npj | Digital Medicine

www.nature.com/npjdigitalmed

REVIEW ARTICLE OPEN

Methods in predictive techniques for mental health status on social media: a critical review

Stevie Chancellor[1] ✉ and Munmun De Choudhury[2]

Social media is now being used to model mental well-being, and for understanding health outcomes. Computer scientists are now using quantitative techniques to predict the presence of specific mental disorders and symptomatology, such as depression, suicidality, and anxiety. This research promises great benefits to monitoring efforts, diagnostics, and intervention design for these mental health statuses. Yet, there is no standardized process for evaluating the validity of this research and the methods adopted in the design of these studies. We conduct a systematic literature review of the state-of-the-art in predicting mental health status using social media data, focusing on characteristics of the study design, methods, and research design. We find 75 studies in this area published between 2013 and 2018. Our results outline the methods of data annotation for mental health status, data collection and quality management, pre-processing and feature selection, and model selection and verification. Despite growing interest in this field, we identify concerning trends around construct validity, and a lack of reflection in the methods used to operationalize and identify mental health status. We provide some recommendations to address these challenges, including a list of proposed reporting standards for publications and collaboration opportunities in this interdisciplinary space.

npj Digital Medicine (2020)3:43 ; https://doi.org/10.1038/s41746-020-0233-7

INTRODUCTION

Researchers in computer science (CS) are using behavioral and linguistic cues from social media data to predict the presence of mood and psychosocial disorders. Since 2013, research can assess the presence of major depression[1–3], suicidality[4–6], eating disorders[7,8], and schizophrenia[9], among others with high accuracy (80–90%). In addition to mental disorders, these approaches are starting to assess related symptomatology, such as self-harm[8], stress[10], and the severity of mental illness[11] without the use of in-person, clinical assessment. These signals are taken from the posting and behavioral history of social media websites and apps, such as Twitter, Reddit, and Facebook[12]. In this article, we adopt the term mental health status (MHS) to capture both mental disorders and these related symptomatology.

The benefits of these computational approaches to understanding MHS could be profound—for new data to supplement clinical care, assessing developing conditions, identifying risky behaviors, providing timely interventions, or reaching populations difficult to access through traditional clinical approaches. In fact, approaches like this have been adopted by platforms such as Facebook for suicide prevention efforts[13,14]. Complementary enthusiasm has surfaced in an emergent area known as "digital psychiatry"[15], which leverages these predictive signals to improve mental health service outcomes.

In this new interdisciplinary space, there are few shared guidelines for what constitutes valid assessment of MHS in social media. Methods and insights for this work are drawn from interdisciplinary areas such as health informatics, machine learning, artificial intelligence, natural language processing, and human-computer interaction. Previous work in these domains has focused on abstract notions of ethics and methodological rigor to understand public health using social media data[16–19]. Reviews and meta-analyses have examined the expression of depression and anxiety in social media[20]; subjective mood, well-being, and mental health in social media[21,22] and other non-clinical texts[23]; and the development of technology more broadly for mental and affective health[24–26]. Nevertheless, recent research has noted a lack of grounded recommendations detailing and evaluating current practices for building algorithms to predict MHS in social media data[16,27].

Given the nascence of this field, we see incredible value in identifying trends in research methods and practices to identifying gaps before they systemically emerge in research paradigms. These issues are important not only as they reflect scholarly research quality, but also because, most importantly, the implications predicting MHS can have on individuals who may be the object of such predictions in clinical care and social media settings.

This article provides a critical review of methods in predicting MHS on social media, identifying 75 papers published between 2013 and 2018. We report on patterns of data annotation and collection, data bias management, pre-processing and feature selection, model selection, and validation. Our results reveal that there are issues in evaluating construct validity to determine and predict MHS that permeate the research process. We argue that this will inhibit reproducibility and extension of this work into practical and clinical domains, and we provide recommendations on how to begin to alleviate these problems.

CORPUS OVERVIEW

Figure 1 shows the years of activity in publication. The first research was published in 2013, with eight papers in total[1,28–34]. This area is showing rapid growth, with 19 papers in 2017[3,8,10,35–49] and 16 in 2018[6,50–64].

We identified the social media platforms in these studies, summarized in Fig. 2. The most popular social media site for this analysis was Twitter, with a substantial portion (30/75) of the

[1]Department of Computer Science, Northwestern University, Evanston, IL, USA. [2]School of Interactive Computing, Georgia Tech, Atlanta, GA 30308, USA. ✉email: stevie@northwestern.edu

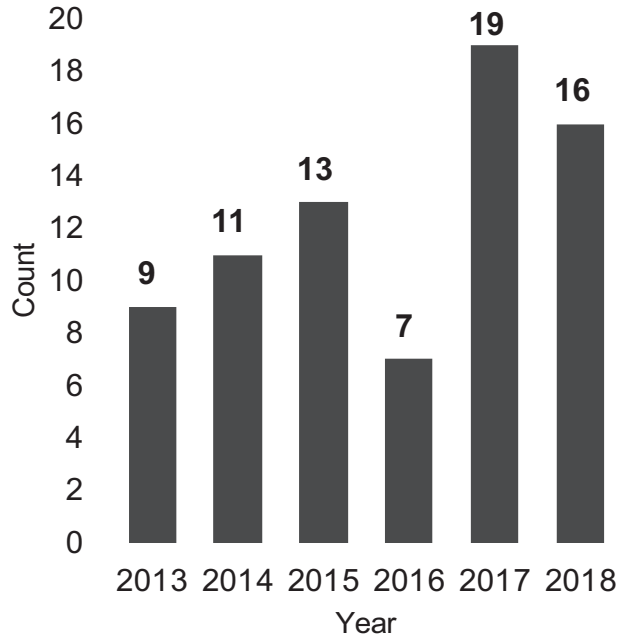

Fig. 1 Publication counts by year. In this graph, we display the publication counts in our corpus from 2013 to 2018.

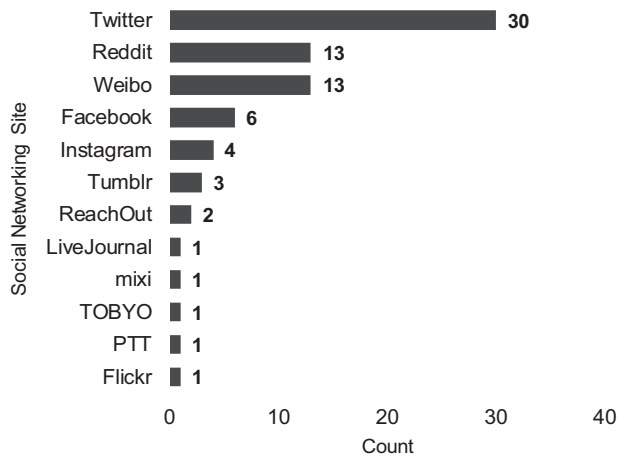

Fig. 2 Publication counts by social networking site. In this graph, we display the counts of publications, organized by the various social networking sites studied.

corpus studying this site (e.g. refs. [65,66]). Other popular sites include Sina Weibo (13)[8,10,39,50,67–75], Reddit (13)[6,41,46,48,52,54,55,58–60,63,64,76], Facebook (6)[33,51,53,56,77,78], Instagram (4)[3,11,38,62], Tumblr (3)[7,44,79], and ReachOut (2)[52,61]. Single papers inspect Flickr[8], PTT[28], mixi[29], LiveJournal[80], and TOBYO Toshoshitsu[81]. Year-over-year, Twitter was the dominant social media site examined in the corpus.

We also identified the representation of languages in publications. The majority of studies are done on English data (54) (e.g. ref. [80]), followed by Chinese (14)[10,28,31,39,50,67–75], Japanese (4)[2,29,32,81], Spanish and Portuguese (1)[82], and two that were not easily identified[38,47].

Disorders and symptomatology

Next, we examined the disorders and symptomatology in each of the 75 papers. Eight papers studied more than one condition[36–38,48,65,83–85], so we report the counts of unique disorders and symptomatology examined in Fig. 3.

