lexical variation in pro-eating disorder communities. In *Proceedings of the 19th ACM conference on computer-supported cooperative work & social computing*. 1201–1213.

[22] Eshwar Chandrasekharan, Mattia Samory, Anirudh Srinivasan, and Eric Gilbert. 2017. 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*. 3175–3187.

[23] Eli Clare. 2017. *Brilliant imperfection: Grappling with cure*. Duke University Press.

[24] Glen Coppersmith, Mark Dredze, and Craig Harman. 2014. Quantifying mental health signals in Twitter. In *Proceedings of the workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality*. 51–60.

[25] Glen Coppersmith, Mark Dredze, Craig Harman, and Kristy Hollingshead. 2015. From ADHD to SAD: Analyzing the language of mental health on Twitter through self-reported diagnoses. In *Proceedings of the 2nd workshop on computational linguistics and clinical psychology: from linguistic signal to clinical reality*. 1–10.

[26] Glen Coppersmith, Craig Harman, and Mark Dredze. 2014. Measuring post traumatic stress disorder in Twitter. In *Eighth international AAAI conference on weblogs and social media*.

[27] Glen Coppersmith, Kim Ngo, Ryan Leary, and Anthony Wood. 2016. Exploratory analysis of social media prior to a suicide attempt. In *Proceedings of the third workshop on computational linguistics and clinical psychology*. 106–117.

[28] Patrick W. Corrigan, Fred E. Markowitz, and Amy C. Watson. 2004. Structural levels of mental illness stigma and discrimination. *Schizophrenia bulletin* 30, 3 (2004), 481–491.

[29] Kaitlin L Costello and Diana Floegel. 2020. "Predictive ads are not doctors": Mental health tracking and technology companies. *Proceedings of the Association for Information Science and Technology* 57, 1 (2020), e250.

[30] Vedant Das Swain, Victor Chen, Shrija Mishra, Stephen M Mattingly, Gregory D Abowd, and Munmun De Choudhury. 2022. Semantic Gap in Predicting Mental Wellbeing through Passive Sensing. In *CHI Conference on Human Factors in Computing Systems*. 1–16.

[31] Norberto Nuno Gomes de Andrade, Dave Pawson, Dan Muriello, Lizzy Donahue, and Jennifer Guadagno. 2018. Ethics and artificial intelligence: suicide prevention on Facebook. *Philosophy & Technology* 31, 4 (2018), 669–684.

[32] Munmun De Choudhury. 2015. Anorexia on tumblr: A characterization study. In *Proceedings of the 5th international conference on digital health 2015*. 43–50.

[33] Munmun De Choudhury, Scott Counts, Eric J Horvitz, and Aaron Hoff. 2014. Characterizing and predicting postpartum depression from shared facebook data. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. 626–638.

