Review

AI for Analyzing Mental Health Disorders Among Social Media Users: Quarter-Century Narrative Review of Progress and Challenges

David Owen[1], MSc; Amy J Lynham[2], PhD; Sophie E Smart[2], PhD; Antonio F Pardiñas[2*], PhD; Jose Camacho Collados[1*], PhD

[1]School of Computer Science and Informatics, Cardiff University, Cardiff, United Kingdom

[2]Centre for Neuropsychiatric Genetics and Genomics, Division of Psychological Medicine and Clinical Neurosciences, School of Medicine, Cardiff University, Cardiff, United Kingdom

[*]these authors contributed equally

Corresponding Author:
David Owen, MSc
School of Computer Science and Informatics
Cardiff University
Abacws
Senghennydd Road
Cardiff, CF24 4AG
United Kingdom
Phone: 44 (0)29 2087 4812
Email: owendw1@cardiff.ac.uk

Abstract

Background: Mental health disorders are currently the main contributor to poor quality of life and years lived with disability. Symptoms common to many mental health disorders lead to impairments or changes in the use of language, which are observable in the routine use of social media. Detection of these linguistic cues has been explored throughout the last quarter century, but interest and methodological development have burgeoned following the COVID-19 pandemic. The next decade may see the development of reliable methods for predicting mental health status using social media data. This might have implications for clinical practice and public health policy, particularly in the context of early intervention in mental health care.

Objective: This study aims to examine the state of the art in methods for predicting mental health statuses of social media users. Our focus is the development of artificial intelligence–driven methods, particularly natural language processing, for analyzing large volumes of written text. This study details constraints affecting research in this area. These include the dearth of high-quality public datasets for methodological benchmarking and the need to adopt ethical and privacy frameworks acknowledging the stigma experienced by those with a mental illness.

Methods: A Google Scholar search yielded peer-reviewed articles dated between 1999 and 2024. We manually grouped the articles by 4 primary areas of interest: datasets on social media and mental health, methods for predicting mental health status, longitudinal analyses of mental health, and ethical aspects of the data and analysis of mental health. Selected articles from these groups formed our narrative review.

Results: Larger datasets with precise dates of participants' diagnoses are needed to support the development of methods for predicting mental health status, particularly in severe disorders such as schizophrenia. Inviting users to donate their social media data for research purposes could help overcome widespread ethical and privacy concerns. In any event, multimodal methods for predicting mental health status appear likely to provide advancements that may not be achievable using natural language processing alone.

Conclusions: Multimodal methods for predicting mental health status from voice, image, and video-based social media data need to be further developed before they may be considered for adoption in health care, medical support, or as consumer-facing products. Such methods are likely to garner greater public confidence in their efficacy than those that rely on text alone. To achieve this, more high-quality social media datasets need to be made available and privacy concerns regarding the use of these data must be formally addressed. A social media platform feature that invites users to share their data upon publication is a possible solution. Finally, a review of literature studying the effects of social media use on a user's depression and anxiety is merited.

(J Med Internet Res 2024;26:e59225) doi: 10.2196/59225

KEYWORDS

mental health; depression; anxiety; schizophrenia; social media; natural language processing; narrative review

Introduction

Background

The Global Burden of Disease study (1990-2019) reports that anxiety disorders, major depressive disorder, and schizophrenia are the main drivers of years lived with disability and disability-adjusted life years across all age groups worldwide [1]. These mental health conditions are a sizable burden on the global population and public health systems. To help alleviate these problems, early intervention is essential [2].

The experiences of those with mental health disorders are often recounted on social media [3]. More broadly, users of Facebook and Reddit express favorable and adverse life events through the medium of text [4,5], and pictorial expressions of sensitive topics such as illness or hardship are becoming increasingly common through image-focused platforms such as Instagram [6]. As a result, methods that harness social media data for prediction of the mental health status of users have burgeoned [7-9]. Research has also spiked following the COVID-19 pandemic [10] and has become a truly interdisciplinary pursuit involving not only computer scientists but also psychologists, psychiatrists, and neuroscientists [11]. The broad idea behind this field is that models underpinned by artificial intelligence (AI) can "predict" a person's "mental health status" (refer to the study by Chancellor et al [12] for a discussion on the meaning of these terms in this literature). A branch of AI that is most appropriate for these methods is natural language processing (NLP), which uses computational techniques to learn, understand, and produce human language content [13]. Text-based dialogue systems, for example, have become a mainstay of NLP research. Their use in assisting people with neurocognitive disorders or mental health conditions is a popular application area. An early system, ELIZA [14], dates back to 1966. It purported to perform the role of a psychotherapist in conversation with a patient and has influenced the design of modern conversational agents such as ChatGPT [15]. In 2024, the potential for adults with dementia to adopt ChatGPT as a memory aid has been explored; it may be able to provide reminders of names, dates, and events, thus easing anxiousness [16]. The mining of text data to help assess a person's mental state has also followed from pre–21st century work. The *Whissell Dictionary of Affect in Language* [17], compiled in 1989 and now available on the web [18], can be used to estimate the mood conveyed in a body of text. This has given rise to modern methods for predicting the mental health status of social media users. Indeed, the huge volume of human language content available on the web, for example in Facebook and Reddit postings, fits very well the technical constraints of NLP techniques and can be straightforwardly processed into model inputs.

Some of the earliest attempts at predicting the mental health statuses of members of web-based communities were done without AI, through manual review of postings and classic statistical analyses. For example, in November 1999, psychiatrists monitored the general psychiatry subforum of the Norwegian web-based forum Doktoronline [19,20]. They observed that users who wrote negatively about their mental health by expressing sadness or resignation typically received positive and constructive responses from other users. Subsequently, these users often sought social support in their local communities. This corroborated previous findings showing that web-based community participation can have positive, real-life consequences for individuals [21,22], a motivation for later attempts at developing automatic health care intervention methods. Haker et al [23] examined the writings of web-based forum users who self-disclosed diagnoses of schizophrenia. They too noted that users with schizophrenia benefited by receiving advice from other users about medications and approaching health care professionals, as well as by receiving empathy and support.

The advent of social media platforms such as Facebook provided further locations for discussion about mental health disorders. Moreno et al [24] recognized that instances of major depressive disorder (depression hereafter) can be challenging to identify, particularly in older adolescents. So, between 2009 and 2010 they sought Facebook profiles of freshman students whose status updates referenced depression symptoms. Such students were then contacted and those who were willing were clinically screened to determine a diagnosis of depression. Students displaying depression symptoms in their status updates were more than twice as likely to be at risk for depression. Furthermore, the status updates referencing depression symptoms were often found to be a means of gathering support or attention, yet the students showed reluctance in seeking help in person. Thus, it was recognized that Facebook depression disclosures could be harnessed to identify those who might have unmet needs for mental health care. This provided an explicit motivation for improving the methods for predicting this disorder early in its course.

This Study

Due to the large volume of literature that exists in this area, which swelled during the COVID-19 pandemic, a review is timely. In this study, we focused on methods that concern the detection of language features presented in the texts of user social media postings. The main aim of our review was to ascertain state-of-the-art methodologies for detecting linguistic features that can be attributed to mental illnesses. This includes cataloging datasets containing "ground truth" (gold standard) labels of mental health status [12], which are available to help fine-tune these methodologies. Ground truths may be obtained from electronic health records (EHRs), clinical questionnaires, or self-disclosure statements of a mental health diagnosis (eg, "I was diagnosed with depression"). We then examined how these methodologies integrate the temporal stochasticity of mental states as reflected by longitudinal studies. We also identified common technical and ethical constraints met in the

development of the reviewed studies. Finally, we will form recommendations for the future direction of AI-based research on mental health.

Methods

Overview

We used Google Scholar to seek peer-reviewed articles published between January 1999 and February 2024. This literature search engine was selected because it is considered the most comprehensive search engine in academia [25-27]. It offers particularly extensive coverage of computer science and informatics, which is the primary discipline of literature that forms this review, outperforming databases like Scopus [28]. Our search aimed to retrieve literature covering the 3 main mental health burdens reported by the Global Burden of Disease study [1]: depression, anxiety, and schizophrenia, which are all common mental disorders. The articles then underwent a manual selection exercise to assign each of them to 1 of the 4 different subject areas that cover important and distinct aspects around mental health research in social media: datasets on social media and mental health, methods for predicting mental health status, longitudinal analyses of mental health, and ethical aspects of the data and analysis of mental health. These subject areas, described in more detail in Textbox 1, underpin the aims of this review described in the Introduction section.

Textbox 1. The subject areas covered in this narrative review.

Datasets on social media and mental health

To develop methods for predicting mental health status or conducting longitudinal analyses, carefully constructed social media datasets are required. We identify publicly available datasets that support this work and the challenges met in constructing them.

Methods for predicting mental health status

Approaches may consider how to detect mental health disorders in social media users and measure attributes of those disorders, such as their severities. We examine the progress in this area against a backdrop of evolving natural language processing technologies.

Longitudinal analyses of mental health

One's mental health state is fluid. We review attempts to gauge mental health state changes at both an individual level and population level. The former may assist in directing personalized health care to people at risk, while the latter may help inform public health policy.

Ethical aspects of the data and analysis of mental health

Research activities in the domain of predicting mental health status inevitably involve the acquisition and processing of personal data. We study the concerns reported among the general population and how they may be ameliorated.

Literature Search and Selection Strategy

A detailed exposition of the literature search and selection strategy, which is informed by Ferrari [29], now follows.

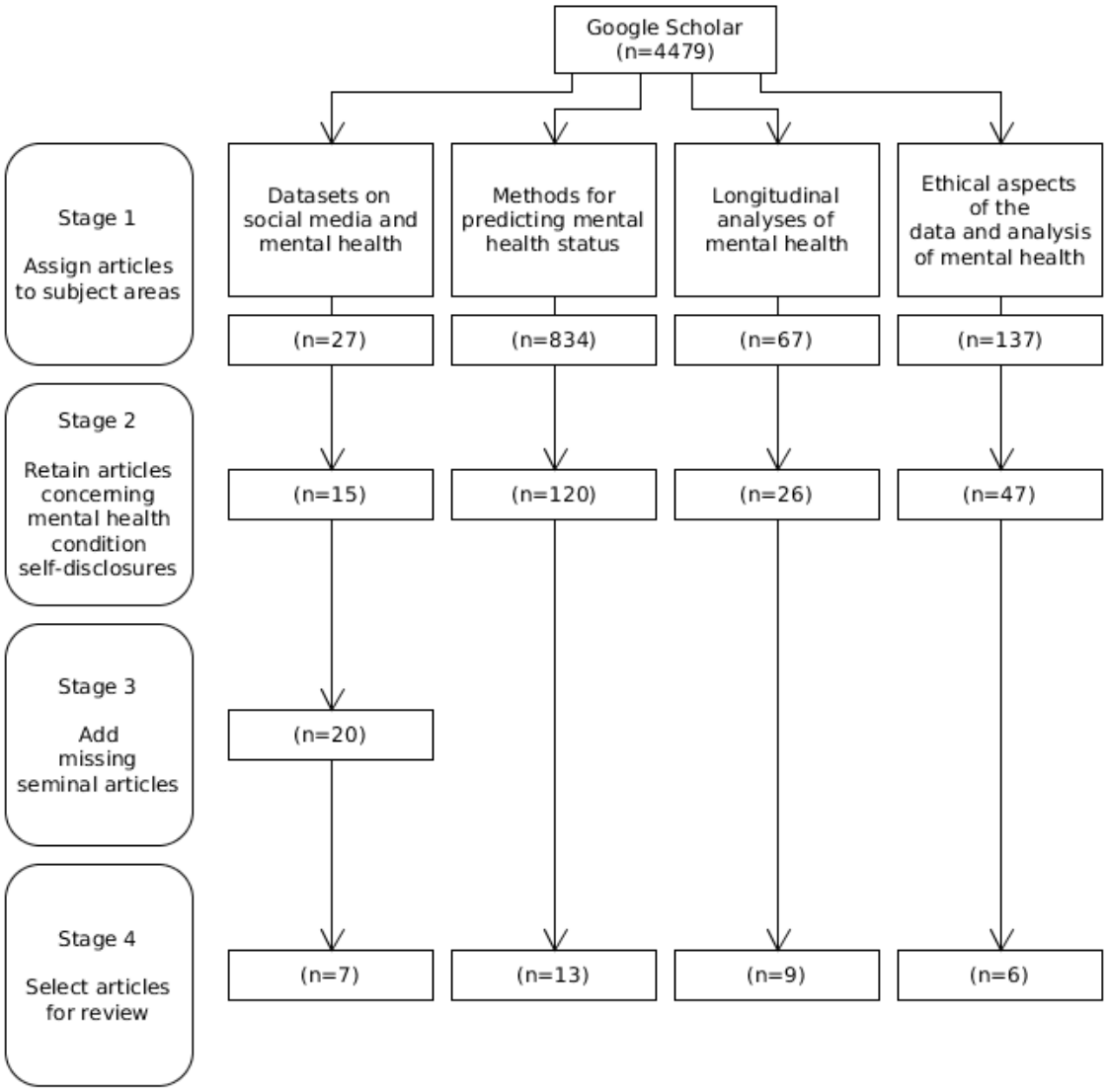
The search string in Textbox 2 was deployed to search article titles. 4479 articles were returned.

The articles then underwent a 4-stage manual sifting exercise as shown in Figure 1.

Textbox 2. Search string used in Google Scholar.

```
(intitle:"mental health"
OR intitle:"mental illness"
OR intitle:"mental disorder"
OR intitle:"psychiatric disorder"
OR intitle:"depression"
OR intitle:"anxiety"
OR intitle:"schizophrenia")
AND
(intitle:"social media"
OR intitle:"forum"
OR intitle:"forums"
OR intitle:"facebook"
OR intitle:"twitter"
OR intitle:"reddit")
```

Figure 1. The 4-stage sifting process for selecting articles for inclusion in the narrative review.



In the first stage, the title and abstract of each article was inspected so that it could be assigned to 1 of the 4 different subject areas shown in Textbox 1. Articles that did not relate to any of the 4 areas of interest were discarded. Duplicate results, preprints, presentation slides, posters, and non–English language articles were also discarded. A total of 1065 articles remained after this stage.

In the second stage, bodies of candidate articles for the 4 subject areas underwent inspection. Only articles where the participants of the studies self-disclosed a diagnosis of depression, anxiety, or schizophrenia were retained. The purpose was to ensure that only studies that used this ground truth were considered.

A third stage was performed that affected only the datasets on social media and mental health subject area. Due to our inherent knowledge of the subject area, we recognized that following the second stage, 5 seminal papers were absent. These articles were not retrieved in the Google Scholar search (Textbox 2). Their omission appeared due to their titles containing only implicit references to social media platforms. For example, the article “RSDD-Time: Temporal Annotation of Self-Reported Mental Health Diagnoses” does not explicitly mention that it concerns social media data. Following completion of the third stage, we arrived at 20 articles in this subject area.

A fourth and final stage involved selecting the articles for review in each subject area. The selection process identified the most pertinent articles over a broad timespan that concerned depression, anxiety, and schizophrenia.

A comprehensive listing of the articles considered at each stage of the exercise is provided in Multimedia Appendix 1.