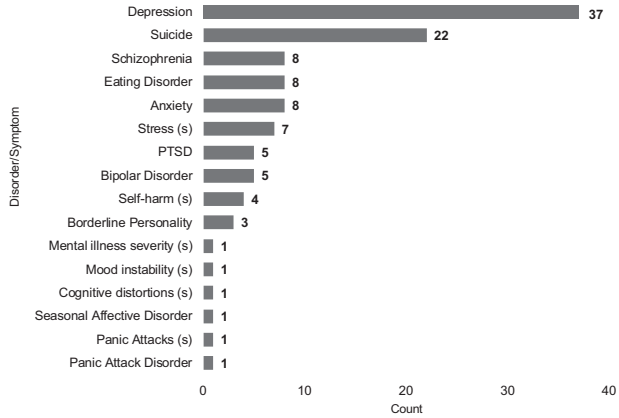

Fig. 3 Publication counts by disorder and symptomatology. In this graph, we display the counts of publications that study specific disorders and symptomatology.

Nearly half of the studies in the dataset (37/75) examined depression. Examples included studying depression generally[28,81,83,86], major depressive disorder[1], postpartum depression[30,78], degree or severity of depression[77], and depression as a risk factor for suicidality[39].

We also found that 22 papers studied suicidality[4–6,29,36,39,48,49,59–61,66,68–70,72,76,80,87,88]. Cases included whether someone has suicidal ideation/is suicidal[29], will attempt suicide[4,36,68], or may shift to suicidal ideation in the future[76]. Other research looked at risk factors for suicide[39,87], using crowdworkers for annotations of suicide risk[6], and distinguishing between suicidal ideation and other discussions of suicide[5].

Eight studies considered eating disorders[7,8,37,38,63,79,82,85], most in the general case[8,37,38,63,82,85], and two focusing on anorexia[7,79]. Another eight examined schizophrenia[9,36,37,43,48,59,85,89]. Lastly, eight studies used social media data to study anxiety, some in the context of other disorders[36,37,48,59,85] and others more specifically[46,54,64].

Other disorders and conditions investigated in our corpus included bipolar disorder (5)[37,48,65,83,84], post-traumatic stress disorder (PTSD) (5)[37,83–85,90], borderline personality disorder (3)[59,65,85], and panic disorder (1)[37].

Many studies also analyzed symptomatology related to mental disorders. This primarily focused on predicting stress (7/75)[10,39,41,67,71,73,75]. We also saw studies on self-harm (4)[48,52,59,91], panic attacks (1)[36], cognitive distortions (1)[44], mood instability (1)[40], and mental illness severity (1)[11].

RESULTS

In this section, we summarize our findings about the corpus. Broadly, the field frames their study design and research investigations around prediction of mental illness from text and behavioral cues extracted from social media data. Almost all papers (69) conceptualized their research questions as a classification problem through binary classification (63/69), such as the categorical distinction between high and low stress[40]. Six papers used multi-class schema instead of binary classification[5,6,11,48,49,52]. Six papers used a model that predicts continuous or discrete values[32,38,53,69,77]. We also found that most studies (47/75) examine the individual/user as the object of prediction, such as predicting suicide risk of a person[39]. Twenty-five studies predicted mental health status per post or aggregated posts (e.g. refs. [11,60]) and then, by proxy, inferring the MHS of the owner of those accounts. One paper examined both[42].

To begin, in the absence of clinical assessment and in-person diagnosis, researchers have adopted alternative signals to identify positive and negative MHS. In this section, we examine questions of construct validity, or how the publications in the corpus validate the presence or absence of MHS.

Establishing ground truth for positive annotation

We identified six methods of annotation for positive sign of MHS.

- Human Assessments (27). Many examinations asked humans to annotate the presence of MHS in a dataset. Domain experts, such as practicing clinicians or psychologists, were often called to annotate or label data[38,70]. For example, one study assessed depression through clinical interviews[31]. In other scenarios, CS researchers conducted the annotations[37,42]. Often, both domain experts and CS researchers partnered to annotate together[43,87]. Finally, some researchers used workers from crowdsourcing sites such as Amazon Mechanical Turk to identify status[5] or verify the veracity of MHS downstream after another protocol[30].
- Community or Network Affiliations (25). Researchers looked for community/network participation (e.g. refs. [48,54]) to indicate MHS. Community participation was used as signal in social networks with formal communities, such as participating in communities about mental health on LiveJournal[80], Reddit[41,46,48,59], or posting in a suicide crisis community/forum[76]. These measures also included network signals such as following another account on Twitter[65,89]. Other studies use the signal of hashtags on apps like Instagram[11,38].
- Self-Disclosure (17). This approaches searched for individuals to state that they suffer from a specific condition or are engaging in behaviors indicative of MHS[4,30,35,36,40,43,47,50,52,67,71,73,78,83,85,89,90]. These were triangulated with specific expressions, like "I was diagnosed with..."[83,90]. Positive annotation included stating that have a specific illness, like post-traumatic stress disorder[83] or schizophrenia[43]. Work also examined self-reports of antidepressant medication usage[35], attempts to take their own life[4], or self-described as being stressed or relaxed[67].
- Administering Screening Questionnaires (14). Another popular technique was administering screening tools and questionnaires to consenting participants[1–3,32,33,39,45,51,53,62,66,69,72,77]. These included screeners that can measure depression, including the Center for Epidemiologic Studies Depression Scale (CES-D)[3,34,51], Beck Depression Inventory (BDI)[1,2,33], Patient Health Questionnaire (PHQ-9)[53,78] and PHQ-8[62], and Zung Self-Rating Depression Scale (SDS)[32]. Screeners were also used for other mental health status, such as suicidality[66,72].
- Keyword use (10). Another approach used the presence of keywords or phrases[10,28,35,42,65,71,75,81,82,88]. Researchers used dictionaries connecting to suicide[88] or stress[10,71]. Researchers also used symptom words and names of disorders on Twitter posts or profiles[42,82], behaviors associated with disorders (like "ultimate goal weight"[8]), or if people use phrases associated with life events (e.g. childbirth)[30].
- Acquired Annotations (9). Several publications acquired annotations from previously published research[31,37,50,57,84,86] or shared tasks[49,52,61].
- News Reports (2). Two studies looked at news reports of who had died by suicide to identify victims' names, then find social media data on these individuals[68,70].
- Medical Diagnostic Codes (1). One research study used the presence of the International Statistical Classification of Diseases and Related Health Problems 10th edition (ICD-10) codes from patient health records to detect depression[56].

Some papers (33/75) took the results of the initial proxy assessments at face value (e.g. refs. [41,46,80]). We noted that acquired datasets were often assumed to have high-quality labels, and the annotations were taken to be accurate[61], as well as the use of screener questionnaires[45,51]. However, most studies (40/55) combined two approaches listed above to acquire a precise sample. Human annotation was a popular follow-up, with the validity of initial results of keyword matching often manually verified by researchers[54,65,88]. Other approaches used human verification to ensure that self-disclosure was genuine[9,42,43]. Two papers combined three ground truth assessment to triangulate MHS[4,66]. There was no reflection across the documents on what ground truth approach was appropriate for establishing construct validity, nor how many approaches combined together would accurately assess the MHS of interest. There was also no validation of applying constructs to social media data, for instance how strongly clinically valid screening questionnaires evaluate social media data.

SOURCE OF CONTROL DATA/NEGATIVE EXAMPLES

Publications also sourced and design negative/control data for predictive tasks—these procedures were often different than the methods to identify positive signs of MHS.

- Validated No MHS (29). Many papers engineered ways to validate that the negative dataset did not contain people with the MHS of interest, e.g. refs. [1,72,73]. This often was taking the lower bounds of cutoff from screening participants with screeners[2,33,51]. Other approaches used an expert to validate that there was an absence of MHS and no concerning symptoms disclosed in social media, such as no diagnosis of schizophrenia[43].
- Random Selection of Control Users (23). Many studies constructed a negative/control dataset from randomly sampled data on the social media platform[4,6,8,9,29,35–37,41,43,46,47,54,59,68,70,79,81,83,85,89–91]. This random sampling came from historical samples of data, like the YFCC100m (Yahoo! Flickr Creative Commons 100 million) dataset[91] or other collections[83]. Others gathered randomly, such as from random Tumblr users[79] or the front page of Reddit[41,54].
- Lack of Mental Health Disclosure (22). These studies used a lack or absence of disclosure of MHS as source for negative data[28,29,37,38,42,45,47,48,50,52,58,63–65,68,76,80,81,87,89]. Examples included sampling people who did not disclose having a condition[65,89] or did not participate in communities related to mental health[28,80].
- Matching Strategies (8). Some research took randomly sampled users and constructed matched samples along demographic/behavioral characteristics of the positively identified users[4,8,9,36,85,89]. This included matching on inferred traits, like age and gender[4,8,85], engagement on the platform[52], or time-matching controls[36]. One study matched controls on health information provided through electronic health records[56].
- Acquired from Other Sources (3). Some research acquired datasets from alternative sources, boosting the size or scope of their dataset with other data[49,57,61].