[34] Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting depression via social media. In *Seventh international AAAI conference on weblogs and social media*.
[35] Alex J DeGrave, Joseph D Janizek, and Su-In Lee. 2021. AI for radiographic COVID-19 detection selects shortcuts over signal. *Nature Machine Intelligence* 3, 7 (2021), 610–619.
[36] Emily Denton, Alex Hanna, Razvan Amironesei, Andrew Smart, Hilary Nicole, and Morgan Klaus Scheuerman. 2020. Bringing the people back in: Contesting benchmark machine learning datasets. *arXiv preprint arXiv:2007.07399* (2020).
[37] Robert F DeVellis and Carolyn T Thorpe. 2021. *Scale development: Theory and applications*. Sage publications.
[38] Fifth Edition et al. 2013. Diagnostic and statistical manual of mental disorders. *Am Psychiatric Assoc* 21, 21 (2013), 591–643.
[39] Sindhu Kiranmai Ernala, Michael L Birnbaum, Kristin A Candan, Asra F Rizvi, William A Sterling, John M Kane, and Munmun De Choudhury. 2019. Methodological gaps in predicting mental health states from social media: triangulating diagnostic signals. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–16.
[40] Sindhu Kiranmai Ernala, Tristan Labetoulle, Fred Bane, Michael L Birnbaum, Asra F Rizvi, John M Kane, and Munmun De Choudhury. 2018. Characterizing audience engagement and assessing its impact on social media disclosures of mental illnesses. In *Twelfth international AAAI conference on web and social media*.
[41] Jerry Alan Fails and Dan R Olsen Jr. 2003. Interactive machine learning. In *Proceedings of the 8th international conference on Intelligent user interfaces*. 39–45.
[42] Jessica L Feuston, Michael Ann DeVito, Morgan Klaus Scheuerman, Katy Weathington, Marianna Benitez, Bianca Z Perez, Lucy Sondheim, and Jed R Brubaker. 2022. " Do You Ladies Relate?": Experiences of Gender Diverse People in Online Eating Disorder Communities. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 1–32.
[43] Jessica L Feuston and Anne Marie Piper. 2018. Beyond the coded gaze: Analyzing expression of mental health and illness on instagram. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–21.
[44] Jessica L Feuston and Anne Marie Piper. 2019. Everyday experiences: small stories and mental illness on Instagram. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–14.
[45] Jessica L Feuston, Alex S Taylor, and Anne Marie Piper. 2020. Conformity of Eating Disorders through Content Moderation. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW1 (2020), 1–28.
[46] Casey Fiesler and Nicholas Proferes. 2018. "Participant" perceptions of Twitter research ethics. *Social Media+ Society* 4, 1 (2018), 2056305118763366.
[47] Michel Foucault. 2013. *History of madness*. Routledge.
[48] Arthur W Frank. 2013. *The wounded storyteller: Body, illness, and ethics*. University of Chicago Press.
[49] R Stuart Geiger, Kevin Yu, Yanlai Yang, Mindy Dai, Jie Qiu, Rebekah Tang, and Jenny Huang. 2020. Garbage in, garbage out? Do machine learning application papers in social computing report where human-labeled training data comes from?. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 325–336.
[50] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence* 2, 11 (2020), 665–673.
[51] Ysabel Gerrard. 2018. Beyond the hashtag: Circumventing content moderation on social media. *New Media & Society* 20, 12 (2018), 4492–4511.
[52] David C Giles and Julie Newbold. 2011. Self-and other-diagnosis in user-led mental health online communities. *Qualitative Health Research* 21, 3 (2011), 419–428.
[53] Aaron Halfaker and R Stuart Geiger. 2020. Ores: Lowering barriers with participatory machine learning in wikipedia. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–37.
[54] Keith Harrigian, Carlos Aguirre, and Mark Dredze. 2020. Do Models of Mental Health Based on Social Media Data Generalize?. In *Proceedings of the 2020 conference on empirical methods in natural language processing: findings*. 3774–3788.
[55] Christina Harrington, Sheena Erete, and Anne Marie Piper. 2019. Deconstructing community-based collaborative design: Towards more equitable participatory design engagements. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–25.
[56] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudik, and Hanna Wallach. 2019. Improving fairness in machine learning systems: What do industry practitioners need?. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–16.
[57] Amina Hussain, Mishal Dar, and Kyle T Ganson. 2022. A post-structural feminist analysis of eating disorders intervention research. *Affilia* 37, 3 (2022), 505–519.
[58] Maia Jacobs, James Clawson, and Elizabeth D Mynatt. 2014. Cancer navigation: opportunities and challenges for facilitating the breast cancer journey. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. 1467–1478.
[59] Jialun Aaron Jiang, Kandrea Wade, Casey Fiesler, and Jed R Brubaker. 2021. Supporting serendipity: Opportunities and challenges for Human-AI Collaboration in qualitative analysis. *Proceedings of the ACM on Human-Computer Interaction*