Results

Overview

Following the 4-stage manual sifting exercise, 35 articles across the 4 subject areas were finally selected for review. The content of these articles covered research activity undertaken between 1999 and 2024 and influential events such as the COVID-19 pandemic. Table 1 describes the articles that were finally included for review. The format of this table is drawn from Szeto et al [30]. A narrative review of these articles is presented in the following 4 sections, which cover each of our 4 subject areas.

Table 1. Articles across the 4 subject areas that were selected for review.

Subject area and article title	Year published	Study population	Summary
Subject area: datasets on social media and mental health			
Predicting Depression via Social Media [31]	2013	Posts by 476 Twitter users who self-reported a diagnosis of depression between September 2011 and June 2012	Development of methods for dataset construction via crowdsourcing and quantifying users' depressive language use during the year before their diagnosis
Quantifying Mental Health Signals in Twitter [32]	2014	Posts published between 2008 and 2013 by 6696 Twitter users with a self-stated diagnosis of a mental health disorder: 394 with bipolar disorder, 441 with depression, 244 with PTSD[a], 159 with seasonal affective disorder, and 5728 controls	Development and evaluation of a method for swift and inexpensive capture of data about a range of mental illnesses
Depression and Self-Harm Risk Assessment in Online Forums [33]	2017	Posts published between January 2006 and October 2016 by 9210 Reddit users with a self-stated diagnosis of depression and 107,274 controls	Development and evaluation of a method for recognizing users with depression from their language use alone
RSDD-Time: Temporal Annotation of Self-Reported Mental Health Diagnoses [34]	2018	Self-reported depression diagnosis posts by 598 Reddit users published between June 2009 and October 2016	Development of methods for rule-based time extraction of depression diagnosis dates and mental health condition state classification
SMHD: A Large-Scale Resource for Exploring Online Language Usage for Multiple Mental Health Conditions [35]	2018	Posts published between January 2006 and December 2017 by 385,476 Reddit users with a self-stated diagnosis of a mental health disorder: 10,098 users with ADHD[b], 8783 users with anxiety, 2911 users with autism, 6434 users with bipolar disorder, 14,139 users with depression, 598 users with eating disorders, 2336 with OCD[c], 2894 with PTSD, 1331 with schizophrenia, and 335,952 controls	Development of methods for recognizing self-reported mental health condition diagnoses and obtaining high-quality labeled data automatically, rather than manually
Mental Health Surveillance over Social Media with Digital Cohorts [36]	2019	Randomly selected posts belonging to 48,000 US Twitter users	Development of methods for automatically inferring characteristics, including gender, ethnicity, and location of randomly collected Twitter users
Overview of eRisk at CLEF 2021: Early Risk Prediction on the Internet (Extended Overview) [37]	2021	Posts published between November 2009 and October 2020 by 80 Reddit users who completed a BDI-II[d] questionnaire	Development of methods for determining the severity of depression in Reddit users
Subject area: methods for predicting mental health status			
Social Media as a Measurement Tool of Depression in Populations [3]	2013	Posts by 117 Twitter users who indicated that they have clinical depression with onset between September 2011 and June 2012 and 157 controls	Development of methods for determining a social media depression index that may serve to gauge levels of depression in populations
Beyond LDA: Exploring Supervised Topic Modeling for Depression-Related Language in Twitter [38]	2015	Posts by approximately 2000 Twitter users, of whom approximately 600 self-identified as having been clinically diagnosed with depression	Investigation into the use of topic models to analyze linguistic signals for detecting depression
Quantifying the Language of Schizophrenia in Social Media [39]	2015	Posts published between 2008 and 2015 by 174 Twitter users who self-reported a diagnosis of schizophrenia	Development of methods for analyzing how the language of schizophrenia can aid in identifying and getting help to people with schizophrenia
Recognizing Depression from Twitter Activity [40]	2015	CES-D[e] questionnaire responses and posts by 209 Twitter users	Development of methods for extracting and using features from the activity histories of Twitter users to estimate the presence of depression

Subject area and article title	Year published	Study population	Summary
A Collaborative Approach to Identifying Social Media Markers of Schizophrenia by Employing Machine Learning and Clinical Appraisals [41]	2017	Posts published between 2012 and 2016 by 146 Twitter users who self-disclosed a diagnosis of schizophrenia and 146 controls	Development of methods for combining linguistic features of Twitter content with clinical appraisals to form a diagnostic tool for identifying individuals with schizophrenia
Detecting depression and mental illness on social media: an integrative review [42]	2017	43 peer-reviewed articles	Literature review of methods for predicting mental illness using social media
Forecasting the onset and course of mental illness with Twitter data [43]	2017	Posts by 105 Twitter users who had a diagnosis of depression and 99 controls. Also, posts by 174 Twitter users who had a diagnosis of PTSD	Development of models to predict the emergence of depression in Twitter users
A text classification framework for simple and effective early depression detection over social media streams [44]	2019	Posts by 135 Reddit users who have depression and 752 controls	Development of a text classification approach for early risk detection concerning social media users with depression, with an emphasis on explainable AI[f]
Towards Preemptive Detection of Depression and Anxiety in Twitter [45]	2020	Posts by 548 Twitter users who self-disclosed having either depression or anxiety and 4102 controls	Development of an LM[g]-based approach for early detection of depression in Twitter users
A Transformers Approach to Detect Depression in Social Media [46]	2021	Posts by 4000 Reddit users who self-disclosed having depression and 4000 controls	Development of transformer-based models for detecting depression in social media users
Characterisation of Mental Health Conditions in Social Media Using Deep Learning Techniques [47]	2022	77 peer-reviewed articles	Literature review of research concerning DL[h] techniques for identifying various mental health conditions from social media data
Utilizing ChatGPT Generated Data to Retrieve Depression Symptoms from Social Media [48]	2023	Posts by 3107 Reddit users	Development of methods for generating synthetic social media data for subsequent use in transformer-based language model depression detection
Prompt-based mental health screening from social media text [49]	2024	Posts by 1684 Twitter users who self-reported a diagnosis of depression and 11,788 controls	Development of methods that use LLM[i] prompting as an aid to mental health screening in social media text
Subject area: longitudinal analyses of mental health			
Feeling bad on Facebook: Depression disclosures by college students on a social networking site [24]	2011	Facebook profiles of 200 university students	Development of methods for determining associations between displayed depression symptoms on Facebook and other demographic or Facebook use characteristics
Towards Assessing Changes in Degree of Depression through Facebook [50]	2014	Status updates and survey responses of 28,749 Facebook users collected between June 2009 and March 2011	Development of a regression model to predict users' degrees of depression based on their Facebook status updates
Small but Mighty: Affective Micropatterns for Quantifying Mental Health from Social Media Language [51]	2017	Posts by 3680 Twitter users with a self-stated diagnosis of a mental health condition: 2271 with generalized anxiety disorder, 687 with eating disorders, 247 prone to panic attacks, 318 with schizophrenia, and 157 who have attempted suicide	An investigation of textual patterns in Tweet sequences occurring over short time windows to ascertain their suitability in quantifying psychological phenomena
Monitoring Online Discussions About Suicide Among Twitter Users With Schizophrenia: Exploratory Study [52]	2018	Posts by 203 Twitter users who self-identified as having schizophrenia and 173 controls	An exploration of the feasibility of monitoring web-based discussions about suicide among Twitter users who self-identify as having schizophrenia
Predicting Depression From Language-Based Emotion Dynamics: Longitudinal Analysis of Facebook and Twitter Status Updates [53]	2018	Status updates and depression severity ratings of 29 Facebook users and 49 Twitter users	A study of the associations between depression severity and emotion word expression on Facebook and Twitter status updates

Subject area and article title	Year published	Study population	Summary
What about Mood Swings: Identifying Depression on Twitter with Temporal Measures of Emotions [54]	2018	Posts by 585 Twitter users who self-reported a diagnosis of depression and 6596 controls	Development of a method for identifying users with or at risk of depression by incorporating measures of 8 emotions as features from Twitter posts over time, including a temporal analysis of these features
Monitoring Depression Trends on Twitter During the COVID-19 Pandemic: Observational Study [55]	2021	Posts of 2575 Twitter users who self-disclosed a diagnosis of depression and 2575 controls	Development of transformer-based DL language models to identify users with depression from their everyday language and to monitor the fluctuation of their depression levels
Using language in social media posts to study the network dynamics of depression longitudinally [56]	2022	Posts by 946 Twitter users who self-reported the dates of any depressive episodes in the past 12 months and the severity of their current depressive symptoms	An investigation into the association between depression severity and text features in Twitter posts
Enabling Early Health Care Intervention by Detecting Depression in Users of Web-Based Forums using Language Models: Longitudinal Analysis and Evaluation [57]	2023	Posts by 56 Reddit users who self-reported a diagnosis of depression and 168 controls	An investigation to determine the time points in the posting history of a person with depression, that are most indicative of their depression
Subject area: ethical aspects of the data and analysis of mental health			
Effectiveness of Social Media Interventions for People With Schizophrenia: A Systematic Review and Meta-Analysis [58]	2016	2 peer-reviewed publications	Literature review of the effectiveness of social media interventions for supporting people with schizophrenia
Ethical issues in using Twitter for population-level depression monitoring: a qualitative study [59]	2016	16 Twitter users with a self-reported diagnosis of depression who participated in a series of focus groups and 10 controls	Cross-sectional survey study of public attitudes toward using Twitter data for mental health monitoring
Social media, big data, and mental health: current advances and ethical implications [60]	2016	62 peer-reviewed articles	Literature review of work that uses social media "big data," NLP[j], and ML[k] for mental health surveillance and the ethical considerations therein
Who is the "Human" in Human-Centered Machine Learning: The Case of Predicting Mental Health from Social Media [61]	2019	55 peer-reviewed articles	Literature review of how scientific articles represent human research participants in human-centered ML
Ethics and Privacy in Social Media Research for Mental Health [62]	2020	35 peer-reviewed articles	Literature review of research that uses social media data in the context of mental health, with reference to the challenges in relation to consent, privacy, and use of such data
Understanding the Role of Social Media–Based Mental Health Support Among College Students: Survey and Semistructured Interviews [63]	2021	101 US university students aged 18 to 24	Web-based survey followed by semistructured interviews to investigate into whether and how social media platforms help meet university students' mental health needs in terms of the social support that they offer

[a]PTSD: posttraumatic stress disorder.

[b]ADHD: attention-deficit/hyperactivity disorder.

[c]OCD: obsessive-compulsive disorder.

[d]BDI-II: Beck Depression Inventory-II.

[e]CES-D: Center for Epidemiologic Studies Depression Scale.

[f]AI: artificial intelligence.

[g]LM: language model.

[h]DL: deep learning.

[i]LLM: large language model.

[j]NLP: natural language processing.

[k]ML: machine learning.

Datasets on Social Media and Mental Health

To develop methods for predicting mental health status, access to high-quality datasets is essential. De Choudhury et al [31] observed in 2013 that previous research had relied heavily on small, homogeneous samples of individuals who gave retrospective self-reports about their mental health, often via surveys. The authors also recognized that a person's posting activity on social media could provide time-stamped insights into their psychological state. To this end, they used crowdsourcing to compile a dataset of tweets belonging to 476 Twitter users who self-reported a diagnosis of depression. The data were subsequently used to analyze linguistic and behavioral patterns, such as symptom mentions and diurnal activity, respectively. While the data were deemed high quality by Coppersmith et al [32], they pointed to the limited size and scope of these data in terms of self-reported diagnoses, which needed to be obtained by manual completion of a questionnaire, namely the Center for Epidemiologic Studies Depression Scale screening test. Therefore, they proposed an automated method for labeled dataset construction, which sought self-reports of mental illness diagnoses on Twitter such as "I was diagnosed with depression." Their yield of >5000 different users conveying such statements between 2008 and 2013 indicated that a low-cost and low-resource method for data collection was possible. However, the authors acknowledged some limitations. First, only Twitter users were captured, a sample not likely representative of the general population but in this sense, similar to other social media datasets. Second, it was not possible to verify that the self-stated diagnoses were genuine or captured the same psychopathology as clinical diagnoses. For example, population biobank data have shown self-reported depression to be less heritable (ie, less of its variance in the population can be attributed to genetic factors) than diagnostically ascertained depression [64]. Nevertheless, this approach has ostensibly provided the foundation for several publicly available and widely used mental health datasets.

Yates et al [33] developed the Reddit Self-reported Depression Diagnosis (RSDD) dataset, which contains the posting histories of 9210 users with a diagnosis of depression revealed by self-report statements, like the ones described earlier. Further populated with 107,274 users without depression for control purposes, RSDD has become an often-used resource in the development of methods for predicting depression [65-70]. It has also propagated the development of 2 sister datasets, RSDD-Time [34] and Self-reported Mental Health Diagnoses (SMHD) [35]. The former was conceived by MacAvaney et al [34] after recognizing that research had largely not examined the temporality of mental health diagnoses. They randomly selected 598 posts from the RSDD dataset that contained the self-reported diagnosis statement of a user with depression and manually annotated them to denote when the diagnosis occurred. Owen et al [57] successfully exploited RSDD and RSDD-Time in a longitudinal study that evidenced a relationship between selected time spans before diagnosis and the sentiment a user exhibits in their postings. However, because many of the annotations in RSDD-Time denote that the diagnosis dates of many of the users cannot be estimated with a reasonable degree of accuracy (eg, the user merely stated that their depression diagnosis occurred "in the past"), the findings were predicated on the posting histories of only 56 users with depression. This highlights a need for much larger datasets where the dates of depression diagnoses are denoted with a high degree of accuracy.

SMHD, meanwhile, was born out of a desire for datasets covering a broad range of mental health disorders. It provided a platform for the development of methods concerning not only depression [71,72] but also suicidal ideation [73], schizophrenia [74], and even multiclass experimental setups involving combinations of anxiety, eating disorders, attention-deficit/hyperactivity disorder, bipolar disorder, and posttraumatic stress disorder [7,75-77]. It was also intended that a wider range of higher positive predictive value patterns be used to collect a greater volume of users who were diagnosed. Such patterns detect diagnosis keywords relevant to each disorder, drawn from the *Diagnostic and Statistical Manual of Mental Disorders* [78]. As a result, SMHD contains 20,406 users who were diagnosed and 335,952 matched controls. Despite these strengths, RSDD and SMHD are limited in their scope because they do not include posts made in mental health subreddits. It is recognized that language used in dedicated mental health subreddits systematically differs from the rest of Reddit [79]. This, and the limitation that they used only simple text patterns such as "I was diagnosed with depression" to collect users with mental health disorders, must be consistently considered in research work as it may introduce a bias to any models developed [80].

Other biases also exist in social media data. For example, most social media platforms, including Facebook, Twitter, and Instagram, have more male users [81]. There is also evidence to suggest that people with higher levels of education and household income are more frequent social media users [82]. To address such biases and improve the representativeness of social media datasets, Amir et al [36] considered a cohort-based approach to dataset construction. That is, they developed a demographic inference pipeline, which sought Twitter users and identified their age, gender, ethnicity, and location to create a subsample that was representative of the wider population. They then leveraged an existing model [83] to ascertain the prevalence of depression and posttraumatic stress disorder across the 48,000 users collected. This is in contrast to identifying users based on self-reported diagnosis statement patterns, which as mentioned, is another potential source of bias. The authors proposed that such use of surveillance-based methods could aid the identification of population-level trends in disorder prevalence. However, they also acknowledged that proper evaluation of these patterns would require disentangling the ways in which social media datasets differ from representative samples of the underlying population. In any case, further development and adoption of surveillance-based methods are constrained by privacy and ethical considerations. For example, it would surely require the permission of social media users before their data could be automatically sought and analyzed en masse, particularly in relation to personally identifiable information (age, gender, ethnicity, and health status). We explore these matters in more depth in the Ethical Aspects of the Data and Analysis of Mental Health section.