Managing data quality and sampling strategies

Next, we report on our study of data quality, or how documents in our corpus curated the dataset for higher quality results. In our corpus, 53/75 studies filtered to manage issues of data bias or quality in their datasets:

- Platform Behavior Thresholds (28). Researchers described removing data for not meeting minimum content or engagement thresholds, e.g. refs. [69,89]. This included behaviors such as having an account on the site of interest[1,78]. Most studies had minimum activity thresholds, such as a minimum number of posts[64,65,83]. Others looked for minimum friends/relationships[29], engagement from others on a thread[61], or platform engagement over time[29,36,52].
- Legitimate Mental Health Mentions (17). These studies validated disclosures of MHS[4,5,7–9,11,34,38,41,76,78,82,83,87–89] Some had strict thresholds on the precision of positive MHS[8,82] or the time frame in which certain behaviors could occur[76]. For instance, one study looked for suicide attempts with discernible dates[4]. Others removed individuals for participating in eating disorder recovery communities, which confounded presence of an active eating disorder[7,11].
- Restriction on Participant Characteristics (14). These studies excluded individuals based on certain characteristics or traits[1,3,33,36,39,40,45,51,53,62,66,72,73,88], such as age[39,73] or posts in English[51,62]. Other studies filtered participants on crowdsourcing sites based on overall approval ratings or a minimum number of previous tasks completed[3,45].
- Quality Control During Online Surveys (7). Another threshold was removing participants for not passing quality control measures on the surveys, especially on surveys given through crowdsourcing sites such as Amazon Mechanical Turk or Crowdflower[3,33,34,51,66,69,72]. This included filtering surveys completed too fast[34,69], who did not pass attention checks during the survey[3,66], or did not finish the survey[33,51,72].
- Removing Spurious Data (6). Other studies removed spurious data[39,66,72,81,88,89], such as duplicate survey responses[39] or gibberish[88]. One study mentioned removing advertisements[81], and two removed spam[81,89].

We did not notice any larger dataset adjustments to account for other kinds of biases, as noted by Olteanu et al.[92] We inspected for whether studies adjusted for sampling bias or confounding factors with limited access APIs, adjusted for other clinically-relevant signals (such as demographics), took alternative data sampling strategies (such as selective rather than random sampling), or removed adversarial content, bots, or outlier/famous accounts (such as celebrities). Other than two papers that removed spam and advertisements[81,89], we did not notice any corrections in the dataset for these factors. We also did not see larger analyses or adjustments to datasets to ensure that the samples were representative or accounted for population-level trends. The only management of these biases were in matching strategies to assemble negative datasets, e.g. refs. [4,9,36].

Variable selection/feature engineering

Next, we examined patterns and characteristics of the data characteristics relevant for prediction. This is often referred to as variable selection or, in the machine learning community as "feature engineering". In all, 42/75 studies reported the total number of features—of those 42 papers, the range of the number of features ranged from 7[11,29] to over 15,000[76].

- Language Features (68/75).
  - Structural/Syntactic (25). We found features that describe the structural or syntactic composition of social media posts, (e.g. refs. [6,72,89], such as the length of the post[39,76], part-of-speech tagging[5], and modality tagging[81]. We also saw counts of specific characters, like emoticons[89]. One study used the length and number of numeric characters in the domain name of a blogging site[72].
  - Character and Word Models (38). These representations of language draw on probabilistic distributions of character and word patterns within text, e.g. refs. [4,32]. This included *n*-gram use[87], character modeling[68], bag-of-words models[2], term-frequency-inverse document frequency (TF-IDF)[28], and word embeddings[38]. We also saw deep learning approaches to modeling language through convolutional neural networks[52].
  - Topical (14). Other documents engineered features using topic modeling to identify meaningful connections between concepts in datasets[2,6,47,49,56,61,69,70,77,83,84,86,87]. This included the popular Latent Dirichlet Allocation (LDA) topic model[84,86], and Brown clustering[9].
  - Linguistic Style (18). Some studies used considered linguistic style and content measures as features[1,6,9,30,34,40,42,43,45,49,59,73,76,78–80,83,91]. Research used style categories from the Linguistic Inquiry and Word Count (LIWC) dictionaries[34,80]. We also noticed the study of readability, coherence, and perplexity measures[9,42], as well as subjectivity measures from MPQA[49] and TextBlob[59].
  - Domain-Specific (13). Studies designed domain-specific linguistic features to evaluate text documents[1,5,42,47,49,58,71,76,80,81,83,89,91]. This included constructing dictionaries or lexicons related to depression[42,58,72], self-harm[91], suicide[5], and stress[71]. This also included assessing user-generated mood tags taken from LiveJournal[80] as well as explicit mentions of medication[1,46]. One study designed features around the final sentence as an indicator of suicidality or intent[49].
  - General Language Measures (18). Papers also described generic language measures[5,6,8–10,39,44,46,49–51,60,61,64,66,69,77,84], such as the LIWC library in its entirety.
- Behavior (37/75).
  - Activity (35). Features also tracked behavioral activity of the individual, e.g. refs. [33–35]. Posting frequencies were a source of interest[4], including volume of posts[76], posting rates[65], and temporal distributions of posting history[28]. Studies also examined platform-specific features, like geotagged posts[33].
  - Interaction (31). Interactions with others on the platform were another common feature source, e.g. refs. [61,67,79]. This included uni-directional follower/followee relationships[47,89] and bi-directional friendships[33]. Papers also examined community membership/affiliation or participation[8,46], platform affordances like Twitter's retweet/quote, mentions/replies features[65], or participation in threads from others[61]. Some other studies examined group membership as a variable[51].
  - Network (6). Studies analyzed the network or graph structures for an individual's social network[1,10,29,35,73,74], including clustering coefficients and homophily[29], strong and weak ties[10], and network size, density, and depth[1,35,73].
  - Domain-Specific (8). In addition to general behavioral features, publications also engineered domain-specific activity measures[1,10,29,31,51,61,72,83]. These features focused on measuring posting between the night hours, quantified as the "insomnia index"[1]. Another paper examined suicide homophily, or the number of friends who had died by suicide[29]. One study used previous evaluations of well-being on a crisis site in the predictive features[61].
- Emotion and Cognition (38/75).
  - Sentiment, Affect, and Valence (36). Many papers examined peoples' expressed mood, sentiment, and intensity of emotion, e.g. refs. [41,53,62]. This was measured with

sentiment scoring mechanisms like ANEW[80], LIWC[7,78], LabMT[62], TextBlob[60,61], and VADER[36]. Other studies examined affect and intensity[30], polarity of emotions on more complex scales[53], or counted the positive and negative emoticons[8,73].
  - Psycholinguistic (11). Researchers also use psycholinguistic evaluations of emotional status from language[7,10,40,43,45,53,67,72,79,80,83], using categories of emotional speech (such as anger or anxiety in LIWC)[80,83].
  - Domain-specific (4). Domain-specific applications of emotion and cognition measurements included measuring personality traits via Big 5[84], behavior theories of anorexia recovery[7], a lexicon of emotional words related to mental distress[6], and Tweets related to depression[42].
- Demographic Features (11). Papers also incorporated data about personal demographics into variable selection[1,33,37,38,50,51,65,72,78,84]. This included age and gender[51,65,72], education, income, and relationship status[1,47]. Some of these were not gathered from individuals in the dataset; rather, they were inferred using computational means[47,84].
- Image Features (8). Researchers extracted visual information from the images of posts[3,10,38,47,50,67,75,91]. This included color themes/Hue-Saturation-Value (HSV) values[3,50,67], if the image includes a face[3], brightness and saturation values[10,47], and the types of colors used[47,75]. This also included data extracted from a convolutional analysis of the images[38,91].

For feature reduction or selection techniques, 26/75 described reducing features to salient ones, such as[5,39,82] The most popular feature reduction technique was dimensionality reduction through Principal Component Analysis (PCA)[77,89]. Other feature selection methods included experimentally removing features[42], deep learning-based reductions through convolution or GRUs (Gated Recurrent Units)[52,58], feature ablation[9], stepwise regression[39], and taking *k*-best features[43].

Algorithm selection

Nearly all papers frame their contributions as predicting MHS; in that vein, most documents choose algorithms from machine learning and statistical modeling, and highlight prediction results in their findings. Two papers chose their algorithms for their ability to assess correlations between features[33,53]. No papers used pseudo-causal or causal approaches to make claims.