5, CSCW1 (2021), 1–23.
[60] Dan Jurafsky. 2000. *Speech & language processing*. Pearson Education India.
[61] Harmanpreet Kaur, Eytan Adar, Eric Gilbert, and Cliff Lampe. 2022. Sensible AI: Re-imagining Interpretability and Explainability using Sensemaking Theory. *arXiv preprint arXiv:2205.05057* (2022).
[62] Harmanpreet Kaur, Daniel McDuff, Alex C Williams, Jaime Teevan, and Shamsi T Iqbal. 2022. "I Didn't Know I Looked Angry": Characterizing Observed Emotion and Reported Affect at Work. In *CHI Conference on Human Factors in Computing Systems*. 1–18.
[63] Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. 2020. Interpreting interpretability: understanding data scientists' use of interpretability tools for machine learning. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–14.
[64] Arthur Kleinman. 2020. *The illness narratives: Suffering, healing, and the human condition*. Basic books.
[65] Kate Loveys, Patrick Crutchley, Emily Wyatt, and Glen Coppersmith. 2017. Small but mighty: affective micropatterns for quantifying mental health from social media language. In *Proceedings of the fourth workshop on computational linguistics and clinical Psychology—From linguistic signal to clinical reality*. 85–95.
[66] Gianluca Maguolo and Loris Nanni. 2021. A critic evaluation of methods for COVID-19 automatic detection from X-ray images. *Information Fusion* 76 (2021), 1–7.
[67] Gabriela Marcu, Nadia Dowshen, Shuvadittya Saha, Ressa Reneth Sarreal, and Nazanin Andalibi. 2016. TreatYoSelf: Empathy-driven behavioral intervention for marginalized youth living with HIV. In *Proceedings of the 10th EAI International Conference on Pervasive Computing Technologies for Healthcare*. 69–76.
[68] Alice E Marwick and Danah Boyd. 2011. I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience. *New media & society* 13, 1 (2011), 114–133.
[69] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. 2019. Reliability and inter-rater reliability in qualitative research: Norms and guidelines for CSCW and HCI practice. *Proceedings of the ACM on human-computer interaction* 3, CSCW (2019), 1–23.
[70] Paul E Meehl. 1999. Clarifications about taxometric method. *Applied and Preventive Psychology* 8, 3 (1999), 165–174.
[71] Margaret Mitchell, Kristy Hollingshead, and Glen Coppersmith. 2015. Quantifying the language of schizophrenia in social media. In *Proceedings of the 2nd workshop on Computational linguistics and clinical psychology: From linguistic signal to clinical reality*. 11–20.
[72] Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kıcıman. 2019. Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data* 2 (2019), 13.
[73] Kayur Patel, James Fogarty, James A Landay, and Beverly L Harrison. 2008. Examining Difficulties Software Developers Encounter in the Adoption of Statistical Machine Learning.. In *AAAI*. 1563–1566.
[74] Jessica A Pater, Oliver L Haimson, Nazanin Andalibi, and Elizabeth D Mynatt. 2016. "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*. 1185–1200.
[75] Jessica A Pater, Lauren E Reining, Andrew D Miller, Tammy Toscos, and Elizabeth D Mynatt. 2019. " Notjustgirls" Exploring Male-related Eating Disordered Content across Social Media Platforms. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–13.
[76] Víctor M Prieto, Sergio Matos, Manuel Alvarez, Fidel Cacheda, and José Luís Oliveira. 2014. Twitter: a good place to detect health conditions. *PloS one* 9, 1 (2014), e86191.
[77] Inioluwa Deborah Raji, I Elizabeth Kumar, Aaron Horowitz, and Andrew Selbst. 2022. The Fallacy of AI Functionality. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. 959–972.
[78] Donghao Ren, Saleema Amershi, Bongshin Lee, Jina Suh, and Jason D Williams. 2016. Squares: Supporting interactive performance analysis for multiclass classifiers. *IEEE transactions on visualization and computer graphics* 23, 1 (2016), 61–70.
[79] Paul Röttger, Bertie Vidgen, Dirk Hovy, and Janet B Pierrehumbert. 2021. Two Contrasting Data Annotation Paradigms for Subjective NLP Tasks. *arXiv preprint arXiv:2112.07475* (2021).
[80] Shiori Sagawa, Aditi Raghunathan, Pang Wei Koh, and Percy Liang. 2020. An investigation of why overparameterization exacerbates spurious correlations. In *International Conference on Machine Learning*. PMLR, 8346–8356.
[81] Johnny Saldaña. 2009. *The coding manual for qualitative researchers*. SAGE Publications.
[82] Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. 2021. "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. In *proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–15.
[83] Morgan Klaus Scheuerman, Alex Hanna, and Emily Denton. 2021. Do datasets have politics? Disciplinary values in computer vision dataset development. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–37.
[84] Morgan Klaus Scheuerman, Jacob M Paul, and Jed R Brubaker. 2019. How computers see gender: An evaluation of gender classification in commercial facial analysis services. *Proceedings of the ACM on Human-Computer Interaction* 3,