Finally in this section, we mention the work of the eRisk Lab [37], which touched upon another important dimension in the support of methods for predicting mental health status. Their 2021 dataset, which comprised Reddit posting histories belonging to 80 users, was accompanied by ground truth data that can aid in the development of methods for gauging the severity of depression. Recorded against each user was a completed Beck Depression Inventory-II (BDI-II) questionnaire, which categorizes the severity of their depression (ranging from minimal to severe). While the dataset proved useful in designing methods for finding associations between language features in the users' postings and their depression severities, the ground truth BDI-II questionnaires provided only the depression severities at the terminuses of the users' posting histories. Because the state of one's mental health is somewhat fluid [34], the dataset may contain users whose depression may have long passed. This is plausible given that one user in the dataset had a posting history spanning >10 years, although it should be noted that this is an anomaly, with the median posting history in the dataset being just >1 year. Furthermore, the dataset's small size in terms of number of users is a major constraint [84,85]. This highlights the difficulty in obtaining copious ground truth data that are traditionally collected via confidential questionnaires.

Table 2 summarizes some important features of the datasets discussed in this section, including the platform, contents, compilation year, acquisition inquiries information, and article title.

Table 2. Datasets discussed in this review that may be obtained from their authors.

Dataset	Platform	Contents	Year compiled	Acquisition inquiries	Article
RSDD[a]	Reddit	116,484 users: 9210 with depression and 107,274 controls	2014	RSDD dataset [86]	Depression and Self-Harm Risk Assessment in Online Forums [33]
RSDD-Time	Reddit	598 users with depression	2018	ir@Georgetown—resources [87]	RSDD-Time: Temporal Annotation of Self-Reported Mental Health Diagnoses [34]
SMHD[b]	Reddit	385,476 users: 10,098 with ADHD[c], 8783 with anxiety, 2911 with autism, 6434 with bipolar disorder, 14,139 with depression, 598 with eating disorders, 2336 with OCD[d], 2894 with PTSD[e], 1331 with schizophrenia, and 335,952 controls	2018	ir@Georgetown—resources—SMHD [88]	SMHD: A Large-Scale Resource for Exploring Online Language Usage for Multiple Mental Health Conditions [35]
2015 Computational Linguistics and Clinical Psychology Shared Task	Twitter	1746 users: 477 with depression, 396 with PTSD, and 873 controls	2015	CLPsych 2015 shared task evaluation [89]	Mental Health Surveillance over Social Media with Digital Cohorts [36]
eRisk 2021 Text Research Collection	Reddit	80 users who completed a BDI-II[f] questionnaire	2021	eRisk 2021 text research collection [90]	Overview of eRisk at CLEF 2021: Early Risk Prediction on the Internet (Extended Overview) [37]

[a]RSDD: Reddit Self-reported Depression Diagnosis.

[b]SMHD: self-reported mental health diagnoses.

[c]ADHD: attention-deficit/hyperactivity disorder.

[d]OCD: obsessive-compulsive disorder.

[e]PTSD: posttraumatic stress disorder.

[f]BDI-II: Beck Depression Inventory-II.

Methods for Predicting Mental Health Status

Background

The methods covered in this review are supported by machine learning (ML). As there is a broad terminology concerning ML, we introduce the relevant terms in Table 3.

Table 3. Machine learning (ML) terms used in this review.

Term	Description
Data representation	
LDA[a] [91]	A technique that can examine a group of documents and produce a series of words, known as a topic, that characterizes those documents. For example, "anatomy, dissection, genomes" may form the topic of a collection of biomedical documents.
LIWC[b] [92]	A text analysis technique that can infer the emotion conveyed in text (eg, positive or negative).
Ontology [93]	A graphical representation of knowledge that is both human-readable and machine-readable. For example, a biomedical ontology might show how different neurological signs and symptoms may be linked to relevant diseases.
Data augmentation [94]	The methods used to increase the size of a dataset by adding slightly modified copies of existing items in the dataset.
Algorithms	
Supervised learning [95]	A type of ML[c] algorithm analogous to human learning from past experiences to gain new knowledge to improve our ability to perform real-world tasks.
SVM[d] [96]	A supervised ML algorithm that learns by assigning labels to objects and can be used, for example, to recognize fraudulent credit card activity.
Random forest [97]	A supervised ML algorithm that combines the output of multiple decision trees to reach a single result.
DL[e] [98]	A type of ML algorithm (supervised or unsupervised) that can produce complex models from data without features (eg, LIWC) needing to be derived as input.
Pretrained models	
LM[f] [99]	An LM is a probability distribution over words or word sequences. LMs learn to predict text that might come before and after other text and thus are used in tasks such as predicting text when writing an email.
BERT[g] [100]	An LM that examines words within text by considering both left-to-right and right-to-left contexts.
ALBERT[h] [101]	A lightweight alternative to BERT that is suitable for use where less computing power is available.
MentalBERT [102]	An LM designed specifically to aid NLP[i] tasks in the mental health care research community.
MentalRoBERTa [102]	An alternative to MentalBERT that can perform predictions in longer left-to-right and right-to-left contexts.
LLM[j] [103]	Large-scale LM designed for NLP tasks such as producing complex text.
GPT [104]	A family of neural network (in that they mimic the workings of the human brain) models that support AI[k]-driven applications for creating content such as text, images, or sound.
ChatGPT [105]	A chatting robot that can provide a detailed response to a question or instruction.
Performance metrics	
Positive predictive value [98]	Of the instances in a dataset predicted by an ML algorithm to have a certain label, positive predictive value denotes how many of them indeed have that label. This is often referred to as precision in the ML literature.
Sensitivity [106]	Of the instances in a dataset with a particular label, sensitivity denotes how many of them were predicted correctly by an ML algorithm. Sensitivity is also known as recall.
F_1-score [107]	The harmonic mean of positive predictive value and sensitivity.
AUROC[l] [108]	Denotes an ML algorithm's performance in terms of distinguishing between labels.

[a]LDA: latent Dirichlet allocation.

[b]LIWC: Linguistic Inquiry and Word Count.

[c]ML: machine learning.

[d]SVM: support vector machine.

[e]DL: deep learning.

[f]LM: language model.

[g]BERT: Bidirectional Encoder Representations from Transformers.

[h]ALBERT: A Lite Bidirectional Encoder Representations from Transformers.

[i]NLP: natural language processing.

[j]LLM: large language model.

[k]AI: artificial intelligence.

[l]AUROC: area under the receiver operating characteristic.

Traditional ML Approaches

In 2013, methods for predicting mental health status from social media data began to emerge [12] and have often involved interdisciplinary teams of computer scientists and clinical psychologists. De Choudhury et al [3] were proponents of supervised learning methods for predicting depression among populations. Exploiting post-level and user-level features from a crowdsourced Twitter dataset, they developed the social media depression index. To do this they used a support vector machine (SVM). The social media depression index could be used to determine the degree of depression manifested by users in their daily tweets. In a US demographic population study, they observed that women were 1.5 times more likely to express signs of depression on social media than men, which marginally exceeded findings from epidemiological surveys on formal diagnoses that suggest the figure to be 1.3 [109]. The overestimation was linked to the greater emotional expressivity of women [110], suggesting that methods more sensitive to language use could help develop more robust models. Such methods include topic modeling via latent Dirichlet allocation (LDA). While this approach has also been used for predicting depression in Twitter users [32], its results have to be taken cautiously as its dataset, in terms of users with depression and control users, was not deemed a representative sample of the population [38]. Later work used LDA-derived features as input to an SVM classifier to discern between users with depression and control users on Twitter [40]. Although the effectiveness of the topic-driven approach was demonstrated to some extent, only a modest result of 35% sensitivity was achieved. In a similar experimental setup for the prediction of depression in Twitter users [43], another traditional ML algorithm, random forests, was deployed using Linguistic Inquiry and Word Count (LIWC) features derived from post text. A commendable area under the receiver operating characteristic score of 87% was achieved and the method was validated by the collection of the mental health histories of its 204 participants via the Center for Epidemiologic Studies Depression Scale questionnaire. Tsugawa et al [40] acknowledged that emerging deep learning algorithms could well advance the methods in this area and were likely to inform future work. We explore these algorithms in the next subsection, Language Models and Transformers. A contemporaneous review also concluded that advances in NLP and ML were making the prospect of large-scale screening of social media for at-risk individuals a near-future possibility [42]. It also cited 2 studies that were influential in dataset design methods [31,32] that we discussed in the Datasets on Social Media and Mental Health section as being likely to help realize this.

By 2019, interest in methods for early prediction of depression had developed due to the recognition that they could help people receive the health care and social support they need sooner than they otherwise might [44]. Burdisso et al [44] designed an algorithm named SS3 that would calculate the degree to which some given text belonged to a certain category. While it could be generalized to any domain, in this case, it was used to classify depressed and control users of the longhand forum Reddit. It demonstrated superior early risk classification performance across several different experimental settings when compared to baselines computed using more traditional algorithms such as SVM. It also demonstrated significantly faster computation times; approximately 20 times faster than SVM. A further aim of SS3 was to provide explainability [111] for its classification decisions. It could display pertinent excerpts of a user's Reddit text, such as "Fact is, I was feeling really depressed and wanting to kill myself," which may assist clinicians. This transparency could not be gleaned from traditional "black box" algorithms such as SVM. SS3 was also hailed as a low-resource method, because, unlike SVM, it does not necessarily need to process the entire input text before returning its classification decision. However, it was acknowledged that because it examines each word of the input text in a singleton fashion, it would not consider potentially crucial 2-word phrases such as "kill myself" in a classification decision.

Language Models and Transformers

The capabilities of language models (LMs) had become well understood in NLP by the start of the 2020s. So, further to the work conducted by Burdisso et al [44], Bidirectional Encoder Representations from Transformers (BERT) and A Lite BERT (ALBERT) were deployed in an early depression prediction task involving tweets that denoted whether a user was with depression or anxiety or with no disorder [45]. As BERT and ALBERT necessarily consider the context of each word they encounter in a classification task, the consideration of n-word phrases is inevitable, thus addressing a matter highlighted by Burdisso et al [44]. In an experimental setting where users with depression and control users were balanced, an F_1-score of 77% was achieved using BERT, compared to an SVM baseline of 65%. In an imbalanced dataset however, which is a more accurate representation of real-world scenarios where these tools could be applied, BERT achieved an F_1-score of 74% compared to SVM's score of 75%. Malviya et al [46] performed a similar experiment where individual posts in a Reddit dataset would be classified as depressed or nondepressed in nature by BERT and traditional baseline algorithms [46]. Once again, strong BERT performance was observed in a balanced experimental setting, therefore, strengthening evidence that further research is needed before LMs could be deployed for this prediction task in more realistic, imbalanced settings. Suggestions include generating synthetic instances to create balance [112] and resampling [113]. A review of deep learning approaches to mental health prediction [47] that postdates both studies [45,46] echoed the need for further work involving much larger datasets while acknowledging the impact of existing datasets that we have already highlighted [33,35].

Some of the most recent methods have harnessed generative AI, principally using GPT [104]. The arrival of generative AI has enhanced opportunities in this domain. We have already noted that the use of quality data is crucial in the pursuit of methods for predicting mental health status. Such data are often scarce and have given rise to data augmentation techniques [94,114]. A slightly different approach involves synthesizing data derived from existing data [115]. In an annual workshop task, a participating team used ChatGPT to synthesize data that would help develop models for identifying BDI-II-recognized depression symptoms conveyed in Reddit posts [48]. Several

thousand apparently suitable texts were generated. For example, to the BDI-II response "I am so sad or unhappy that I can't stand it," ChatGPT formed the text "I'm so overwhelmed by sadness that I can barely function anymore." However, it was found that models for linking such texts to appropriate BDI-II responses performed more strongly with respect to real data rather than their synthesized counterparts. It was suggested that the synthesized texts were overly detailed and complex, thus confounding LMs used in the subsequent classification exercise. One LM used was MentalRoBERTa [102], which is trained on real Reddit data. More judicious use of ChatGPT such that it produces less detailed texts that are more semantically similar to the BDI-II responses was proposed as follow-up work. A further use of a GPT has been in the automatic trisection [49] of the SetembroBR Twitter corpus of users with depression and control users [116]. The GPT was prompted to label each tweet as having either high, medium, or low relevance to mental health. The labeled dataset was then used as an input to a bag-of-words classifier and its prediction performance was compared with that of a BERT-derived baseline produced by an earlier study [117]. While this approach was markedly low resource and improved the baseline result by 5% in terms of sensitivity, it was acknowledged that improved prompting of the GPT, perhaps by using a more formal definition of depression, might see further improved sensitivity. Therefore, large LM (LLM) supported GPTs have shown potential for aiding mental health prediction in a variety of ways. For that potential to be fully realized, computer scientists need to consider how GPT prompting techniques can be optimized in each context.

Considerations for Schizophrenia

Finally in this section, we examine the literature's coverage of schizophrenia. In a 2015 study by Mitchell et al [39], LDA was applied in a Twitter dataset with the goal of distinguishing between users with schizophrenia and controls. Key findings were that irrealis mood (denoted by the use of uncertain terms such as "think" or "believe") [118] and flat affect (due to lack of emoticon use) [119] were prevalent in the posts of people with schizophrenia. A limitation of their dataset was that users' self-statements of schizophrenia diagnoses could not be verified, which is a problem in this field of research as psychotic symptoms might preclude people from believing in their diagnoses [120,121]. In any case, people with schizophrenia may be reluctant to disclose their diagnoses on social media because they are likely to receive stigmatized responses [122,123]. Birnbaum et al [41] attempted more accurate identification on Twitter using a human-machine partnered approach. Self-reported schizophrenia statements were scrutinized for their authenticity by a psychiatrist and a graduate-level mental health clinician. The ML-derived model subsequently developed was able to distinguish between users with schizophrenia and controls with 87% sensitivity. Despite this, the authors acknowledged that truly confirming the diagnosis of a user who makes a self-disclosure statement is not possible without access to the user's EHRs.

Longitudinal Analyses of Mental Health

Studies discussed so far have tended to predict a person's mental health at a particular point in time. However, a person's mental health state is not static [34]. Indeed, it has been argued that inferences derived from sample-level "snapshots" of mental health states might not lead to reliable predictions of the individual-level variation in these states through time [124]. Therefore, research has also examined temporal profiles of mental health disorders and symptoms. A 2011 study considered US college students' Facebook status updates and their potential for exhibiting content that may reveal symptoms of depression [24]. It was noted that opportunities for recognition and treatment of depression were being missed, particularly among college students [125,126]. Therefore, Facebook, a social media platform that had become well-established among the student population [127], presented innovative opportunities to identify college students at risk. A manual exercise saw the collection of Facebook status updates of 200 students that spanned 1 year. Human annotators then scrutinized each post, denoting a depressive symptom if deemed present according to the *Diagnostic and Statistical Manual of Mental Disorders* criteria [128]. A quarter of profiles exhibited at least 1 depressive symptom (as inferred through the use of terms like "hopeless" or "giving up"). This evidence that Facebook may allow the identification of at-risk students would be a precursor to future longitudinal analyses.