There was high diversity in algorithm selection, of which 73/75 papers reported on their algorithm of choice. The most popular predictive algorithm was Support Vector Machines, used by 24 projects[1,2,6,8,9,28,30,34,39–42,51,54,55,60,68,70,79,79,81,86–89,93]. Fifteen studies used logistic regression[4,11,29,44,56,60,61,63,64,72,72,73,76,78,80,82]. Next was Random Forest at seven papers in the corpus[3,5,36,43,45,65,72], and one who used a Rotation Forest (a boosted version of Random Forest)[5]. We also saw the use of decision trees (2)[35,66], Naïve Bayes (2)[31,82], and XGBoost[49]. Finally, we found the use of regression techniques for some studies (8)[7,33,53,62,69,77,90,90]. This included the use of linear regressions[62,69,77], log-linear regression[83,90], correlational analyses[33,53], and survival analysis/Cox regression[7].

Deep learning has been a more recent trend, with 14 papers using this technique[10,37,38,46–48,50,52,57–59,67,75,91]. Some papers used more straightforward deep neural networks[8,46,67], some with additional convolutional layers[48], or recurrent neural networks[58,59]. Other research adopted a multitask neural network to share information between prediction tasks[37,71].

How were these algorithms selected for use? In all, 41/75 papers described their process for selecting their algorithm of choice. The vast majority of algorithms (30/41) were selected because they performed the best, e.g. refs. [3,34,50], experimentally chosen across several algorithmic options[34,42]. Other reasons offered were the suitability of the model to the research task, such as sharing knowledge between tasks[37], stability of model training[52], interpretable features for clinicians and other stakeholders[63,66], or dropout impacting the use of standard regression techniques[7]. Others drew from theoretical and practical reasons to select their models[5], such as the "no free lunch theorem"[44].

Validating algorithms and reporting performance

72/75 papers reported how they validated the models, the most popular of which was using *k*-fold cross validation. Fifty-four papers use this technique, with a *k* ranging from 5[40], 10[82], 20[62] to leave-one-out[39,66]. Another common technique (20/72) was holding out blind data as a test set and reporting performance[4,11,42,43,48,50,52,57–59,63,73,76,77,86–89,91]; held-out dataset size ranged from 10%[88] to 30–40%[69,91]. Next were multiple experimental runs of the model (14/72)[1–3,10,30,34,45,47,50,51,60,67,72,79], ranging from 5[45] to 1000[79] runs. Three studies used model fit measures to validate the fit of the model, such as deviance for regression fit[7,11,29] and feature relevance or curation techniques like stepwise regression to prevent overfitting[29,32].

Many papers combined multiple validation techniques, the most common was cross-validating their test data and reporting results on a held-out dataset[30,88] or pairing cross-validation with multiple experimental runs[34,72].

Finally, 70/75 papers reported performance in a way that can be evaluated and benchmarked by other research. The best performance tended to be measured on machine learning metrics such as accuracy[46,51,80], precision and recall[45,86,89], F1 (a harmonic between precision and recall)[52,76], and area under the curve (AUC)[56,62]. We occasionally found the use of regression-oriented measures, such as root mean squared error (RMSE)[69] and R^2[77]. We very rarely saw use of popular metrics from other domains, such as sensitivity, specificity (or false positive/negative rates), and positive and negative predictive value[37]—the machine-learning oriented metrics dominated reporting standards.

Essential reporting of prediction technique details

Last, we studied the reporting of essential information required to reproduce a predictive algorithm, which are de facto minimum standards to evaluate an approach. We identified five crucial factors essential to running any regression model or machine learning approach. These are: the number of samples/data points, number of variables/features, the predictive approach (either a specific algorithm or regression type), a method for validation, and the metric used to evaluate performance. We then counted the number of papers that explicitly reported on these five items:

- 71/75 number of samples/data points.
- 42/75 number of variables/features.
- 73/75 algorithm or regression of choice.
- 72/75 at least one validation method.
- 70/75 explicit performance or fit metrics.

We noticed that the most commonly omitted variable was the dimensionality or number of variables in the feature/variable space. For those that omitted this information, studies would describe what features were being included (such as word embeddings representation of the social media posts, or language models built on top of the post content), yet not include the size or number of their feature vectors. In five papers, we had difficulty assessing the performance of the selected regression or classification algorithm because the authors included this information on poorly-labeled graphs or figures. It was not possible in these graphs to assess the precise performance or fit of the model to the data, and we were forced to estimate from bar charts' bands of performance, i.e. (80–85% F1).

Finally, we studied the patterns of reporting for all minimum standards across the dataset. If each paper is examined for the presence of these five traits, only 32/75 papers, or 42%, successfully reported all five measures. If we examined for four of five criteria, 67/75 papers, or 89%, reported on at least four of five criteria.

DISCUSSION

Our results demonstrate the variety of study design, methods techniques, and reporting schema to understand mental health status (MHS) of individuals through their social media data. Despite these innovations in techniques, we noticed concerning trends around construct validity with the identification and prediction of MHS in our corpus. Shadish et al. define construct validity as "making inferences from the sampling particulars of a study to the higher-order constructs they represent"[94]—said another way, this type of experimental validity maps theoretical constructs of knowledge to the observed phenomenon within the dataset. The challenges of construct validity in observational social media research in particular has been recognized[92,95,96]. These issues of construct validity risks deviating from known clinical and diagnostic criteria for MHS that ultimately may limit the reproducibility and application of this research.

Concerns around construct validity

In our dataset, there was limited explication on the theoretical/clinical grounding of the MHS of interest, beginning with clearly defining what mental health concern is being measured, and how it is operationalized within the research.

Specifically, many papers did not leverage established theories in clinical science or clinical psychology to establish or ground the status they investigated or specifically defined the construct itself. For example, five studies examine the concept of anxiety[36,37,46,48,54,59,64,85], though none operationalize what they mean when they study this particular disorder. Anxiety as a concept is overloaded—it is a category of nervous disorders, symptomatology that can influence other mental disorders, a transient emotion that people experience, and lay usage referring to emotional states and/or traits of a person. We see similar patterns for the notion of depression—it is frequently and subtly implied that the authors are referring to major depressive disorder; yet, these definitions are rarely explicated.

More ambiguities arise when documents establish positive and negative sources of data for identifying examples to pass to a predictive system. In our Results, we identified numerous innovations in techniques for positively identifying MHS—from hashtag use, e.g. #depression), follower networks, and digital administration of screening questionnaires like CES-D to consenting participants. However, in the documents, we rarely see reflection or evaluation of whether the new technique may measure the construct of interest. For example, the use of hashtags is a unique way to identify discussions of depression, but does it accurately identify those who suffer from major depressive disorder or is it another group of people interested in the topic? For less precise measurements, such as mood or stress, hashtags may be a valuable signal, but their application to diagnostic-level criteria is as of yet untested. Similar ambiguities on evaluating negative or "control" datasets also appear, as few studies establish that the research team was able to identify a lack of MHS in their populations. Even in the case of clinically-grounded approaches such as screening questionnaires, the papers do not establish the strength of the relationships between screening for MHS and the variables of interest.

These unstable constructs permeate through the experimental design, data collection and designing and selecting models. Rarely is reflection or justification provided that explain the selection and reduction of variables/features, data bias corrections, or algorithm selection. We see this gap manifest in what is reported for validation of predictive algorithms—only 32 of 75 papers reported explicitly five minimum standards for reproducing these algorithms. Additionally, we saw very limited use of causal analysis approaches or techniques to establish stronger relationships between the variables on social media and the MHS of interest, such as controlling for confounding factors or adjusting for sampling biases.

These challenges with construct validity jeopardize the credibility of identifying MHS and the replication of these studies in the future. As Ernala et al. also found in their explorations of schizophrenia prediction on social media[27], the operationalization of identifying MHS is not connected to theoretically or clinically rigorous definitions of mental health, nor is the new method of identification formally or causally validated. Without construct validity being established, it is hard to know if the studies in our corpus indeed measure MHS in ways that may be useful for other audiences, such as clinicians, or if they are in fact measuring something else. Ernala et al. also showed that it is possible that we are measuring a complementary population of those interested in mental illness, of which a subset will likely have diagnoses[27]. However, if the implications of the work are being framed for clinical audiences and adoption, there must be stronger validation of the constructs in the research to be applied to clinical practices.