CSCW (2019), 1–33.
[85] Andrew Scull. 2015. *Madness in Civilization: A Cultural History of Insanity, from the Bible to Freud, from the Madhouse to Modern Medicine*. Princeton University Press. 432 pages.
[86] Indira Sen, Fabian Flöck, Katrin Weller, Bernd Weiß, and Claudia Wagner. 2021. A total error framework for digital traces of human behavior on online platforms. *Public Opinion Quarterly* 85, S1 (2021), 399–422.
[87] C Estelle Smith, Bowen Yu, Anjali Srivastava, Aaron Halfaker, Loren Terveen, and Haiyi Zhu. 2020. Keeping community in the loop: Understanding wikipedia stakeholder values for machine learning-based systems. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.
[88] Ian Stewart, Stevie Chancellor, Munmun De Choudhury, and Jacob Eisenstein. 2017. # anorexia,# anarexia,# anarexyia: Characterizing online community practices with orthographic variation. In *2017 IEEE International Conference on Big Data (Big Data)*. IEEE, 4353–4361.
[89] TS Szasz. 1974. The myth of mental illness: Foundations of a theory of personal conduct. HarperPerennial.
[90] Anja Thieme, Danielle Belgrave, and Gavin Doherty. 2020. Machine learning in mental health: A systematic review of the HCI literature to support the development of effective and implementable ML systems. *ACM Transactions on Computer-Human Interaction (TOCHI)* 27, 5 (2020), 1–53.
[91] Philip S Wang, Patricia A Berglund, Mark Olfson, and Ronald C Kessler. 2004. Delays in initial treatment contact after first onset of a mental disorder. *Health services research* 39, 2 (2004), 393–416.
[92] Tao Wang, Markus Brede, Antonella Ianni, and Emmanouil Mentzakis. 2017. Detecting and characterizing eating-disorder communities on social media. In *Proceedings of the Tenth ACM International conference on web search and data mining*. 91–100.
[93] Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel S Weld. 2019. Errudite: Scalable, reproducible, and testable error analysis. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 747–763.
[94] Andrew Yates, Arman Cohan, and Nazli Goharian. 2017. Depression and self-harm risk assessment in online forums. *arXiv preprint arXiv:1709.01848* (2017).
[95] Anon Ymous, Katta Spiel, Os Keyes, Rua M Williams, Judith Good, Eva Hornecker, and Cynthia L Bennett. 2020. " I am just terrified of my future"—Epistemic Violence in Disability Related Technology Research. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–16.
[96] Jun Yuan, Jesse Vig, and Nazneen Rajani. 2022. iSEA: An Interactive Pipeline for Semantic Error Analysis of NLP Models. In *27th International Conference on Intelligent User Interfaces*. 878–888.
[97] Haiyi Zhu, Bowen Yu, Aaron Halfaker, and Loren Terveen. 2018. Value-sensitive algorithm design: Method, case study, and lessons. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–23.

Received January 2023; revised April 2023; accepted May 2023