Schwartz et al [50] sought to gauge how the level of depression changes among Facebook users during a calendar year. Their method involved the extraction of 1-to-3-word terms, LDA-derived topics, and LIWC categories from the status updates of >28,000 users. A regression model was developed that indicated a significantly higher degree of depression among users during winter months than in summer months, which is compatible with observations made in the psychiatry literature [129]. A baseline model that considered only the average sentiment across each user's status updates was outperformed in terms of accuracy almost 3-fold, although the optimal model only exceeded 30% [130]. By comparison, Loveys et al [51] conducted experiments predicting mental health statuses during much shorter time spans, hours in fact. Tweets belonging to >2500 users who self-stated a diagnosis of either anxiety or schizophrenia were automatically labeled with either positive, neutral, or negative sentiment. For each user, the changes (or otherwise) in terms of sentiment across 3 subsequent tweets that occur within any 3-hour window were observed. These observations were dubbed "micropatterns." It was noted that users with schizophrenia were less likely to show emotional variability between tweets than control users, which perhaps demonstrates a deficit in affective expression, a known schizophrenia symptom [131]. Users with anxiety were less likely to make consecutive positive tweets than controls, again consistent with psychological findings [132]. However, the micropatterns did not contain sufficient details to indicate the severity of the mental health disorders but enriching the automatic labeling process by considering linguistic features other than sentiment (eg, terms that may be mapped to specific symptoms) may help in this respect.

Emotions and their changing nature over a series of web-based postings have also been studied. Seabrook et al [53] considered whether "emotion dynamics" in Twitter and Facebook may provide early indicators for depression risk. The feasibility of using emotion variability and instability as an indicator of depression severity, measured by the Patient Health Questionnaire-9 [133], was explored. It was hypothesized that self-reported depression severity would be positively associated with negative emotion word variability and instability across status updates. Status updates and depression severity ratings of 29 Facebook users and 49 Twitter users were collected. MoodPrism [134] would gauge the emotion of their status updates and the severity of depression (via Patient Health Questionnaire-9) over a 1-year period. Results suggested that instability in the negative emotion expressed on Facebook provides insight into the presence of depression symptoms for social media users. Also, greater variability of negative emotion expression on Twitter may, in fact, be protective for mental health. However, these observations were constrained by the users' tweets being unavailable for manual inspection due to privacy reasons. Therefore, no manual verification was possible, and the results are essentially unreproducible. Another study from 2018 also considered emotion expressions on Twitter for their use in predicting depression [54]. In total, 8 basic emotions (anger, disgust, fear, happiness, sadness, surprise, shame, and confusion) were sought in the tweets of 585 users with depression across a 4-month period. The average intensity of each emotion was calculated via the EMOTIVE ontology [135] and used in a time-series analysis of each user. This analysis in turn helped build ML-based classifiers for labeling previously unseen Twitter users as being either depressed or not. In the best-performing setup with a random forests classifier, 87% sensitivity using temporal features was achieved compared to 71% using simple LIWC features. This suggests that the changes in an individual's emotions over time show potential in identifying users with depression. Fine-grained consideration of the language used in tweets, such as tentative (eg, "maybe") and temporal-related terms, may not only predict its presence but also its severity [56].

The emergence of transformer-based LMs coincided with the onset of the COVID-19 pandemic. It was no coincidence that interest grew in methods for monitoring population-level depression on social media at that time and that LMs would feature. In one study, tweets dated between March 3 and May 22, 2020, were collected regarding users who self-disclosed having depression [55]. The goal was to develop a model for monitoring the fluctuation of depression levels of different groups as COVID-19 propagated. Using the BERT-like model XLNet [136] and a geographical aggregation of users in the dataset, they demonstrated how depression levels fluctuated between the aforementioned dates in New York, California, Florida, and the United States as a whole. It was observed that depression levels in all 4 geographical areas were similar during the pandemic, with a steady increase after the announcement of the United States National Emergency on March 13, a modest decrease after April 23, followed by a steep increase after May 10. The overall depression score of Florida was substantially lower than the United States average and the other 2 states, possibly because it has a lower depression level overall compared to the average US level irrespective of the pandemic. These findings were constrained by the fact that only Twitter users were considered, who therefore are not fully representative of the population. In a further use of LMs, Owen et al [57] aimed to determine how far in advance of a Reddit user's depression diagnosis their postings were most indicative of their condition. Overall, 56 users with depression and 168 controls were acquired from an intersection of the RSDD [33] and RSDD-Time datasets [34]. BERT and a specialist LM, MentalBERT [102], considered all user posts in increasingly large temporal bands up to 24 weeks (approximately 6 months) before the diagnosis dates of users with depression. The LMs achieved F_1-scores of 0.726 and 0.715, respectively, when 12 weeks of postings were considered, suggesting therefore that the most poignant language used by users with depression occurs in the final 3 months before their eventual diagnosis. The reason for the specialist LM performing less effectively than its general counterpart may be explained by the fact that the former is trained on text found in mental health subreddits, and such postings are not included in RSDD. Findings were tempered by the fact that the diagnosis dates were mere estimates, as explained in the discussion of RSDD-Time in the Datasets on Social Media and Mental Health section. In any case, it was posited that a multimodal classification approach might provide more robust results. For example, a Reddit user's upvotes or downvotes for posts may also be predictive of their mental health state.

We conclude this section by again exploring what the literature has covered in the realm of schizophrenia. Hswen et al [52] investigated the language used by Twitter users with schizophrenia to observe whether it would help assess suicide risk [52]. They examined the frequency of suicide-related tweets, paying particular attention to the times of such tweets. They hypothesized that Twitter users who self-identify as having schizophrenia would be significantly more likely to post tweets containing suicide terms when compared to Twitter users from the general population, thereby reflecting the elevated risk of suicide observed among individuals with schizophrenia in real-world settings. The tweets of 203 users with schizophrenia and 173 control users covering a 200-day period were collected. Only tweets that contained the words suicide or suicidal were targeted because, perhaps not surprisingly, the term suicide is frequently contained in suicide-related conversations [137,138]. Crucially, the time of day of each tweet was recorded. A logistic regression model predicted that the users with schizophrenia showed significantly greater odds of tweeting about suicide compared with control users (odds ratio 2.15, 95% CI 1.42-3.28). Considering the times of tweets, the frequency of conversations about suicide on Twitter correlated significantly with discussions about depression and anxiety, another trend that is consistent with established data [139,140]. However, similar to the studies discussed previously [39,41], the inability to be able to verify the diagnoses of the users with schizophrenia was cited as a main limitation.

Ethical Aspects of the Data and Analysis of Mental Health

When constructing datasets, developing methods, and performing longitudinal analyses to aid mental health prediction, people's privacy ought to be considered. In 2016, Mikal et al [59] sought to determine the attitudes of Twitter users toward the platform's use in population health monitoring. Their qualitative study focused on depression. A focus group was formed of Twitter users, some of whom had previously received a diagnosis of depression while others had not. The group was canvassed for their opinions on the prospect of machine-driven health monitoring and their privacy expectations thereon. Broadly speaking, participants were supportive of the use of publicly available data for health monitoring activities, provided that the user identities were concealed. The concerns about the reliability of methods that use crude keyword searches and the misleading findings they could yield were also noted. An incorrect labeling of depression for a user whose identity is revealed would be considered stigmatizing according to participants. The study was only indicative because the group comprised just 26 Twitter users of a narrow demographic (predominantly male with an average age of 26.9 years). However, a concurrent study by Conway and O'Connor [60] gleaned further evidence of fears regarding such stigmatization.

Nicholas et al [62] address similar privacy matters. They note that the introduction of the General Data Protection Regulation in Europe and popular scandals such as Cambridge Analytica's use of Facebook data brought data privacy into sharp focus. User concerns are many and varied. Some users fear that the research findings may affect credit card applications [141] and employment prospects and attract stigma [142]. Fears are compounded by evidence that deidentified data can be reidentified using materials published alongside research articles [143]. Indeed, the desire for anonymity appears particularly widely held, which echoes the findings by Mikal et al [59] and is reinforced by Vornholt and De Choudhury [63]. Therefore, obtaining explicit user consent for the use of their data is considered crucial. A possible route is via acceptance of social media platform terms and conditions. However, as these may not be read and understood [144], this may not constitute informed consent. One solution is to explicitly invite users to donate their social media data for research purposes [145]. Another proposal is a feature that enables users to opt in or out of their data being used as they post it [146].

A matter has also been identified regarding the terminology used in this area of mental health research. Chancellor et al [61] reviewed how human participants are referred to in literature for predicting mental health status using social media data. Common traits were seen across 55 articles. For example, introductions often refer to human participants as "individuals" and "people," but technical sections then refer to them as "samples" and "data," respectively. It is argued that this may present risks to scientific rigor and the populations the research aims to help. Inconsistent terminology may cause misunderstandings regarding study design thus affecting reproducibility of results. Depersonalization and dehumanization may be another byproduct [147]. This may cause individuals and communities to become stigmatized, echoing the findings of the studies discussed earlier. To alleviate this, it is suggested that more human-centered methods such as participatory design should be considered where interviews and field studies are conducted. However, this is at odds with the challenges highlighted in the Datasets on Social Media and Mental Health section where acquiring sizable datasets through such methods is largely intractable [32].

With respect to schizophrenia, Välimäki et al [58] determined via their review that the perceptions and risks of social media interventions are largely unexplored. However, there are suggestions that some clinicians fear that the use of web-based peer support without professional moderation may cause anxiety in the bearer of the disorder [148]. Cognitive deficits in people with schizophrenia can inhibit the development of digital skills [149], evidencing clinicians' misgivings.

Discussion

Principal Findings

We have seen that there is growing interest in methods for predicting mental health status using social media data, particularly those that involve NLP. Enthusiasm has been notable since the COVID-19 pandemic when interest in remote monitoring of individual- and population-level mental states grew. Indeed, the search strategy followed for this review yielded more articles in the years 2020 to 2021 than in the previous 20 years; 917 and 903, respectively (Multimedia Appendix 1). Methods have progressed from those that use features from text as input to traditional ML algorithms, to increasingly sophisticated approaches using transformer-based LMs and now, LLMs. The research community has endeavored to provide social media data to support this work and to do so in ways that are increasingly sensitive to ethical and privacy concerns of the participants involved.

Our review has not only shown depression to be the most common condition reported in publicly available datasets, it also highlights the need for much larger samples where contextual information on this and other conditions, such as a date for the diagnosis and not just its presence, is denoted to a high degree of accuracy. Having such data would likely strengthen results found in longitudinal studies, most of which have focused on depression as well, providing more opportunities for predictions before an eventual diagnosis is formalized [57]. Obtaining such ground truth data via traditional confidential questionnaires is time-consuming and intrusive from the participant's point of view [134]. A solution may involve obtaining consented access to EHRs to accompany the users' social media postings, as piloted by Eichstaedt et al [150]. Indeed, this means of verification is crucial in studies that consider schizophrenia because diagnosis self-disclosure statements, although having high sensitivity [151], may lack specificity [120,121]. In any case, social media data obtained also needs to be broadened to better support NLP methods. For example, Reddit datasets should routinely include postings from mental health subreddits in addition to other subreddits [79]. This would help ensure that LMs pretrained on such data are less prone to biases that may dampen the effectiveness of methods developed thereon [80]. LLM-driven technologies,

such as ChatGPT and its successors, will likely underpin methods in the immediate future. However, a fledgling attempt involving Reddit posts found that models were better able to detect BDI-II-measured depression symptoms using authentic data rather than LLM (GPT-3) synthesized data [48]. It was suggested that improved prompt manipulation is needed to produce synthesized data that are less stilted. Another role for LLMs may be in the automatic labeling of mental health dataset instances. Ramos dos Santos and Paraboni [49] produced evidence that an LLM (GPT-3.5) can perform promisingly (72% sensitivity) when distinguishing between tweets of users who may have depression and users who likely do not. LLMs may eventually offer a far less costly alternative to dataset labeling than manual approaches. Psychiatry literature suggests LLM performance in these settings could be improved by prescribing potentially time-consuming trials to learn what prompts are best suited for specific tasks [152,153]. Instruction fine-tuning is one such proposal for improving LLM performance. LLMs including GPTs are trained on very large, nondomain specific datasets such as Wikipedia. However, further training an LLM on smaller, domain-specific datasets may enhance its performance in that domain. For example, when comparing the performance of a nonfine-tuned LLM and its fine-tuned counterpart, Xu et al [154] measured a 23.4% increase in accuracy across 6 different mental health prediction tasks involving Reddit data. However, fine-tuning ought to be performed using a wider range of domain-specific datasets, which is advisable to reduce biases in the resulting LLM.

With respect to population-level and individual-level longitudinal studies, we found the analysis of emotions conveyed in social media posts to be an underrepresented topic of research in this area. Consideration of fine-grained language features may also help to better predict depression severity over time [56]. In fact, the most promising approaches will probably involve those that augment NLP; multimodal methods that consider nontext features from social media activities are expected to help provide richer findings. In Twitter, this may involve consideration of user geolocations and profile images. For example, Ghosh et al [155] attempted to distinguish between users with depression and users without depression by considering their profile images and the text of their profile descriptions. A classifier that used features from the profile image outperformed a baseline classifier that used only features from the profile description by approximately 10% in terms of the F_1-score. While profile images may be predictive of users' mental health statuses to an extent, there are confounding factors that these multimodal methods must address. For example, people with depression are likely on social media platforms to display positive-looking pictures (including profile images) as opposed to negative-looking ones, according to Ghosh et al [155]. This perhaps counterintuitive phenomenon has been dubbed "smiling depression" and training of multimodal models with larger, labeled datasets is needed so that they may become more discerning in these conditions. Semwal et al [156] have also evidenced in similar experimental settings that information contained in tweet text and profile images complement one another and ought to be used in alliance. They recorded that their multimodal model outperformed their best-performing textual and image-only models by 3.5% and 27.1%, respectively, in terms of F_1-score. Therefore, the conclusion was that images seem to contain significant information regarding a user's mental health status, thus motivating further study in mental health status prediction. Meanwhile in Reddit, multimodal methods may involve time-aware consideration of user posts. One study considered the relative time between posts as a feature for distinguishing between Reddit and Twitter users with and without depression [157]. Obtaining an F_1-score of 0.93 with Reddit and 0.87 with Twitter, it was concluded that a time-aware approach to classification is more effective where posting frequency is relatively high. The supposition is that the concise nature of Twitter posts, compared to the often much lengthier posts on Reddit, contributes to users posting more frequently on Twitter. A further study considered a multimodal approach with emphasis on emojis, again in the task of distinguishing between users with and without depression on both Twitter and Reddit [158]. With F_1-scores of 0.80 and 0.95 being achieved for Twitter and Reddit, respectively, it could be concluded, given the 2 studies that have just been outlined, that different multimodal approaches will be suitable for different platforms.

The advent of multimodal approaches may also help allay a privacy-related concern that our review has brought to the fore. The public has expressed concerns about methods for predicting mental health status that harness primitive keyword searches due to the risk of unreliable output. Naturally, a social media user may be affronted at receiving an incorrect diagnosis of depression, anxiety, or schizophrenia [59,60]. Multimodal approaches that more accurately capture people's real-life behaviors are thus being pursued [159]. It is not only methods that need to improve to gain public confidence; more fundamentally, the means of collecting data for use in any study need to be more explicit and have user consent. Inviting users to grant access to their social media data for research purposes on a large scale, perhaps at the point that they publish a social media posting, could become widespread [160]. However, such invitations must be accessible to a wide demographic. Privacy literacy, which describes one's understanding of the risks of sharing information on social media, is considered more prevalent among women than men, for example [161].