For replication, imprecise reporting of study details, such as variable selection criteria, can cause inappropriate or erroneous conclusions to be drawn from the results. For those unfamiliar with machine learning but are interested in the potential of these approaches, these gaps in reporting standards can imply that undisclosed researcher discretion guided the decision-making process, when, in fact, there are guided ways to approach problem solving in machine learning and artificial intelligence.

These gaps and unstable constructs may limit clinical and public health adoption of social media predictions of MHS. Many papers in the corpus indicate in their Introductions the potential for social media to augment clinical intake or assessment, the active management of mental disorder, guiding interventions, or accessing hard-to-reach populations[16]. However, with unstable construct validity and unclear methods documentation, the techniques in these papers may not be adopted for these purposes, as clinicians may not believe the measures are reliable for their patient populations. This may limit their adoption into real-world treatment protocols and designs.

Moving toward better practices in research

In light of these findings, we are hopeful that researchers can adopt practices that would facilitate better validity of their measures and correspondingly influence downstream adoption into clinical practice. There have been calls by researchers from within social media and health research to consider these factors[16,19,27], as well as broader calls around operationalizing constructs and abstraction in machine learning[97]. Workshops and symposia across disciplinary boundaries are emerging, designed to support more collaborative rigorous practices within this new area

Several studies within our corpus had strong construct validity that may serve as models in the dataset for best practices. Construct validity necessitates connection to clinically or theoretically-grounded practices—so grounding how MHS in these areas is operationalized is very important. This could be done in several ways. First, researchers could draw on relevant literature from domains like clinical psychiatry and psychology to anchor their approach, as De Choudhury et al. clearly defined the clinical research on major depressive disorder, then assessed it via administering screeners (like CES-D) to participants[1]. Similarly, Eichstaedt et al. used ICD-10 codes for diagnosis to establish the

Table 1. Our recommendations for standards for reporting for methods and study design.

Proposed standards for study design and methods reporting	
Ground truth validation procedures for all data	Explicit number of features/variables
Data source (API, scraping, etc.)	Variable/feature reduction techniques
Bias mitigation and sampling strategies	Algorithm used in best-performing scenario
Number of data points/samples	Hyperparameter tuning procedures
Source of all features/variables	Validation metrics
Error analysis and explanation	Explicit performance evaluation measures

presence of MHS, then asked participants consent to examine their Facebook data for signs of depression[56]. We also advocate for collaborations with domain experts to guide the operationalization process for MHS; domain insights and guidance would be brought into the explication of the clinical terms to the social media context. In another paper, Burnap et al. partner with an expert on suicidality to build a classifier that distinguishes between six kinds of Tweets about suicide, ranging from those indicating legitimate disclosures of suicidality to awareness campaigns[5].

We encourage this new area of research to be mindful of reporting practices within papers to facilitate better replicability and scholarship. These issues may be caused in part because of the interdisciplinarity of the area and lower awareness around the adoption of predictive models in research domains without background in machine learning or statistical modeling[26]. We believe that the concerning reporting practices across the corpus can easily be rectified with better reporting standards for data collection, annotation of data, and statistical modeling. In that vein, in Table 1, we propose several reporting standards that could be adopted by the area to provide clarity. These extend beyond our minimum reporting requirements, and include opportunities for better reporting of positive and negative signs of MHS, data bias and sampling strategies, and feature selection. We also believe that better reporting standards will avoid potential traps in erroneous conclusions being drawn without sufficient evidence or risky causal language being used, strengthening the quality of the research from this emergent area. This list is not intended to be an all-encompassing proposal for the field; in fact, the field should work to establish practices and guidelines for effective use of machine learning and predictive techniques in this domain area beyond these ideas.

We also advocate for the establishment of practices and norms by this nascent field of research through stronger connections to the traditions of clinical psychiatry. Domain experts like clinical psychiatrists, researchers in medicine, social workers with experience in mental illness, and other experts have valuable knowledge to direct this research to be more rigorous and accurately assess the constructs we claim to measure. As the field moves towards generalizing these findings to new social media platform or new opportunities for practice, it is essential that psychometric, especially construct validity is carefully maintained throughout these practices. Looking towards complementary fields like mobile health[98,99], bioinformatics[100], these areas have prioritized critical inquiry and reflection into their practices and have brought in clinical collaborators on their projects. This may also mean drawing on the methods of other areas to establish better validity, such as experiments, controlled study designs, and randomized control trials. By working with domain experts and adopting practices from this space, the research will improve as it is better able to "measure what we think [the concepts] measure"[92][p. 5].

In conclusion, we offered a critical analysis of the methods, study design, and results reporting in 75 papers that predict mental health status on social media data. Our review identified key areas of similarity and trends within the field around data annotation and bias, pre-processing and feature selection, and model selection and validation measures. We also uncovered gaps in reporting procedures for minimum standards for methods validation, and gaps in precision in identifying the mental health status of interest. We hope that this meta-review provides the field guidance on the methods of interest in this space and guides researchers towards better reporting standards to encourage more reproducible and replicable science in this important area.

METHOD

Constructing a literature review corpus across disciplinary boundaries is challenging because of the methods of publication. Unlike other fields which rely on journals, the most common venues for publication in CS are conference proceedings. When we tested our initial search strategy through standard indexing services, journal entries were robustly indexed; yet there were large gaps in conferences known to be important in these subfields across professional organizations (e.g. AAAI, ACL, ACM, NIPS/NeurIPS, AMIA). Initial experiments with keyword searches through engines like Google Scholar yielded over 200,000 candidate papers, which is intractable for searching.

To manage these challenges, our search consisted of 41 hand-selected venues (both conferences and journals) that "seeded" our search. Then, we used search terms to filter for candidate papers in these venues. Finally, we sampled the references of candidates once to identify any missing research. We found 75 papers in total —more extensive details of our process are included in the Supplementary Information.

Search strategy

Two sets of keywords were developed to search in pair-wise fashion: those for mental health and those for social media. For mental health, 16 terms were identified, related to generic terms for mental health and disorders, the most common mood and psychosocial disorders, and symptomatology (e.g. stress, psychosis). This was informed by prior work[20,21] and the DSM-V[101]. For social media, we searched for eight terms, including general terms for social media as well as three popular social networks, Facebook, Twitter, and Instagram. A list of our keywords can be found in Table 2.

To overcome the challenges mentioned above about indexing, 41 English venues were identified that could publish research on predicting MHS using social media data. This included a large set of CS conference venues across many sub-areas, general interest journals, and proceedings in health informatics and data science. A full list of these venues can be found in the Supplementary Information, Table 3.

We used three different search engines to ensure robust coverage across these venues, given our above indexing concerns. We used the Association of Computing Machinery (ACM) Digital Library for ACM journals and conferences, Google Scholar using the Publish or Perish software[93] for other conference publications, and Web of Science for journals. One venue (CLPsych) was not

Table 2. Keywords for literature search.

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

indexed correctly by any search engine, so we manually searched the proceedings for matching keywords in the title and abstract. Using these strategies, we identified 4420 manuscripts that matched our keyword pairs.

Filtering strategy

The manuscripts were filtered to only include peer-reviewed, original, and archival studies published between 2008 and 2017, dovetailing with the emergence of academic research on social media[102]. Certain kinds of publications were excluded, as they did not conform to our standards for originality: meta and literature reviews, commentaries and opinions, case studies, shared tasks, and non-archived submissions to CS conferences. After deduplication and filtering, this resulted in 2344 manuscripts.

Next, we manually filtered by title and abstract, removing items obviously not relevant to mental health or social media. Examples of mismatches included other health conditions, such as cancer, and data sources like electronic health records. This screening of titles/abstracts resulted in 87 papers.

Finally, all 87 papers were read and fully screened with the following criteria for MHS:

1. They must address mental health in clinically specific ways. This meant studying a mood or psychosocial disorder (e.g. depression), given symptoms from the DSM-V[101] about disorders (e.g. suicide), or the generalized severity of mental disorders (e.g. moderate vs. severe depression). We excluded papers about subjective mood, well-being, happiness, or general emotions not directly related to mental disorder diagnosis (e.g. angry or happy). We also excluded papers about mental disorders and conditions that are not mood or psychosocially oriented (e.g. ADHD, autism spectrum disorder)[101].
2. The paper's method must focus on quantitative prediction. This included regression analysis, machine learning, and time series analysis.
3. The paper must study social media data, which we define as websites or apps that allow users to post/maintain content and profiles and interact/develop social networks and communities with others around said content[12,92,102]. Current examples would be Facebook, Twitter, Reddit, and Tumblr. We excluded other digital data traces, such as search engines, SMS/texting datasets, and fitness or mood trackers—these areas represent important areas for exploration but were out of scope for our study.
4. The prediction must be made on an individual. If a paper made predictions on individuals that were then aggregated for another purpose, we included these in our analysis.