Finally, our literature search returned many articles that consider the effects of social media use on a user's levels of depression and anxiety (Multimedia Appendix 1). A primary hypothesis, greatly debated in specialist literature [162], is that extended or otherwise distinct patterns of social media use may cause or exacerbate these mental health disorders. This was not the subject area of this study, but our results on the volume of published articles suggest that this related matter perhaps merits a review of its own.

Potential Clinical Applications

With reference to the research covered in this review, we now consider the potential clinical applications of using AI on social media data. These include (1) evaluating data at a population level to inform health care delivery and policy making, (2) identifying and providing access to support and interventions for those at risk of developing mental health problems, and (3) monitoring existing individual patients to detect and intervene

at early signs of relapse [163]. The third application area was underrepresented in methods for predicting mental health status literature.

At a population level, AI and NLP may be used to navigate large volumes of data to inform clinical needs in a particular area, to identify changing patterns of mental illness across populations and time, to better understand patients' experiences and perceptions of health services, and to identify patterns of risky behaviors among certain demographics (eg, young people accessing accounts linked to proanorexia or encouraging self-harm). As noted earlier, NLP was used to evaluate large volumes of social media data during the COVID-19 pandemic and identify the specific concerns of people living with mental illness, including health anxieties, loneliness, and suicidality [164]. This type of information can be used to inform resource allocation in health services and the development of government policies. Crucially, this analysis can be performed relatively quickly (particularly compared to traditional research methods), which is essential during periods of instability, such as a public health crisis, where decisions need to be made rapidly.

At an individual level, AI may be used to identify people at risk of or living with mental health problems and enable organizations to provide early intervention support. There are some concerns regarding consent, data use, and privacy, as noted in the Ethical Aspects of the Data and Analysis of Mental Health section. Interestingly, while both young people and mental health professionals somewhat agree that social media companies should use AI to proactively detect users at risk of suicide or self-harm and signpost them helpful information and resources, they felt more strongly that AI capabilities should be used to promote helpful content such as psychoeducation [165]. In addition, there are logistical challenges to doing this, such as how individual data collected by global platforms can be harnessed by localized health care providers to support care.

Despite these challenges, social media has been proven to be a useful tool to identify relevant individuals for research, including delivering interventions to young people living with eating disorders [166] and who have been exposed to suicide (eg, a friend or family member had died by suicide or attempted suicide) [165] and using Facebook data to detect relapse in patients with schizophrenia [167]. As an example, Birnbaum et al [167] used LIWC on extracted Facebook archives and concurrent medical records for participants with psychosis. Researchers built an individual-centric classifier to predict re-admission to the hospital due to exacerbation of psychotic symptoms. However, the sensitivity of the prediction model was low (38%) indicating that the algorithm only identified a small proportion of all those who relapsed. Furthermore, the algorithm was applied to retrospective Facebook archives and paired with retrospective medical records, all with explicit consent from participants. The use of social media data to prospectively predict relapse in patients is likely to be considerably more challenging. As the authors noted, patients may change their social media behavior if they are aware that they are being actively monitored by their care team.

While the AI-driven mental health status prediction methods outlined may appear to lend themselves readily for use in clinical practice, there are limitations that need to be addressed before they are adopted. A chief limitation, as already mentioned in this review, is the likelihood of bias in methods based on data that do not represent diverse populations [168,169]. Thus, they may not be able to account, for example, for the fact that mental health conditions may present differently in different people. This is challenging to overcome because the field of mental health care is limited in its access to large, high-quality datasets. Compounding these limitations is the fact that the underlying biological processes of mental health disorders are still poorly understood meaning that models must be bootstrapped from observations rather than be derived from first principles. Indeed, the nature of decision-making in mental health care can be far more complex than that of other clinical areas. Indicative of this is the fact that the specific and objective task of tumor identification from an image is already successfully supported by AI-driven methods [170]. Mental health care therefore desires AI-driven methods that are transparent, explainable, and able to provide guidance to clinicians [26,169,171].

Limitations

We have reviewed the literature in what we deemed 4 chief areas in the realm of predicting mental health status. There are opportunities for greater depth of coverage in these areas and they could be the subject of review articles of their own. There is also scope for a greater breadth of coverage that could fuel follow-up studies. For example, our coverage has primarily considered research related to NLP, with occasional deference to multimodal alternatives. Visual computing provides techniques applicable to data from predominantly image-based platforms, such as Instagram [6,9]. Experts in computer vision may therefore be able to provide greater insight here.

Being a narrative review, the nature of article selection and analysis is somewhat subjective. To mollify this, we used a well-defined search and selection process that borrowed features often used in systematic reviews [29] (Methods section).

In addition, we only considered articles in which the participants of the studies self-reported a diagnosis of depression, anxiety, or schizophrenia; however, more widely any sort of information garnered from a social media posting should be treated like a self-report. While this confers a certainty that the input reflects the experiences and beliefs of the social media user, providing the opportunity to automatically accrue large datasets that have information about mental health statuses, this approach also has weaknesses that have been explored in the psychopathological literature [172]. For example, compared to a manually compiled and curated dataset, there are likely to be more false positive instances of nearly any common diagnosis, although there may also be false negative instances or controls that do in fact bear a mental health disorder [32] are also possible. In the case of schizophrenia, the condition itself might be partly responsible for the unreliability of self-reports, creating an even larger weakness for automatically constructed datasets as previously highlighted.

We should also mention that the social media platforms covered in this review, including Facebook, Twitter, and Reddit, are ostensibly English-language platforms. This coverage is perhaps by virtue of our literature search and selection strategy, which

excluded non-English language articles. Therefore, we acknowledge that the findings presented in this paper may well not apply to non-English language platforms such as Weibo [173] and VK [174], which are Chinese and Russian language platforms, respectively. A complementary narrative review that considers social media platforms concerning these languages and cultures could form future work.

Finally, we highlight a theme that has recurred throughout this review, which is that of biases in predicting mental health status research. Addressing these biases, or at least being aware of them, is crucial for ensuring accurate and generalizable findings. This review has concerned predominantly English-language social media platforms, which in turn, largely reflect Western culture. Therefore, when such findings are reported in the literature it must be ceded that they might not generalize to social media platforms that predominantly reflect Eastern culture. In any case, there are other platform-related biases that must be considered; certain platforms may be used largely by certain demographics. We have already noted that on platforms such as Facebook, Twitter, and Instagram, male users are in the majority [81] and that social media users are generally well-educated and affluent [82]. Cohort-based strategies for dataset construction have been trialed to account for these biases [36]. There are also user-oriented biases that may distort datasets. A user's posting habits may change over time and convey a distorted view of their life and experiences [50]. This behavior may be influenced by reports published in traditional print media on the negative consequences of social media use [175]. It may also be influenced by the proliferation of use-limiting tools, which encourage users to choose carefully the personal information they share on social media platforms [176]. On a collective scale, certain users may post content significantly more frequently than others, creating imbalances in datasets and subsequent models. This is evident in 2 of the datasets we have covered [86,90]. Data augmentation is one approach that may alleviate this problem [94,114], while another includes data synthesis via LLMs [48]. Lastly, we should mention confirmation bias, which involves people's tendency to seek data that support their beliefs and ignore or distort data contradicting them [177]. Wherever possible, a selection of appropriate datasets ought to be used in experimental setups so that conclusions are better balanced. In general, it is suggested that future research in the domain of mental health status prediction should seek and report data biases to enhance the reliability of findings [27].

Conclusions

The research area of predicting mental health status is receiving much attention, particularly in recent years. The COVID-19 era appears to have been the catalyst for the expanding interest. Further work needs to be completed with respect to methods for predicting mental health status before they may be considered sufficiently reliable for clinical purposes. We have documented public misgivings about text-only approaches, particularly those that rely on keyword searches. We have also acknowledged that image-based social media platforms such as Instagram are in wide use. Therefore, to help gain public confidence, methods will likely need to be multimodal. That is, they will need to generalize to text-, voice-, image-, and video-based social media data. The pursuit is merited to help relieve strain on health care and mental health services. In fact, the integration of automated early health care intervention methods and traditional methods may be advantageous.

This work cannot take place in a vacuum; however, due consideration must be given to the ethical concerns regarding the collection and use of social media users' data. Consent from users needs to be sought, perhaps by providing them with the opportunity to donate their social media data or by allowing them to choose to share their data for research purposes on a post-by-post basis. In any event, the purposes of collecting such data ought to be made clear to users through transparent data use agreements. Then, when data are subsequently compiled into datasets for public release, anonymization of the user accounts they contain is essential.

Acknowledgments

AJL was supported by the National Centre for Mental Health, a collaboration between Cardiff, Swansea, and Bangor Universities, funded by the Welsh government through Health and Care Research Wales. SES and AFP were supported by the European Union's Horizon 2020 research and innovation program under grant agreement 964874. AFP was also supported by a Medical Research Council Programme Grant (MR/Y004094/1). JCC was supported by a United Kingdom Research and Innovation Future Leaders Fellowship.

Data Availability

Data sharing is not applicable to this paper as no datasets were generated or analyzed during this study.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Verbose exposition of the 4-stage literature selection process.
[XLSX File (Microsoft Excel File), 272 KB-Multimedia Appendix 1]

References

1. GBD 2019 Mental Disorders Collaborators. Global, regional, and national burden of 12 mental disorders in 204 countries and territories, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet Psychiatry. Feb 2022;9(2):137-150. [FREE Full text] [doi: 10.1016/S2215-0366(21)00395-3] [Medline: 35026139]
2. Picardi A, Lega I, Tarsitani L, Caredda M, Matteucci G, Zerella MP, et al. SET-DEP Group. A randomised controlled trial of the effectiveness of a program for early detection and treatment of depression in primary care. J Affect Disord. Jul 01, 2016;198:96-101. [doi: 10.1016/j.jad.2016.03.025] [Medline: 27015158]
3. De CM, Counts S, Horvitz E. Social media as a measurement tool of depression in populations. In: Proceedings of the 5th Annual ACM Web Science Conference. 2013. Presented at: WebSci '13; May 2-4, 2013:47-56; Paris, France. URL: https://dl.acm.org/doi/10.1145/2464464.2464480 [doi: 10.1145/2464464.2464480]
4. Ammari T, Schoenebeck S. Networked empowerment on Facebook groups for parents of children with special needs. In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems. 2015. Presented at: CHI '15; April 18-23, 2015:2805-2814; Seoul, Republic of Korea. URL: https://dl.acm.org/doi/10.1145/2702123.2702324 [doi: 10.1145/2702123.2702324]
5. Tadesse MM, Lin H, Xu B, Yang L. Detection of depression-related posts in Reddit social media forum. IEEE Access. 2019;7:44883-44893. [doi: 10.1109/access.2019.2909180]
6. Andalibi N, Ozturk P, Forte A. Sensitive self-disclosures, responses, and social support on Instagram: the case of #Depression. In: Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computi. 2017. Presented at: CSCW '17; February 25-March 1, 2017:1485-1500; Portland, OR. URL: https://dl.acm.org/doi/10.1145/2998181.2998243 [doi: 10.1145/2998181.2998243]
7. Dinu A, Moldovan A. Automatic detection and classification of mental illnesses from general social media texts. In: Proceedings of the 2021 International Conference on Recent Advances in Natural Language Processing. 2021. Presented at: RANLP '21; Sep 1-3, 2021:358-366; Virtual Event. URL: https://aclanthology.org/2021.ranlp-1.41.pdf [doi: 10.26615/978-954-452-072-4_041]
8. Singh A, Singh J. Automation of detection of social network mental disorders – a review. IOP Conf Ser Mater Sci Eng. Jan 01, 2021;1022(1):012008. [FREE Full text] [doi: 10.1088/1757-899x/1022/1/012008]
9. Muhammad KA. Unveiling the emotional and psychological states of Instagram users: a deep learning approach to mental health analysis. Inf Sci Lett. May 01, 2023;12(5):1877-1890. [doi: 10.18576/isl/120531]
10. Ganguly C, Nayak S, Gupta A. Mental health impact of COVID-19 and machine learning applications in combating mental disorders: a review. In: Jain S, Pandey K, Jain P, Seng KP, editors. Artificial Intelligence, Machine Learning, and Mental Health in Pandemics. New York, NY. Academic Press; Jan 2022:1-51.
11. Holmes EA, O'Connor RC, Perry VH, Tracey I, Wessely S, Arseneault L, et al. Multidisciplinary research priorities for the COVID-19 pandemic: a call for action for mental health science. Lancet Psychiatry. Jun 2020;7(6):547-560. [FREE Full text] [doi: 10.1016/S2215-0366(20)30168-1] [Medline: 32304649]
12. Chancellor S, Birnbaum M, Caine E, Silenzio V, De CM. A taxonomy of ethical tensions in inferring mental health states from social media. In: Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency. 2019. Presented at: FAT* '19; January 29-31, 2019:79-88; Atlanta, GA. URL: https://dl.acm.org/doi/10.1145/3287560.3287587 [doi: 10.1145/3287560.3287587]
13. Hirschberg J, Manning CD. Advances in natural language processing. Science. Jul 17, 2015;349(6245):261-266. [doi: 10.1126/science.aaa8685] [Medline: 26185244]
14. Weizenbaum J. ELIZA—a computer program for the study of natural language communication between man and machine. Commun ACM. Jan 1966;9(1):36-45. [doi: 10.1145/365153.365168]
15. Rajaraman V. From ELIZA to ChatGPT: history of human-computer conversation. Reson. Jun 21, 2023;28(6):889-905. [doi: 10.1007/s12045-023-1620-6]
16. Gilman ES, Kot S, Engineer M, Dixon E. Training adults with mild to moderate dementia in ChatGPT: exploring best practices. In: Proceedings of the 29th International Conference on Intelligent User Interface. 2024. Presented at: IUI '24; March 18-21, 2024:101-106; Greenville, SC. URL: https://dl.acm.org/doi/abs/10.1145/3640544.3645230 [doi: 10.1145/3640544.3645230]
17. Whissell CM. The dictionary of affect in language. In: Plutchik R, Kellerman H, editors. The Measurement of Emotions. New York, NY. Academic Press; 1989:113-131.
18. Whissell dictionary of affect in language. Spirituality & The Brain. URL: https://www.god-helmet.com/wp/whissel-dictionary-of-affect/index.htm [accessed 2024-08-02]
19. Welcome to Dokteronline. Dokteronline. URL: https://www.dokteronline.com/ [accessed 2024-02-29]
20. Johnsen JK, Rosenvinge JH, Gammon D. Online group interaction and mental health: an analysis of three online discussion forums. Scand J Psychol. Dec 2002;43(5):445-449. [doi: 10.1111/1467-9450.00313] [Medline: 12500784]
21. McKenna KY, Bargh JA. Coming out in the age of the internet: identity "demarginalization" through virtual group participation. Pers Soc Psychol. 1998;75(3):681-694. [doi: 10.1037//0022-3514.75.3.681]
22. Thorn P, La Sala L, Hetrick S, Rice S, Lamblin M, Robinson J. Motivations and perceived harms and benefits of online communication about self-harm: an interview study with young people. Digit Health. 2023;9:20552076231176689. [FREE Full text] [doi: 10.1177/20552076231176689] [Medline: 37252260]

23. Haker H, Lauber C, Rössler W. Internet forums: a self-help approach for individuals with schizophrenia? Acta Psychiatr Scand. Dec 2005;112(6):474-477. [doi: 10.1111/j.1600-0447.2005.00662.x] [Medline: 16279878]
24. Moreno MA, Jelenchick LA, Egan KG, Cox E, Young H, Gannon KE, et al. Feeling bad on Facebook: depression disclosures by college students on a social networking site. Depress Anxiety. Jun 2011;28(6):447-455. [FREE Full text] [doi: 10.1002/da.20805] [Medline: 21400639]
25. Winn JG, Hao T, Hardaway JW, Oh H. Redefining searching in non-medical sciences systematic reviews: the ascendance of Google Scholar as the primary database. J Libr Inf Sci. Jun 16, 2024;5:1. [doi: 10.1177/09610006241256393]
26. Martín-Martín A, Thelwall M, Orduna-Malea E, Delgado López-Cózar E. Google Scholar, Microsoft Academic, Scopus, Dimensions, Web of Science, and OpenCitations' COCI: a multidisciplinary comparison of coverage via citations. Scientometrics. 2021;126(1):871-906. [FREE Full text] [doi: 10.1007/s11192-020-03690-4] [Medline: 32981987]
27. Gusenbauer M. Google Scholar to overshadow them all? Comparing the sizes of 12 academic search engines and bibliographic databases. Scientometrics. Nov 10, 2018;118(1):177-214. [doi: 10.1007/s11192-018-2958-5]
28. Martín-Martín A, Orduna-Malea E, Thelwall M, López-Cózar E. Google Scholar, Web of Science, and Scopus: a systematic comparison of citations in 252 subject categories. J Informetr. Nov 2018;12(4):1160-1177. [doi: 10.1016/j.joi.2018.09.002]
29. Ferrari R. Writing narrative style literature reviews. Med Writ. Dec 23, 2015;24(4):230-235. [doi: 10.1179/2047480615z.000000000329]
30. Szeto MD, Barber C, Ranpariya VK, Anderson J, Hatch J, Ward J, et al. Emojis and Emoticons in health care and dermatology communication: narrative review. JMIR Dermatol. 2022;5(3):e33851. [FREE Full text] [doi: 10.2196/33851] [Medline: 36405493]
31. De Choudhury M, Gamon M, Counts S, Horvitz E. Predicting depression via social media. Proc Int AAAI Conf Weblogs Soc Media. Aug 03, 2021;7(1):128-137. [doi: 10.1609/icwsm.v7i1.14432]
32. Coppersmith G, Dredze M, Harman C. Quantifying mental health signals in Twitter. In: Proceedings of the 2014 Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality. 2014. Presented at: CLCP-LSCR '14; June 27, 2014:51-60; Baltimore, MD. URL: https://aclanthology.org/W14-3207.pdf [doi: 10.3115/v1/w14-3207]
33. Yates A, Cohan A, Goharian N. Depression and self-harm risk assessment in online forums. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. 2017. Presented at: EMNLP '17; September 7-11, 2017:2968-2978; Copenhagen, Denmark. URL: https://aclanthology.org/D17-1322.pdf [doi: 10.18653/v1/d17-1322]
34. MacAvaney S, Desmet B, Cohan A, Soldaini L, Yates A, Zirikly A, et al. RSDD-time: temporal annotation of self-reported mental health diagnoses. In: Proceedings of the 5th Workshop on Computational Linguistics and Clinical Psychology. 2018. Presented at: CLPsych '18; June 5, 2018:168-173; New Orleans, LA. URL: https://aclanthology.org/W18-0618.pdf [doi: 10.18653/v1/w18-0618]
35. Cohan A, Desmet B, Yates A, Soldaini L, MacAvaney S, Goharian N. SMHD: a large-scale resource for exploring online language usage for multiple mental health conditions. In: Proceedings of the 27th International Conference on Computational Linguistics. 2018. Presented at: COLING '18; August 20-26, 2018:1485-1497; Santa Fe, MX. URL: https://aclanthology.org/C18-1126.pdf [doi: 10.18653/v1/w18-0618]
36. Amir S, Dredze M, Ayers J. Mental health surveillance over social media with digital cohorts. In: Proceedings of the 6th Workshop on Computational Linguistics and Clinical Psychology. 2019. Presented at: CLPsych '19; June 6, 2019:114-120; Minneapolis, MN. URL: https://aclanthology.org/W19-3013.pdf [doi: 10.18653/v1/w19-3013]
37. Parapar J, Martín-Rodilla P, Losada D, Crestani F. Overview of eRisk 2021: early risk prediction on the internet. In: Proceedings of the 12th International Conference on Experimental IR Meets Multilinguality, Multimodality, and Interaction. 2021. Presented at: CLEF '21; September 21-24, 2021:324-344; Virtual Event. URL: https://link.springer.com/chapter/10.1007/978-3-030-85251-1_22 [doi: 10.1007/978-3-030-85251-1_22]
38. Resnik P, Armstrong W, Claudino L, Nguyen T, Nguyen V, Boyd-Graber J. Beyond LDA: exploring supervised topic modeling for depression-related language in Twitter. In: Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology. 2015. Presented at: CLPsych '15; June 5, 2015:99-107; Denver, CO. URL: https://aclanthology.org/W15-1212.pdf [doi: 10.3115/v1/w15-1212]
39. Mitchell M, Hollingshead K, Coppersmith G. Quantifying the language of schizophrenia in social media. In: Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology. 2015. Presented at: CLPsych '15; June 5, 2015:11-20; Denver, CO. URL: https://aclanthology.org/W15-1202.pdf [doi: 10.3115/v1/w15-1202]
40. Tsugawa S, Kikuchi Y, Kishino F, Nakajima K, Itoh Y, Ohsaki H. Recognizing depression from Twitter activity. In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems. 2015. Presented at: CHI EA '15; April 18-23, 2015:3187-3196; Seoul, Republic of Korea. URL: https://dl.acm.org/doi/10.1145/2702123.2702280 [doi: 10.1145/2702123.2702280]
41. Birnbaum ML, Ernala SK, Rizvi AF, De Choudhury M, Kane JM. A collaborative approach to identifying social media markers of schizophrenia by employing machine learning and clinical appraisals. J Med Internet Res. Aug 14, 2017;19(8):e289. [FREE Full text] [doi: 10.2196/jmir.7956] [Medline: 28807891]
42. Guntuku SC, Yaden DB, Kern ML, Ungar LH, Eichstaedt JC. Detecting depression and mental illness on social media: an integrative review. Curr Opin Behav Sci. Dec 2017;18:43-49. [doi: 10.1016/j.cobeha.2017.07.005]

43. Reece AG, Reagan AJ, Lix KL, Dodds PS, Danforth CM, Langer EJ. Forecasting the onset and course of mental illness with Twitter data. Sci Rep. Oct 11, 2017;7(1):13006. [FREE Full text] [doi: 10.1038/s41598-017-12961-9] [Medline: 29021528]
44. Burdisso SG, Errecalde M, Montes-y-Gómez M. A text classification framework for simple and effective early depression detection over social media streams. Expert Syst Appl. Nov 2019;133:182-197. [doi: 10.1016/j.eswa.2019.05.023]
45. Owen D, Camacho-Collados J, Anke LE. Towards preemptive detection of depression and anxiety in Twitter. In: Proceedings of the 5th Social Media Mining for Health Applications. 2020. Presented at: SMM4H '20; December 12, 2020:82-89; Barcelona, Spain. URL: https://aclanthology.org/2020.smm4h-1.12.pdf
46. Malviya K, Roy B, Saritha S. A transformers approach to detect depression in social media. In: Proceedings of the 2021 International Conference on Artificial Intelligence and Smart Systems. 2021. Presented at: ICAIS '21; March 25-27, 2021:718-723; Coimbatore, India. URL: https://ieeexplore.ieee.org/document/9395943 [doi: 10.1109/icais50930.2021.9395943]
47. Sharma T, Panchendrarajan R, Saxena A. Characterisation of mental health conditions in social media using deep learning techniques. In: Hong TP, Serrano-Estrada L, Saxena A, Biswas A, editors. Deep Learning for Social Media Data Analytics. Cham, Switzerland. Springer; 2022:157-176.
48. Bucur AM. Utilizing ChatGPT generated data to retrieve depression symptoms from social media. arXiv. Preprint posted online July 5, 2023. [FREE Full text]
49. Ramos dos Santos W, Paraboni I. Prompt-based mental health screening from social media text. arXiv. Preprint posted online January 11, 2024. [FREE Full text] [doi: 10.5753/brasnam.2024.1879]
50. Schwartz H, Eichstaedt J, Kern M, Park G, Sap M, Stillwell D, et al. Towards assessing changes in degree of depression through Facebook. In: Proceedings of the 2014 Workshop on Computational Linguistics and Clinical Psychology. 2014. Presented at: CLPsych '14; June 27, 2014:118-125; Baltimore, MD. URL: https://aclanthology.org/W14-3214.pdf [doi: 10.3115/v1/w14-3214]
51. Loveys K, Crutchley P, Wyatt E, Coppersmith G. Small but mighty: affective micropatterns for quantifying mental health from social media language. In: Proceedings of the 4th Workshop on Computational Linguistics and Clinical Psychology. 2017. Presented at: CLPsych '17; August 3, 2017:85-95; Vancouver, BC. URL: https://aclanthology.org/W17-3110.pdf [doi: 10.18653/v1/w17-3110]
52. Hswen Y, Naslund JA, Brownstein JS, Hawkins JB. Monitoring online discussions about suicide among Twitter users with schizophrenia: exploratory study. JMIR Ment Health. Dec 13, 2018;5(4):e11483. [FREE Full text] [doi: 10.2196/11483] [Medline: 30545811]
53. Seabrook EM, Kern ML, Fulcher BD, Rickard NS. Predicting depression from language-based emotion dynamics: longitudinal analysis of Facebook and Twitter status updates. J Med Internet Res. May 08, 2018;20(5):e168. [FREE Full text] [doi: 10.2196/jmir.9267] [Medline: 29739736]
54. Chen X, Sykora M, Jackson T, Elayan S. What about mood swings? Identifying depression on Twitter with temporal measures of emotions. In: Proceedings of the 2018 Web Conference. 2018. Presented at: WWW '18; April 23-27, 2018:1653-1660; Lyon, France. URL: https://dl.acm.org/doi/fullHtml/10.1145/3184558.3191624 [doi: 10.1145/3184558.3191624]
55. Zhang Y, Lyu H, Liu Y, Zhang X, Wang Y, Luo J. Monitoring depression trends on Twitter during the COVID-19 pandemic: observational study. JMIR Infodemiology. 2021;1(1):e26769. [FREE Full text] [doi: 10.2196/26769] [Medline: 34458682]
56. Kelley SW, Gillan CM. Using language in social media posts to study the network dynamics of depression longitudinally. Nat Commun. Feb 15, 2022;13(1):870. [FREE Full text] [doi: 10.1038/s41467-022-28513-3] [Medline: 35169166]
57. Owen D, Antypas D, Hassoulas A, Pardiñas AF, Espinosa-Anke L, Collados JC. Enabling early health care intervention by detecting depression in users of web-based forums using language models: longitudinal analysis and evaluation. JMIR AI. Mar 24, 2023;2:e41205. [FREE Full text] [doi: 10.2196/41205] [Medline: 37525646]
58. Välimäki M, Athanasopoulou C, Lahti M, Adams CE. Effectiveness of social media interventions for people with schizophrenia: a systematic review and meta-analysis. J Med Internet Res. Apr 22, 2016;18(4):e92. [FREE Full text] [doi: 10.2196/jmir.5385] [Medline: 27105939]
59. Mikal J, Hurst S, Conway M. Ethical issues in using Twitter for population-level depression monitoring: a qualitative study. BMC Med Ethics. Apr 14, 2016;17:22. [FREE Full text] [doi: 10.1186/s12910-016-0105-5] [Medline: 27080238]
60. Conway M, O'Connor D. Social media, big data, and mental health: current advances and ethical implications. Curr Opin Psychol. Jun 2016;9:77-82. [FREE Full text] [doi: 10.1016/j.copsyc.2016.01.004] [Medline: 27042689]
61. Chancellor S, Baumer EP, De Choudhury M. Who is the "human" in human-centered machine learning: the case of predicting mental health from social media. Proc ACM Hum Comput Interact. Nov 07, 2019;3(CSCW):1-32. [doi: 10.1145/3359249]
62. Nicholas J, Onie S, Larsen ME. Ethics and privacy in social media research for mental health. Current Psychiatry Reports. Nov 23, 2020;22(12):84. [doi: 10.1007/s11920-020-01205-9] [Medline: 33225404]
63. Vornholt P, De Choudhury M. Understanding the role of social media-based mental health support among college students: survey and semistructured interviews. JMIR Ment Health. Jul 12, 2021;8(7):e24512. [FREE Full text] [doi: 10.2196/24512] [Medline: 34255701]