This process generated 44 papers for analysis. Finally, to comprehensively expand our dataset beyond our 41 venues, we conducted a snowball sampling of related papers to extend the corpus of these 44 papers, identified from the bibliographic details from the citations, detailed in the Supplementary Information. This process identified 11 new papers, in turn providing 55 papers for inclusion in the review. In September 2019, we updated the dataset to search for 2018 data. This process and snowball sample identified 20 new papers, bringing the total number of papers in our corpus to 75. A full list of the documents, and details of our data collection process, are included in the Supplementary Information, Table 1.

Analysis technique

We developed a priori a rubric for analyzing the manuscripts that included both descriptive, quantitative, and qualitative criteria, influenced by prior work[20,21,92,103] and our understandings of the research space. This rubric had over 100 items, including data collection methods and pre-processing strategies, accuracy and baseline thresholds, results reporting mechanisms, and the presence of commentary on certain study design choices and implications of the research. We also recorded qualitative notes for analytical insights and thematic observations. To test the robustness of our rubric, we randomly selected four manuscripts of our corpus to annotate before beginning. We adjusted the rubric for additional reporting categories based on the results of our trial annotation. The relevant portions of our rubric design can be found in the Supplementary Information, Table 2.

We then conducted a close reading of all 75 papers in our corpus, annotating the rubric and identifying corpus-wide trends. The entire dataset was read and coded twice by the first author to standardize the coding process, each time in a random order. We then met and discussed the emergent themes and findings, which constitute our analysis.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

DATA AVAILABILITY

All data generated and papers analysed during this study are included in this published article (and its Supplementary Information files).

Received: 22 July 2019; Accepted: 17 January 2020;
Published online: 24 March 2020

REFERENCES

1. De Choudhury, M., Gamon, M., Counts, S. & Horvitz, E. Predicting depression via social media. *ICWSM* **2**, 128–137 (AAAI, 2013).
2. Tsugawa, S. et al. Recognizing depression from twitter activity. In *Proc. ACM Conference on Human Factors in Computing Systems (CHI)*. 3187–3196 (ACM, 2015).
3. Reece, A. G. & Danforth, C. M. Instagram photos reveal predictive markers of depression. *EPJ Data Science* **6**, 1–34 (2017).
4. Coppersmith, G., Ngo, K., Leary, R. & Wood, A. Exploratory analysis of social media prior to a suicide attempt. In *Proc. 3rd Workshop on Computational Linguistics and Clinical Psychology*. 106–117 (ACL, 2016).
5. Burnap, P., Colombo, W. & Scourfield, J. Machine Classification and analysis of suicide-related communication on Twitter. In *Proc. ACM Conf. of HyperText (HT)*. 75–84. (ACM, 2015).
6. Shing, H.-C. et al. Expert, crowdsourced, and machine assessment of suicide risk via online postings. In *Proc. 5th Workshop on Computational Linguistics and Clinical Psychology* 25–36 (ACL, 2018).
7. Chancellor, S., Mitra, T. & De Choudhury, M. Recovery Amid Pro-Anorexia: Analysis of Recovery in Social Media. In *Proc. ACM Conference on Human Factors in Computing Systems (CHI)* 2111–2123 (ACM, 2016).
8. Wang, T., Brede, M., Ianni, A. & Mentzakis, E. Detecting and Characterizing Eating-Disorder Communities on Social Media. In *Proc. 10th ACM International Conference on Web Search and Data Mining (WSDM)* 91–100 (ACM, 2017). https://doi.org/10.1145/3018661.3018706.
9. Mitchell, M., Hollingshead, K. & Coppersmith, G. Quantifying the language of schizophrenia in social media. In *Proc. 2nd Workshop on Computational*

Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality 11–20 (ACL, 2015).
10. Lin, H. et al. Detecting stress based on social interactions in social networks. *IEEE Transac. Knowl. Data Eng.* **29**, 1820–1833 (IEEE, 2017).
11. Chancellor, S., Lin, Z. J. J., Goodman, E. L., Zerwas, S. & De Choudhury, M. Quantifying and Predicting Mental Illness Severity in Online Pro-Eating Disorder Communities. In *Proc. 19th ACM Conference of Computer Supported Cooperative Work (CSCW)* 1169–1182 (ACM, 2016). http://dl.acm.org/citation.cfm?doid=2818048.2819973.
12. Ellison, N.B. & Boyd, Danah M. Sociality through social network sites. In *The Oxford handbook of internet studies* (Oxford University Press, 2013).
13. Vincent, J. Facebook is using AI to spot users with suicidal thoughts and send them help. https://www.theverge.com/2017/11/28/16709224/facebook-suicidal-thoughts-ai-help. (2017).
14. Muriello, D., Donahue, L., Ben-David, D., Ozertem, U. & Shilon, R. Under the hood: Suicide prevention tools powered by AI. https://code.fb.com/ml-applications/under-the-hood-suicide-prevention-tools-powered-by-ai/. (2018).
15. Torous, J., Keshavan, M. & Gutheil, T. Promise and perils of digital psychiatry. *Asian journal of psychiatry* **10**, 120–122 (2014).
16. Chancellor, S., Birnbaum, M., Caine, E., Silenzio, V. & De Choudhury, M. A taxonomy of ethical tensions in inferring mental health states from social media. in *Proc. Conference on Fairness, Accountability, and Transparency (FAT*)* (ACM, 2019).
17. Conway, M. & O'Connor, D. Social media, big data, and mental health: current advances and ethical implications. *Curr. Opin. Psycholog.* **9**, 77–82 (2016).
18. Paul, M. J. & Dredze, M. Social monitoring for public health. *Synthesis Lectures on Information Concepts, Retrieval, and Services* Vol. 9, 1–183 (Morgan & Claypool Publishers, 2017).
19. Benton, A., Coppersmith, G. & Dredze, M. Ethical research protocols for social media health research. In *Proc. of the First ACL Workshop on Ethics in Natural Language Processing*, 94–102 (ACL, 2017).
20. Seabrook, E. M., Hons, B., Kern, M. L. & Rickard, N. S. Social networking sites, depression, and anxiety: a systematic review. *JMIR Ment. Health* **3**, e50 (2016).
21. Wongkoblap, A., Vadillo, M. A. & Curcin, V. Researching mental health disorders in the era of social media: Systematic review. *J. Med. Internet Res.* **19**, e228 (2017).
22. Guntuku, S. C., Yaden, D. B., Kern, M. L., Ungar, L. H. & Eichstaedt, J. C. Detecting depression and mental illness on social media: an integrative review. *Curr. Opin. Behav. Sci.* **18**, 43–49 (2017).
23. Calvo, R., Milne, D., Hussain, M. S. & Christensen, H. Natural language processing in mental health applications using non-clinical texts. *Nat. Lang. Eng.* 1–37 (2017).
24. Hicks, J. L. et al. Best practices for analyzing large-scale health data from wearables and smartphone apps. *NPJ Digit. Med.* **2**, 45 (2019).
25. Sanches, P. et al. HCI and Affective Health: Taking stock of a decade of studies and charting future research directions. In *Proc. ACM Conference on Human Factors in Computing Systems (CHI)* (ACM, 2019).
26. Chancellor, S., Baumer, E. P.S. & De Choudhury, M. Who is the "human" in human-centeredmachine learning: the case of predicting mental health from social media. *Proc. ACM Hum.-Comput. Interact.* **3**, 147–1 (ACM, 2019).
27. Ernala, S. K. et al. Methodological Gaps in Predicting Mental Health States from Social Media: Triangulating Diagnostic Signals. In *Proc. ACM Conference on Human Factors in Computing Systems (CHI)* (ACM, 2019).
28. Shen, Y.-c., Kuo, T.-t., Yeh, I.-n., Chen, T.-t. & Lin, S.-d. Exploiting Temporal Information in a Two-Stage Classification Framework for Content-Based Depression. In *Proc. 17th Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 276–288 (Springer-Verlag Berlin Heidelberg, 2013).
29. Masuda, N., Kurahashi, I. & Onari, H. Suicide Ideation of Individuals in Online Social Networks. *PLoS ONE* **8**, e62262 (2013).
30. De Choudhury, M., Counts, S. & Horvitz, E. Predicting postpartum changes in emotion and behavior via social media. In *Proc. ACM Conference on Human Factors in Computing Systems (CHI)* 3267–3276 (ACM, 2013).
31. Wang, X. et al. A depression detection model based on sentiment analysis in micro-blog social network. In *Proc. Pacific-Asia Conference on Knowledge Discovery and Data Mining* 201–213 (Springer-Verlag Berlin Heidelberg, 2013).
32. Tsugawa, S. et al. On estimating depressive tendency of twitter users from their tweet data. *IEEE Virtual Reality*. **2**, 29–32 (IEEE, 2013).
33. Park, S., Lee, S. W., Kwak, J., Cha, M. & Jeong, B. Activities on Facebook reveal the depressive state of users. *J. Med. Internet Res.* **15**, 1–15 (2013).
34. De Choudhury, M., Counts, S. & Horvitz, E. Social Media As a Measurement Tool of Depression in Populations. In *Proc. 5th Annual ACM Web Science Conference* (WebSci) 47–56 (ACM, 2013).
35. Vedula, N. & Parthasarathy, S. Emotional and Linguistic Cues of Depression from Social Media. in *Proc. 2017 International Conference on Digital Health* 127–136 (ACM, 2017). https://doi.org/10.1145/3079452.3079465.
36. Loveys, K., Crutchley, P., Wyatt, E. & Coppersmith, G. Small but Mighty: Affective Micropatterns for Quantifying Mental Health from Social Media Language. In *Proc. Fourth Workshop on Computational Linguistics and Clinical Psychology* 85–95 (ACL, 2017).
37. Benton, A., Mitchell, M. & Hovy, D. Multitask learning for mental health conditions with limited social media data. In *Proc. 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1*. http://www.aclweb.org/anthology/E17-1015. (ACL, 2017).
38. Zhou, Y., Zhan, J. & Luo, J. Predicting Multiple Risky Behaviors via Multimedia Content. In *Proc. International Conference on Social Informatics* (Springer International, 2017).
39. Cheng, Q., Li, T. M. H., Kwok, C.-L. L., Zhu, T. & Yip, P. S. F. Assessing suicide risk and emotional distress in Chinese social media: a text mining and machine learning study. *J. Med. Internet Res.* **19**, 1–10 (2017).
40. Saha, K., Chan, L., De Barbaro, K., Abowd, G.D. & De Choudhury, M. Inferring Mood Instability on Social Media by Leveraging Ecological Momentary Assessments. In *Proc. ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)* (ACM, 2017).
41. Saha, K. & De Choudhury, M. Modeling Stress with Social Media Around Incidents of Gun Violence on College Campuses. In *Proc. ACM on Human-Computer Interaction* 1–92 (ACM, 2017).
42. Jamil, Z., Inkpen, D., Buddhitha, P. & White, K. Monitoring Tweets for Depression to Detect At-risk Users. In *Proc. Fourth Workshop on Computational Linguistics and Clinical Psychology* 32–40 (ACL, 2017).
43. Birnbaum, M. L., Ernala, S. K., Rizvi, A. F., De Choudhury, M. & Kane, J. M. A collaborative approach to identifying social media markers of schizophrenia by employing machine learning and clinical appraisals. *J. Med. Internet Res.* **19**, e289 (2017).
44. Simms, T. et al. Detecting cognitive distortions through machine learning text analytics. In *Proc. 2017 IEEE International Conference on Healthcare Informatics (ICHI)*. http://ieeexplore.ieee.org/abstract/document/8031202/. (IEEE, 2017).
45. Reece, A. G. et al. Forecasting the onset and course of mental illness with Twitter data. *Sci. Rep.* **7**, 13006 (2017).
46. Shen, J. H. & Rudzicz, F. Detecting anxiety on Reddit. In *Proc. Fourth Workshop on Computational Linguistics and Clinical Psychology*. 58–65 (ACL, 2017).
47. Shen, G. et al. Depression detection via harvesting social media: A multimodal dictionary learning solution. In *Proc. Twenty-Sixth International Joint Conference on Artificial Intelligence* (IJCAI) (IJCAI, 2017).
48. Gkotsis, G. et al. Characterisation of mental health conditions in social media using Informed Deep Learning. *Sci. Rep.* **7**, 45141 (2017).
49. Cohan, A., Young, S., Yates, A. & Goharian, N. Triaging content severity in online mental health forums. *J. Assoc. Inform. Sci. Technol.* **68**, 2675–2689 (2017).
50. Shen, T. et al. Cross-domain depression detection via harvesting social media. In *Proc. Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI)* 1611–1617 (IJCAI, 2018).
51. Wongkoblap, A., Vadillo, M. A. & Curcin, V. A multilevel predictive model for detecting social network users with depression. In *Proc. 2018 IEEE International Conference on Healthcare Informatics (ICHI)* 130–135 (IEEE, 2018).
52. Yates, A., Cohan, A. & Goharian, N. Depression and self-harm risk assessment in online forums. In *Proc. 2017 Conference on Empirical Methods in Natural Language Processing* 2968–2978 (ACL, 2017).
53. Seabrook, E. M., Kern, M. L., Fulcher, B. D. & Rickard, N. S. Predicting depression from language-based emotion dynamics: longitudinal analysis of facebook and twitter status updates. *J. Med. Internet Res.* **20**, e168 (2018).
54. Dutta, S., Ma, J. & De Choudhury, M. Measuring the impact of anxiety on online social interactions. In *Proc. International AAAI Conference on Web and Social Media* (ICWSM) (AAAI, 2018).
55. Pirina, I. & Çöltekin, Ç. Identifying depression on reddit: The effect of training data. In *Proc. 2018 EMNLP Workshop SMM4H: The 3rd Social Media Mining for Health Applications Workshop & Shared Task* 9–12 (Association for Computational Linguistics, 2018).
56. Eichstaedt, J. C. et al. Facebook language predicts depression in medical records. *Proc. Natl Acad. Sci. USA* **115**, 11203–11208 (2018).
57. Orabi, A. H., Buddhitha, P., Orabi, M.H. & Inkpen, D. Deep learning for depression detection of twitter users. In *Proc. Fifth Workshop on Computational Linguistics and Clinical Psychology* 88–97 (Association for Computational Linguistics, 2018).
58. Sadeque, F., Xu, D. & Bethard, S. Measuring the latency of depression detection in social media. In *Proc. Eleventh ACM International Conference on Web Search and Data Mining (WSDM)* 495–503 (ACM, 2018).
59. Ive, J., Gkotsis, G., Dutta, R., Stewart, R. & Velupillai, S. Hierarchical neural model with attention mechanisms for the classification of social media text related to mental health. In *Proc. Fifth Workshop on Computational Linguistics and Clinical Psychology* 69–77 (ACL, 2018).

60. Aladağ, A. E., Muderrisoglu, S., Akbas, N. B., Zahmacioglu, O. & Bingol, H. O. Detecting suicidal ideation on forums: proof-of-concept study. *J. Medical Internet Res.* **20**, e215 (2018).
61. Soldaini, L., Walsh, T., Cohan, A., Han, J. & Goharian, N. Helping or hurting? predicting changes in users' risk of self-harm through online community interactions. In *Proc. Fifth Workshop on Computational Linguistics and Clinical Psychology* 194–203 (Association for Computational Linguistics, 2018).
62. Ricard, B. J., Marsch, L. A., Crosier, B. & Hassanpour, S. Exploring the utility of community-generated social media content for detecting depression: an analytical study on instagram. *J. Med. Internet Res.* **20**, e11817 (2018).
63. Chancellor, S., Hu, A. & De Choudhury, M. Norms matter: contrasting social support around behavior change in online weight loss communities. In *Proc. 2018 CHI Conference on Human Factors in Computing Systems* (ACM, 2018).
64. Ireland, M. & Iserman, M. Within and between-person differences in language used across anxiety support and neutral reddit communities. In *Proc. Fifth Workshop on Computational Linguistics and Clinical Psychology* 182–193 (ACL, 2018).
65. Saravia, E., Chang, C. H., De Lorenzo, R. J. & Chen, Y. S. MIDAS: Mental illness detection and analysis via social media. In Proc. *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)* 1418–1421 (ACM, 2016).
66. Braithwaite, S. R., Giraud-Carrier, C., West, J., Barnes, M. D. & Hanson, C. L. Validating machine learning algorithms for twitter data against established measures of suicidality. *JMIR Ment. Health* **3**, e21 (2016).
67. Lin, H. et al. User-level psychological stress detection from social media using deep neural network. In *Proc. 22nd ACM international conference on Multimedia* 507–516 (ACM, 2014).
68. Huang, X. et al. Detecting Suicidal Ideation in Chinese Microblogs with Psychological Lexicons. In *Proc. 2014 IEEE 11th Intl Conf on Ubiquitous Intelligence & Computing and 2014 IEEE 11th Intl Conf on Autonomic & Trusted Computing and 2014 IEEE 14th Intl Conf on Scalable Computing and Communications and Its Associated Workshops (UIC-ATC-ScalCom)* Vol. 2014, 844–849 (IEEE, 2014).
69. Zhang, L. et al. Using Linguistic Features to Estimate Suicide Probability of Chinese Microblog Users. In *Proc. International Conference on Human Centered Computing* 549–559 (Springer, 2015).
70. Huang, X. et al. Topic Model for Identifying Suicidal Ideation in Chinese Microblog. In *Proc. Pacific Asia Conference on Language, Information and Computation* 553–562. http://www.aclweb.org/anthology/Y15-1064. (ACL, 2015).
71. Lin, H., Jia, J., Nie, L., Shen, G. & Chua, T.-S. What Does Social Media Say about Your Stress?. In *Proc.Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-16)* http://www.ijcai.org/Proceedings/16/Papers/531.pdf. (IJCAI, 2016).
72. Guan, L., Hao, B., Cheng, Q., Yip, P. S. F. & Zhu, T. Identifying Chinese microblog users with high suicide probability using internet-based profile and linguistic features: classification model. *JMIR Ment. Health* **2**, e17 (2015).
73. Zhao, L., Jia, J. & Feng, L. Teenagers' stress detection based on time-sensitive micro-blog com- ment/response actions. In *Proc. IFIP International Conference on Artificial Intelligence in Theory and Practice* 26–36 (IFIP, 2015).
74. Wang, X., Zhang, C. & Sun, L. An improved model for depression detection in micro-blog social network. In *2013 IEEE 13th International Conference on Data Mining Workshops*, 80–87 (IEEE, 2013).
75. Lin, H. et al. Psychological stress detection from cross-media microblog data using deep sparse neural network. In *Proc. 2014 IEEE International Conference on Multimedia and Expo (ICME)* 1–6 (IEEE, 2014).
76. De Choudhury, M., Kiciman, E., Dredze, M., Coppersmith, G. & Kumar, M. Discovering shifts to suicidal ideation from mental health content in social media. *Proc. ACM Conference on Human Factors in Computing Systems (CHI)*. **2016**, 2098–2110 (ACM, 2016).
77. Schwartz, H. A. et al. Towards assessing changes in degree of depression through facebook. In *Proc. Workshop on Computational Linguistics and Clinical Psychology* 118–125 (Association for Computational Linguistics, 2014).
78. De Choudhury, M., Counts, S., Horvitz, E. J. & Hoff, A. Characterizing and Predicting Postpartum Depression from Shared Facebook Data. In *Proc. 17th ACM Conference on Computer supported cooperative work & social computing (CSCW)* 626–638 (ACM, 2014).
79. De Choudhury, M. Anorexia on Tumblr : A Characterization Study on Anorexia. In *Proc. 5th International Conference on Digital Health* 43–50 (ACM, 2015).
80. Nguyen, T., Phung, D., Dao, B., Venkatesh, S. & Berk, M. Affective and content analysis of online depression communities. *IEEE Trans. Affect. Comput.* **5**, 217–226 (IEEE, 2014).
81. Nakamura, T., Kubo, K., Usuda, Y. & Aramaki, E. Defining patients with depressive disorder by using textual information. In *Proc. 2014 AAAI Spring Symposium Series* (AAAI, 2014).
82. Prieto, V.M., Matos, S., Alvarez, M., Cacheda, F. & Oliveira, J.L. Twitter: a good place to detect health conditions. *PLoS One.* **9** https://doi.org/10.1371/journal.pone.0086191 (2014).
83. Coppersmith, G., Dredze, M. & Harman, C. Quantifying Mental Health Signals in Twitter. In *Proc. Workshop on Computational Linguistics and Clinical Psychology* Vol. 2014, 51–60 (Association for Computational Linguistics, 2014).
84. Preotiuc-Pietro, D. et al. The Role of Personality, Age and Gender in Tweeting about Mental Illnesses. In *Proc. 2nd Workshop on Computational Linguistics and Clinical Psychology* 21–30 (Association for Computational Linguistics, 2015).
85. Coppersmith, G., Dredze, M., Harman, C., Holli and Hollingshead, K. From ADHD to SAD: Analyzing the language of mental health on Twitter through self-reported diagnoses. In *Proc. 2nd Workshop on Computational Linguistics and Clinical Psychology* 1–10 (Association for Computational Linguistics, 2015).
86. Resnik, P. et al. Beyond LDA: exploring supervised topic modeling for depression-related language in Twitter. In *Proc. 2nd Workshop on Computational Linguistics and Clinical Psychology* Vol. 2014, 99–107 (Association for Computational Linguistics, 2015).
87. Homan, C. M. et al. Toward Macro-Insights for Suicide Prevention: Analyzing Fine-Grained Distress at Scale. In *Proc. Workshop on Computational Linguistics and Clinical Psychology* 107 (ACL, 2014).
88. O'Dea, B. et al. Detecting suicidality on twitter. *Internet Interv.* **2**, 183–188 (2015).
89. McManus, K., Mallory, E. K., Goldfeder, R. L., Haynes, W. A. & Tatum, J. D. Mining Twitter data to improve detection of schizophrenia. *AMIA* **2015**, 122–126 (2015).
90. Coppersmith, G., Harman, C. & Dredze, M.H. Measuring post traumatic stress disorder in Twitter. In *Proc. Eighth International AAAI Conference on Weblogs and Social Media* 579–582 (AAAI, 2014).
91. Wang, Y. et al. Understanding and Discovering Deliberate Self-harm Content in Social Media. In *Proc. WWW* 93–102 (WWW, 2017).
92. Olteanu, A., Castillo, C., Diaz, F. & Kiciman, E. Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data* **2**, 13 (2019).
93. Harzing, Anne-Wil. Publish or perish. Tarma Software Research Pty Limited, (1997).
94. Shadish, W. R., Cook, T. D. & Campbell, D. T. In *Experimental and quasi-experimental designs for generalized causal inference* (Houghton Mifflin Company, 2002).
95. Howison, J., Wiggins, A. & Crowston, K. Validity issues in the use of social network analysis with digital trace data. *J. Assoc. Inform. Syst.* **12**, 2 (2011).
96. Lazer, D. Issues of construct validity and reliability in massive, passive data collections. In *The City Papers: An Essay Collection from The Decent City Initiative* (2015).
97. Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S. & Vertesi, J. Fairness and abstraction in sociotechnical systems. In *Proc. Conference on Fairness, Accountability, and Transparency (FAT*)* 59–68 (ACM, 2019).
98. Stowell, E. et al. Designing and Evaluating mHealth Interventions for Vulnerable Populations. In *Proc. ACM Conference on Human Factors in Computing Systems (CHI)*1–17 (ACM, 2018).
99. Arora, S., Yttri, J. & Nilsen, W. Privacy and security in mobile health (mhealth) research. *Alcohol Res.: Curr. Rev.* **36**, 143 (2014).
100. Luo, W. et al. Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view. *J. medical Internet Res.* **18**, e323 (2016).
101. Association, A. P. et al. *Diagnostic and statistical manual of mental disorders (DSM-5®)* (American Psychiatric Pub, 2013).
102. Boyd, Danah & Ellison, N. B. Social network sites: definition, history, and scholarship. *J. Comput.-Mediated Commun.* **13**, 210–230 (2007).
103. Liberati, A. et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: Explanation and elaboration. *PLoS Med.* **6**, e1000100 (2009).

ACKNOWLEDGEMENTS

This work was done in part while the first author was employed at the Georgia Institute of Technology. This work is in part supported by an NIH Grant #R01GM112697.

AUTHOR CONTRIBUTIONS

S.C. gathered the documents, filtered the dataset, conducted the close readings, and conducted the analysis of the dataset. Both S.C. and M.D.C. conceptualized the project, discussed the findings to identify thematic trends, and reviewed the paper.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information is available for this paper at https://doi.org/10.1038/s41746-020-0233-7.

Correspondence and requests for materials should be addressed to S.C.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.