64. Cai N, Revez JA, Adams MJ, Andlauer TF, Breen G, Byrne EM, MDD Working Group of the Psychiatric Genomics Consortium, et al. Minimal phenotyping yields genome-wide association signals of low specificity for major depression. Nat Genet. Apr 2020;52(4):437-447. [FREE Full text] [doi: 10.1038/s41588-020-0594-5] [Medline: 32231276]
65. Cong Q, Feng Z, Li F, Xiang Y, Rao G, Tao C. X-A-BiLSTM: a deep learning approach for depression detection in imbalanced data. In: Proceedings of the 2018 IEEE International Conference on Bioinformatics and Biomedicine. 2018. Presented at: BIBM '18; December 3-6, 2018:1624-1627; Madrid, Spain. URL: https://ieeexplore.ieee.org/document/8621230 [doi: 10.1109/bibm.2018.8621230]
66. Marnauzs S, Kalita J. A domain independent social media depression detection model. In: The 33rd AAAI Conference on Artificial Intelligence. 2019. Presented at: AI '19; January 27-February 1, 2019:50-55; Honolulu, HI. URL: https://faculty.uccs.edu/jkalita/wp-content/uploads/sites/45/2024/02/00Proceedings2019Optimized.pdf#page=62
67. Trifan A, Semeraro D, Drake J, Bukowski R, Oliveira JL. Social media mining for postpartum depression prediction. Stud Health Technol Inform. Jun 16, 2020;270:1391-1392. [doi: 10.3233/SHTI200457] [Medline: 32570674]
68. Bucur AM, Cosma A, Dinu LP. Early risk detection of pathological gambling, self-harm and depression using BERT. arXiv. Preprint posted online June 30, 2021. [FREE Full text]
69. Ali J, Ngo DQ, Bhattacharjee A, Maiti T, Singh T, Mei J. Depression detection: text augmentation for robustness to label noise in self-reports. In: Bertolaso M, Capone L, Rodríguez-Lluesma C, editors. Digital Humanism: A Human-Centric Approach to Digital Technologies. Cham, Switzerland. Springer; 2022:81-103.
70. Kulkarni H, MacAvaney S, Goharian N, Frieder O. Knowledge augmentation for early depression detection. In: Proceedings of the 7th International Workshop on Health Intelligence. 2023. Presented at: W3PHAI '23; February 13-14, 2023:175-191; Washington, DC. [doi: 10.1007/978-3-031-36938-4_14]
71. Souza VB, Nobre J, Becker K. Characterization of anxiety, depression, and their comorbidity from texts of social networks. In: Proceedings of the 35th Brazilian Symposium on Databases. 2022. Presented at: SBBD '20; September 28-October 1, 2020:121-132; Virtual Event. URL: https://sol.sbc.org.br/index.php/sbbd/article/view/13630/13478 [doi: 10.5753/sbbd.2020.13630]
72. Chen Z, Yang R, Fu S, Zong N, Liu H, Huang M. Detecting reddit users with depression using a hybrid neural network SBERT-CNN. In: Proceedings of the 2023 IEEE 11th International Conference on Healthcare Informatics. 2023. Presented at: ICHI '23; June 26-29, 2023:193-199; Houston, TX. URL: https://www.computer.org/csdl/proceedings-article/ichi/2023/026300a193/1SN7B0FKEmI [doi: 10.1109/ichi57859.2023.00035]
73. Buddhitha P, Inkpen D. Multi-task learning to detect suicide ideation and mental disorders among social media users. Front Res Metr Anal. 2023;8:1152535. [FREE Full text] [doi: 10.3389/frma.2023.1152535] [Medline: 37138946]
74. Khan M, Sakib S, Habib A, Hossain M. A machine learning and deep learning approach to classify mental illness with the collaboration of natural language processing. In: Proceedings of the 2022 International Conference on Information and Communication Technology for Development. 2022. Presented at: ICICTD '22; July 29-30, 2022:83-94; Khulna, Bangladesh. URL: https://link.springer.com/chapter/10.1007/978-981-19-7528-8_7 [doi: 10.1007/978-981-19-7528-8_7]
75. Zanwar S, Wiechmann D, Qiao Y, Kerz E. Exploring hybrid and ensemble models for multiclass prediction of mental health status on social media. In: Proceedings of the 13th International Workshop on Health Text Mining and Information Analysis. 2022. Presented at: LOUHI '22; December 7, 2022:184-196; Abu Dhabi, United Arab Emirates. URL: https://aclanthology.org/2022.louhi-1.21.pdf [doi: 10.18653/v1/2022.louhi-1.21]
76. Sekulic I, Strube M. Adapting deep learning methods for mental health prediction on social media. In: Proceedings of the 5th Workshop on Noisy User-generated Text. 2019. Presented at: WNUT '19; November 4, 2019:322-327; Hong Kong, China. URL: https://aclanthology.org/D19-5542.pdf [doi: 10.18653/v1/d19-5542]
77. Borba de Souza V, Campos Nobre J, Becker K. DAC stacking: a deep learning ensemble to classify anxiety, depression, and their comorbidity from Reddit texts. IEEE J Biomed Health Inform. Jul 2022;26(7):3303-3311. [doi: 10.1109/JBHI.2022.3151589] [Medline: 35230959]
78. American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders (DSM-5-TR). Washington, DC. American Psychiatric Association; 2013.
79. Ireland M, Iserman M. Within and between-person differences in language used across anxiety support and neutral reddit communities. In: Proceedings of the 5th Workshop on Computational Linguistics and Clinical Psychology. 2018. Presented at: CLPsych '18; June 5, 2018:182-193; New Orleans, LA. URL: https://aclanthology.org/W18-0620.pdf [doi: 10.18653/v1/w18-0620]
80. Harrigian K, Aguirre C, Dredze M. Do models of mental health based on social media data generalize? In: Proceedings of the 2020 Findings of the Association for Computational Linguistics. 2020. Presented at: EMNLP '20; November 16-20, 2020:3774-3788; Virtual Event. URL: https://aclanthology.org/2020.findings-emnlp.337.pdf [doi: 10.18653/v1/2020.findings-emnlp.337]
81. The 2024 social media demographics guide. Khoros. URL: https://khoros.com/resources/social-media-demographics-guide [accessed 2024-04-29]
82. Hruska J, Maresova P. Use of social media platforms among adults in the United States—behavior on social media. Societies. Mar 23, 2020;10(1):27. [doi: 10.3390/soc10010027]

83. Amir S, Coppersmith G, Carvalho P, Silva M, Wallace B. Quantifying mental health from social media with neural user embeddings. arXiv. Preprint posted online April 30, 2017. [FREE Full text]
84. Wu S, Qiu Z. A RoBERTa-based model on measuring the severity of the signs of depression. In: Proceedings of the 2021 Conference and Labs of the Evaluation Forum. 2021. Presented at: CLEF '21; September 21-24, 2021:1-10; Bucharest, Romania. URL: https://ceur-ws.org/Vol-2936/paper-86.pdf
85. Inkpen D, Skaik R, Buddhitha P, Angelov D, Fredenburgh M. uOttawa at eRisk 2021: automatic filling of the Beck's depression inventory questionnaire using deep learning. In: Proceedings of the 2021 Conference and Labs of the Evaluation Forum. 2021. Presented at: CLEF '21; September 21-24, 2021:966-980; Bucharest, Romania. URL: https://ceur-ws.org/Vol-2936/paper-79.pdf
86. Reddit Self-reported Depression Diagnosis (RSDD) dataset. Georgetown IR Lab. 2023. URL: https://georgetown-ir-lab.github.io/emnlp17-depression/ [accessed 2024-04-05]
87. Resources. Georgetown IR Lab. 2023. URL: https://ir.cs.georgetown.edu/resources/ [accessed 2024-04-05]
88. Resources - SMHD. Georgetown IR Lab. 2023. URL: https://ir.cs.georgetown.edu/resources/smhd.html [accessed 2024-04-05]
89. Dredze M. CLPsych 2015 shared task evaluation. Johns Hopkins Whiting School of Engineering. URL: https://www.cs.jhu.edu/~mdredze/clpsych-2015-shared-task-evaluation/ [accessed 2024-04-05]
90. eRisk 2021 text research collection. Bibtex. URL: https://tec.citius.usc.es/ir/code/eRisk2021.html [accessed 2024-04-05]
91. What is topic modeling?: a beginner's guide. Levity. URL: https://levity.ai/blog/what-is-topic-modeling [accessed 2024-04-29]
92. Pennebaker J, Boyd R, Jordan K, Blackburn K. The development and psychometric properties of LIWC2015. The University of Texas at Austin. 2015. URL: https://repositories.lib.utexas.edu/server/api/core/bitstreams/b0d26dcf-2391-4701-88d0-3cf50ebee697/content [accessed 2024-04-29]
93. Hier DB, Brint SU. A neuro-ontology for the neurological examination. BMC Med Inform Decis Mak. Mar 04, 2020;20(1):47. [FREE Full text] [doi: 10.1186/s12911-020-1066-7] [Medline: 32131804]
94. Chen J, Tam D, Raffel C, Bansal M, Yang D. An empirical survey of data augmentation for limited data learning in nlp. Trans Assoc Comput Linguist. 2023;11:191-211. [doi: 10.1162/tacl_a_00542]
95. Liu B. Supervised learning. In: Liu B, editor. Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data. Cham, Switzerland. Springer; 2011.
96. Noble WS. What is a support vector machine? Nat Biotechnol. Dec 2006;24(12):1565-1567. [doi: 10.1038/nbt1206-1565] [Medline: 17160063]
97. Breiman L. Random forests. Mach Learn. 2001;45:5-32. [doi: 10.1023/A:1010933404324]
98. Zhong G, Wang L, Ling X, Dong J. An overview on data representation learning: from traditional feature learning to recent deep learning. J Fin Data Sci. Dec 2016;2(4):265-278. [doi: 10.1016/j.jfds.2017.05.001]
99. What is BERT and how is it used in AI? H2O.ai. 2023. URL: https://h2o.ai/wiki/bert/ [accessed 2024-02-18]
100. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2019. Presented at: NAACL-HLT '19; June 2-7, 2019:4171-4186; Minneapolis, Minnesota. URL: https://aclanthology.org/N19-1423.pdf
101. Lan Z, Chen M, Goodman S, Gimpel K, Sharma P, Soricut R. Albert: a lite bert for self-supervised learning of language representations. arXiv. Preprint posted online September 26, 2019. [FREE Full text]
102. Ji S, Zhang T, Ansari L, Fu J, Tiwari P, Cambria E. MentalBERT: publicly available pretrained language models for mental healthcare. In: Proceedings of the 13th Language Resources and Evaluation Conference. 2022. Presented at: LREC '22; June 20-25, 2022:7184-7190; Marseille, France. URL: https://aclanthology.org/2022.lrec-1.778.pdf
103. Jo A. What ChatGPT and generative AI mean for science. Nature. Feb 9, 2023;614(1):214-216. [FREE Full text]
104. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. Language models are few-shot learners. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. 2020. Presented at: NIPS '20; December 6-12, 2020:1877-1901; Vancouver, BC. URL: https://dl.acm.org/doi/abs/10.5555/3495724.3495883
105. Wu T, He S, Liu J, Sun S, Liu K, Han Q, et al. A brief overview of ChatGPT: the history, status quo and potential future development. IEEE/CAA J Autom Sin. May 2023;10(5):1122-1136. [doi: 10.1109/jas.2023.123618]
106. Monaghan TF, Rahman SN, Agudelo CW, Wein AJ, Lazar JM, Everaert K, et al. Foundational statistical principles in medical research: sensitivity, specificity, positive predictive value, and negative predictive value. Medicina (Kaunas). May 16, 2021;57(5):503. [FREE Full text] [doi: 10.3390/medicina57050503] [Medline: 34065637]
107. Shung KP. Accuracy, precision, recall or F1? Towards Data Science. 2018. URL: https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9 [accessed 2024-04-18]
108. Narkhede S. Understanding AUC - ROC curve. Towards Data Science. 2018. URL: https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5 [accessed 2024-08-18]
109. Kessler RC, Berglund P, Demler O, Jin R, Koretz D, Merikangas KR, et al. National Comorbidity Survey Replication. The epidemiology of major depressive disorder: results from the National Comorbidity Survey Replication (NCS-R). JAMA. Jun 18, 2003;289(23):3095-3105. [doi: 10.1001/jama.289.23.3095] [Medline: 12813115]
110. Fujita F, Diener E, Sandvik E. Gender differences in negative affect and well-being: the case for emotional intensity. J Pers Soc Psychol. Sep 1991;61(3):427-434. [doi: 10.1037//0022-3514.61.3.427] [Medline: 1941513]

111. Software and systems engineering "software testing" part 11: guidelines on the testing of AI-based systems. International Organization for Standardization. 2023. URL: https://www.iso.org/obp/ui/en/#iso:std:iso-iec:tr:29119:-11:ed-1:v1:en:term:3.1.31 [accessed 2024-04-12]
112. Shaikh S, Daudpota SM, Imran AS, Kastrati Z. Towards improved classification accuracy on highly imbalanced text dataset using deep neural language models. Appl Sci (Basel). Jan 19, 2021;11(2):869. [doi: 10.3390/app11020869]
113. Younes Y, Mathiak B. Handling class imbalance when detecting dataset mentions with pre-trained language models. In: Proceedings of the 5th International Conference on Natural Language and Speech Processing. 2022. Presented at: ICNLSP '22; December 16-17, 2022:79-88; Trento, Italy. URL: https://aclanthology.org/2022.icnlsp-1.9.pdf
114. Bayer M, Kaufhold MA, Buchhold B, Keller M, Dallmeyer J, Reuter C. Data augmentation in natural language processing: a novel text generation approach for long and short text classifiers. Int J Mach Learn Cybern. Apr 12, 2022;14(1):135-150. [FREE Full text] [doi: 10.1007/s13042-022-01553-3] [Medline: 35432623]
115. Li B, Hou Y, Che W. Data augmentation approaches in natural language processing: a survey. AI Open. 2022;3:71-90. [doi: 10.1016/j.aiopen.2022.03.001]
116. Santos WR, de Oliveira RL, Paraboni I. SetembroBR: a social media corpus for depression and anxiety disorder prediction. Lang Resour Eval. Jan 11, 2023;58(1):273-300. [doi: 10.1007/s10579-022-09633-0]
117. Santos W, Yoon S, Paraboni I. Mental health prediction from social media text using mixture of experts. IEEE Latin Am Trans. Jun 2023;21(6):723-729. [doi: 10.1109/tla.2023.10172137]
118. Andrews GJ. Co-creating health's lively, moving frontiers: brief observations on the facets and possibilities of non-representational theory. Health Place. Nov 2014;30:165-170. [doi: 10.1016/j.healthplace.2014.09.002] [Medline: 25282125]
119. Gur RE, Kohler CG, Ragland JD, Siegel SJ, Lesko K, Bilker WB, et al. Flat affect in schizophrenia: relation to emotion processing and neurocognitive measures. Schizophr Bull. Apr 2006;32(2):279-287. [FREE Full text] [doi: 10.1093/schbul/sbj041] [Medline: 16452608]
120. Rickelman BL. Anosognosia in individuals with schizophrenia: toward recovery of insight. Issues Ment Health Nurs. 2004;25(3):227-242. [doi: 10.1080/01612840490274741] [Medline: 14965844]
121. Schizophrenia. National Alliance on Mental Illness. 2023. URL: https://www.nami.org/About-Mental-Illness/Mental-Health-Conditions/Schizophrenia [accessed 2024-03-19]
122. Li A, Jiao D, Liu X, Zhu T. A comparison of the psycholinguistic styles of schizophrenia-related stigma and depression-related stigma on social media: content analysis. J Med Internet Res. Apr 21, 2020;22(4):e16470. [FREE Full text] [doi: 10.2196/16470] [Medline: 32314969]
123. Robinson P, Turk D, Jilka S, Cella M. Measuring attitudes towards mental health using social media: investigating stigma and trivialisation. Soc Psychiatry Psychiatr Epidemiol. Jan 2019;54(1):51-58. [FREE Full text] [doi: 10.1007/s00127-018-1571-5] [Medline: 30069754]
124. Meyer-Lindenberg A. The non-ergodic nature of mental health and psychiatric disorders: implications for biomarker and diagnostic research. World Psychiatry. Jun 2023;22(2):272-274. [FREE Full text] [doi: 10.1002/wps.21086] [Medline: 37159352]
125. Zivin K, Eisenberg D, Gollust SE, Golberstein E. Persistence of mental health problems and needs in a college student population. J Affect Disord. Oct 2009;117(3):180-185. [doi: 10.1016/j.jad.2009.01.001] [Medline: 19178949]
126. Eisenberg D, Golberstein E, Gollust SE. Help-seeking and access to mental health care in a university student population. Med Care. Jul 2007;45(7):594-601. [doi: 10.1097/MLR.0b013e31803bb4c1] [Medline: 17571007]
127. Burke M, Kraut R, Marlow C. Social capital on Facebook: differentiating uses and users. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. 2011. Presented at: CHI '11; May 7-12, 2011:571-580; Vancouver, BC. URL: https://dl.acm.org/doi/10.1145/1978942.1979023 [doi: 10.1145/1978942.1979023]
128. Bell CC. DSM-IV: diagnostic and statistical manual of mental disorders. JAMA. Sep 14, 1994;272(10):828-829. [doi: 10.1001/jama.1994.03520100096046]
129. Westrin A, Lam RW. Seasonal affective disorder: a clinical update. Ann Clin Psychiatry. 2007;19(4):239-246. [doi: 10.1080/10401230701653476] [Medline: 18058281]
130. Mohammad S, Kiritchenko S, Zhu X. NRC-Canada: building the state-of-the-art in sentiment analysis of tweets. In: Proceedings of the 2nd Joint Conference on Lexical and Computational Semantics (*SEM) & Proceedings of the 7th International Workshop on Semantic Evaluation. 2013. Presented at: SIGLEX '13; June 14-15, 2013:321-327; Atlanta, GA. URL: https://aclanthology.org/S13-2053.pdf [doi: 10.3115/v1/s14-2077]
131. Messinger JW, Trémeau F, Antonius D, Mendelsohn E, Prudent V, Stanford AD, et al. Avolition and expressive deficits capture negative symptom phenomenology: implications for DSM-5 and schizophrenia research. Clin Psychol Rev. Feb 2011;31(1):161-168. [FREE Full text] [doi: 10.1016/j.cpr.2010.09.002] [Medline: 20889248]
132. Bar-Haim Y, Lamy D, Pergamin L, Bakermans-Kranenburg MJ, van IJzendoorn MH. Threat-related attentional bias in anxious and nonanxious individuals: a meta-analytic study. Psychol Bull. Jan 2007;133(1):1-24. [doi: 10.1037/0033-2909.133.1.1] [Medline: 17201568]
133. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med. Sep 2001;16(9):606-613. [FREE Full text] [doi: 10.1046/j.1525-1497.2001.016009606.x] [Medline: 11556941]

134. Rickard N, Arjmand H, Bakker D, Seabrook E. Development of a mobile phone app to support self-monitoring of emotional well-being: a mental health digital innovation. JMIR Ment Health. Nov 23, 2016;3(4):e49. [FREE Full text] [doi: 10.2196/mental.6202] [Medline: 27881358]
135. Sykora MD, Jackson TW, O'Brien A, Elayan S. Emotive ontology: extracting fine-grained emotions from terse, informal messages. Int J Comput Sci Inf Sec. 2013;8(2):6. [FREE Full text]
136. Yang Z, Dai Z, Yang Y, Carbonell J, Salakhutdinov RR, Le QV. Xlnet: Generalized autoregressive pretraining for language understanding. arXiv. Preprint posted online June 19, 2019. [FREE Full text]
137. Jashinsky J, Burton SH, Hanson CL, West J, Giraud-Carrier C, Barnes MD, et al. Tracking suicide risk factors through Twitter in the US. Crisis. 2014;35(1):51-59. [doi: 10.1027/0227-5910/a000234] [Medline: 24121153]
138. Fodeh S, Goulet J, Brandt C, Hamada A. Leveraging Twitter to better identify suicide risk. Proc Mach Learn Res. 2017;69:1-7. [FREE Full text]
139. Hawton K, Sutton L, Haw C, Sinclair J, Deeks JJ. Schizophrenia and suicide: systematic review of risk factors. Br J Psychiatry. Jul 2005;187:9-20. [doi: 10.1192/bjp.187.1.9] [Medline: 15994566]
140. Sareen J, Cox BJ, Afifi TO, de Graaf R, Asmundson GJ, ten Have M, et al. Anxiety disorders and risk for suicidal ideation and suicide attempts: a population-based longitudinal study of adults. Arch Gen Psychiatry. Nov 2005;62(11):1249-1257. [doi: 10.1001/archpsyc.62.11.1249] [Medline: 16275812]
141. Golder S, Ahmed S, Norman G, Booth A. Attitudes toward the ethics of research using social media: a systematic review. J Med Internet Res. Jun 06, 2017;19(6):e195. [FREE Full text] [doi: 10.2196/jmir.7082] [Medline: 28588006]
142. Ford E, Curlewis K, Wongkoblap A, Curcin V. Public opinions on using social media content to identify users with depression and target mental health care advertising: mixed methods survey. JMIR Ment Health. Nov 13, 2019;6(11):e12942. [FREE Full text] [doi: 10.2196/12942] [Medline: 31719022]
143. Liu G, Wang C, Peng K, Huang H, Li Y, Cheng W. SocInf: membership inference attacks on social media health data with machine learning. IEEE Trans Comput Soc Syst. Oct 2019;6(5):907-921. [doi: 10.1109/tcss.2019.2916086]
144. Reidenberg JR, Breaux T, Cranor LF, French B, Grannis A, Graves J, et al. Disagreeable privacy policies: mismatches between meaning and users understanding. SSRN Journal. Preprint posted online March 31, 2014. [FREE Full text] [doi: 10.2139/ssrn.2418297]
145. Sleigh J. Experiences of donating personal data to mental health research: an explorative anthropological study. Biomed Inform Insights. Jun 27, 2018;10:1178222618785131. [FREE Full text] [doi: 10.1177/1178222618785131] [Medline: 30013355]
146. Beninger K, Fry A, Jago N, Lepps H, Nass L, Silvester H. Research using social media; users' views. NatCen Social Research. Feb 20, 2014. URL: https://www.researchgate.net/profile/Kelsey-Beninger/publication/261551701_Research_using_Social_Media_Users'_Views/links/0c96053497fed9ac11000000/Research-using-Social-Media-Users-Views.pdf [accessed 2024-02-18]
147. Haslam N. Dehumanization: an integrative review. Pers Soc Psychol Rev. 2006;10(3):252-264. [doi: 10.1207/s15327957pspr1003_4] [Medline: 16859440]
148. Kaplan K, Salzer MS, Solomon P, Brusilovskiy E, Cousounis P. Internet peer support for individuals with psychiatric disabilities: a randomized controlled trial. Soc Sci Med. Jan 2011;72(1):54-62. [doi: 10.1016/j.socscimed.2010.09.037] [Medline: 21112682]
149. Spanakis P, Wadman R, Walker L, Heron P, Mathers A, Baker J, et al. Measuring the digital divide among people with severe mental ill health using the essential digital skills framework. Perspect Public Health. Jan 2024;144(1):21-30. [FREE Full text] [doi: 10.1177/17579139221106399] [Medline: 35929589]
150. Eichstaedt JC, Smith RJ, Merchant RM, Ungar LH, Crutchley P, Preoţiuc-Pietro D, et al. Facebook language predicts depression in medical records. Proc Natl Acad Sci U S A. Oct 30, 2018;115(44):11203-11208. [FREE Full text] [doi: 10.1073/pnas.1802331115] [Medline: 30322910]
151. Woolway GE, Legge SE, Lynham A, Smart SE, Hubbard L, Daniel ER, et al. Assessing the validity of a self-reported clinical diagnosis of schizophrenia. medRxiv. Preprint posted online December 8, 2023. [FREE Full text] [doi: 10.1101/2023.12.06.23299622] [Medline: 38106032]
152. Cheng S, Chang C, Chang W, Wang H, Liang C, Kishimoto T, et al. The now and future of ChatGPT and GPT in psychiatry. Psychiatry Clin Neurosci. Nov 2023;77(11):592-596. [FREE Full text] [doi: 10.1111/pcn.13588] [Medline: 37612880]
153. Grabb D. The impact of prompt engineering in large language model performance: a psychiatric example. J Med Artif Intell. Oct 2023;6:20-25. [doi: 10.21037/jmai-23-71]
154. Xu X, Yao B, Dong Y, Gabriel S, Yu H, Hendler J, et al. Mental-llm: leveraging large language models for mental health prediction via online text data. Proc ACM Interact Mob Wearable Ubiquitous Technol. Mar 06, 2024;8(1):1-32. [doi: 10.1145/3643540]
155. Ghosh S, Ekbal A, Bhattacharyya P. What does your bio say? Inferring Twitter users' depression status from multimodal profile information using deep learning. IEEE Trans Comput Soc Syst. Oct 2022;9(5):1484-1494. [doi: 10.1109/tcss.2021.3116242]
156. Semwal N, Suri M, Chaudhary D, Gorton I, Kumar B. Multimodal analysis and modality fusion for detection of depression from Twitter data. In: Proceedings of the 39th Annual AAAI Conference on Artificial Intelligence. 2023. Presented at:

AAAI '23; February 25-March 4, 2023:1-5; Philadelphia, PA. URL: https://amulyayadav.github.io/AI4SG2023/images/31.pdf
157. Bucur AM, Cosma A, Rosso P, Dinu LP. It's just a matter of time: detecting depression with time-enriched multimodal transformers. In: Proceedings of the 45th European Conference on Information Retrieval on Advances in Information Retrieval. 2023. Presented at: ECIR '23; April 2-6, 2023:200-215; Dublin, Ireland. URL: https://link.springer.com/chapter/10.1007/978-3-031-28244-7_13 [doi: 10.1007/978-3-031-28244-7_13]
158. Zhang H, Wang H, Han S, Li W, Zhuang L. Detecting depression tendency with multimodal features. Comput Methods Programs Biomed. Oct 2023;240:107702. [doi: 10.1016/j.cmpb.2023.107702] [Medline: 37531689]
159. Khoo LS, Lim MK, Chong CY, McNaney R. Machine learning for multimodal mental health detection: a systematic review of passive sensing approaches. Sensors (Basel). Jan 06, 2024;24(2):348. [doi: 10.3390/s24020348] [Medline: 38257440]
160. van Driel II, Giachanou A, Pouwels JL, Boeschoten L, Beyens I, Valkenburg PM. Promises and pitfalls of social media data donations. Commun Methods Meas. Sep 12, 2022;16(4):266-282. [doi: 10.1080/19312458.2022.2109608]
161. Choi S. Privacy literacy on social media: its predictors and outcomes. Int J Hum Comput Interact. Apr 18, 2022;39(1):217-232. [doi: 10.1080/10447318.2022.2041892]
162. Orben A, Przybylski AK. The association between adolescent well-being and digital technology use. Nat Hum Behav. Feb 2019;3(2):173-182. [doi: 10.1038/s41562-018-0506-1] [Medline: 30944443]
163. Torous J, Bucci S, Bell IH, Kessing LV, Faurholt-Jepsen M, Whelan P, et al. The growing field of digital psychiatry: current evidence and the future of apps, social media, chatbots, and virtual reality. World Psychiatry. Oct 2021;20(3):318-335. [FREE Full text] [doi: 10.1002/wps.20883] [Medline: 34505369]
164. Low DM, Rumker L, Talkar T, Torous J, Cecchi G, Ghosh SS. Natural language processing reveals vulnerable mental health support groups and heightened health anxiety on reddit during COVID-19: observational study. J Med Internet Res. Oct 12, 2020;22(10):e22635. [FREE Full text] [doi: 10.2196/22635] [Medline: 32936777]
165. La Sala L, Pirkis J, Cooper C, Hill NT, Lamblin M, Rajaram G, et al. Acceptability and potential impact of the #chatsafe suicide postvention response among young people who have been exposed to suicide: pilot study. JMIR Hum Factors. May 19, 2023;10:e44535. [FREE Full text] [doi: 10.2196/44535] [Medline: 37204854]
166. Kasson E, Vázquez MM, Doroshenko C, Fitzsimmons-Craft EE, Wilfley DE, Taylor CB, et al. Exploring social media recruitment strategies and preliminary acceptability of an mHealth tool for teens with eating disorders. Int J Environ Res Public Health. Jul 28, 2021;18(15):7979. [FREE Full text] [doi: 10.3390/ijerph18157979] [Medline: 34360270]
167. Birnbaum ML, Ernala SK, Rizvi AF, Arenare E, R Van Meter A, De Choudhury M, et al. Detecting relapse in youth with psychotic disorders utilizing patient-generated and patient-contributed digital data from Facebook. NPJ Schizophr. Oct 07, 2019;5(1):17. [FREE Full text] [doi: 10.1038/s41537-019-0085-9] [Medline: 31591400]
168. Alowais SA, Alghamdi SS, Alsuhebany N, Alqahtani T, Alshaya AI, Almohareb SN, et al. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. BMC Med Educ. Sep 22, 2023;23(1):689. [FREE Full text] [doi: 10.1186/s12909-023-04698-z] [Medline: 37740191]
169. Singh V, Sarkar S, Gaur V, Grover S, Singh OP. Clinical practice guidelines on using artificial intelligence and gadgets for mental health and well-being. Indian J Psychiatry. Jan 2024;66(Suppl 2):S414-S419. [FREE Full text] [doi: 10.4103/indianjpsychiatry.indianjpsychiatry_926_23] [Medline: 38445270]
170. Lee EE, Torous J, De Choudhury M, Depp CA, Graham SA, Kim H, et al. Artificial intelligence for mental health care: clinical applications, barriers, facilitators, and artificial wisdom. Biol Psychiatry Cogn Neurosci Neuroimaging. Sep 2021;6(9):856-864. [FREE Full text] [doi: 10.1016/j.bpsc.2021.02.001] [Medline: 33571718]
171. Maddox TM, Rumsfeld JS, Payne PR. Questions for artificial intelligence in health care. JAMA. Jan 2019;321(1):31-32. [doi: 10.1001/jama.2018.18932] [Medline: 30535130]
172. Tiego J, Martin EA, DeYoung CG, Hagan K, Cooper SE, Pasion R, et al. HiTOP Neurobiological Foundations Work Group. Precision behavioral phenotyping as a strategy for uncovering the biological correlates of psychopathology. Nat Ment Health. May 2023;1(5):304-315. [FREE Full text] [doi: 10.1038/s44220-023-00057-5] [Medline: 37251494]
173. Home page. Weibo. 2023. URL: https://m.weibo.cn/ [accessed 2024-07-21]
174. Home page. VK. 2023. URL: https://vk.com/ [accessed 2024-07-21]
175. Just how harmful is social media? Our experts weigh-in. Columbia University Mailman School of Public Health. Sep 27, 2021. URL: https://www.publichealth.columbia.edu/news/just-how-harmful-social-media-our-experts-weigh [accessed 2024-08-31]
176. Kozyreva A, Lewandowsky S, Hertwig R. Citizens versus the internet: confronting digital challenges with cognitive tools. Psychol Sci Public Interest. Dec 2020;21(3):103-156. [FREE Full text] [doi: 10.1177/1529100620946707] [Medline: 33325331]
177. Peters U. What is the function of confirmation bias? Erkenn. Apr 20, 2020;87(3):1351-1376. [doi: 10.1007/s10670-020-00252-1]

Abbreviations

AI: artificial intelligence

ALBERT: A Lite Bidirectional Encoder Representations from Transformers
BDI-II: Beck Depression Inventory-II
BERT: Bidirectional Encoder Representations from Transformers
EHR: electronic health record
LDA: latent Dirichlet allocation
LIWC: Linguistic Inquiry and Word Count
LLM: large language model
LM: language model
ML: machine learning
NLP: natural language processing
RSDD: Reddit Self-reported Depression Diagnosis
SMHD: self-reported mental health diagnoses
SVM: support vector machine

Edited by G Eysenbach; submitted 23.04.24; peer-reviewed by Y Hua, KW Tay; comments to author 05.07.24; revised version received 08.09.24; accepted 01.10.24; published 15.11.24

Please cite as:
Owen D, Lynham AJ, Smart SE, Pardiñas AF, Camacho Collados J
AI for Analyzing Mental Health Disorders Among Social Media Users: Quarter-Century Narrative Review of Progress and Challenges
J Med Internet Res 2024;26:e59225
URL: https://www.jmir.org/2024/1/e59225
doi: 10.2196/59225
PMID: