Paper type: Review

Artificial Intelligence for Analyzing Mental Health Disorders in Social Media: A 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, UK
[2] Centre for Neuropsychiatric Genetics and Genomics, Division of Psychological Medicine and Clinical Neurosciences, School of Medicine, Cardiff University, UK

* These authors contributed equally

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 examines the state of the art in methods for predicting mental health statuses of social media users. Our focus is the development of AI-driven methods, particularly Natural Language Processing (NLP), for analyzing large volumes of written text. We also detail constraints affecting research in this area. These include the dearth of high-quality public data sets for methodological benchmarking and the need to adopt ethical and privacy frameworks acknowledging the stigma and vulnerability of those affected by mental illness.

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

Results: Larger data sets with precise dates of subjects' diagnoses are needed to support the development of methods for predicting mental health status, particularly in severe disorders such as schizophrenia. Inviting participants 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 NLP 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. For this to be achieved, more high-quality social media data sets need to be made available and privacy concerns regarding the use of this 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.

Keywords: mental health, depression, anxiety, schizophrenia, social media, natural language processing, narrative review

Introduction

The Global Burden of Disease study (1990-2019) reports that anxiety disorders, major depressive disorder (MDD), 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].

Experiences of those affected by these 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 the prediction of mental health status in its users have burgeoned [7,8,9]. Research has also spiked following the COVID-19 pandemic [10] and has become a truly interdisciplinary pursuit involving not only computer scientists, but 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" (see [12] for a discussion on the meaning of these terms in this literature). A branch of AI most apt 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 to 1966. It purported to perform the role of 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]. 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 mental health status of social media users. Indeed, the huge volume of human language content available online, 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 mental health statuses of members of online communities were done without AI, through manual review of postings and classic statistical analyses. For example, during November 1999, psychiatrists monitored the general psychiatry sub-forum 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, affected users often sought social support in their local communities. This corroborated previous findings showing that online 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 affected users 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 MDD (“depression” hereafter) can be challenging to identify, particularly in older adolescents. So, between 2009 and 2012 they sought Facebook profiles of freshman students whose status updates referenced depression symptoms. Such students were then contacted and, where 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 to seeking help in person. Thus, it was recognized that Facebook depression disclosures could be harnessed to identify those who might have unmet needs of mental health care. This provided an explicit motivation for improving the methods for predicting this disorder early in its course.

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 focus on methods that concern the detection of language features presented in the texts of user social media postings. A main aim of our review is to ascertain state-of-the-art methodologies for detecting linguistic features that can be attributed to mental illnesses. This includes cataloging data sets 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 (EHR), clinical questionnaires, or self-disclosure statements of a

mental health diagnosis (eg, “I was diagnosed with depression”), for example. We then examine how these methodologies integrate the temporal stochasticity of mental states as reflected by longitudinal studies. We also identify 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

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 in academia [25,26,27]. It offers particularly extensive coverage of computer science and informatics, which is the primary discipline of the literature that forms this review, outperforming the likes of Scopus [28]. Our search aimed to retrieve literature covering the three 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 one of four different subject areas that cover important and distinct aspects around mental health research in social media: Data sets on Social Media and Mental Health, Methods for Predicting Mental Health Status, Longitudinal Analyses on Mental Health, and Ethical Aspects on 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.

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

Data sets on Social Media and Mental Health
To develop methods for predicting mental health status or conducting longitudinal analyses, carefully constructed social media data sets are required. We identify publicly available data sets 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 NLP technologies.

Longitudinal Analyses on 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 on 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 amongst the general population and how they may be ameliorated.

We selected the most pertinent articles that also covered a broad time span. A detailed exposition of the literature search and selection strategy, which is informed by Ferrari [29], can be found in Multimedia Appendix 1.

Results

Overview

Following the four-stage manual sifting exercise, 35 articles across the four 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 four sections, which cover each of our four subject areas.

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

	Number	Article title	Year published	Study population	Summary
Subject area: Data sets on Social Media and Mental Health					
	1	Predicting Depression via Social Media [31]	2013	Posts of 476 Twitter users who self-report a diagnosis of depression between September 2011 and June 2012	Development of methods for data set construction via crowdsourcing and quantifying subjects' depressive language use during the year before their diagnosis
	2	Quantifying Mental Health Signals in Twitter [32]	2014	Posts published between 2008 and 2013 of 6696 Twitter users with a self-stated	Development and evaluation of a method for swift and inexpensive capture of data about a range of mental illnesses

				diagnosis of a mental health disorder: 394 with bipolar disorder 441 with depression 244 with post-traumatic stress disorder (PTSD) 159 with seasonal affective disorder 5728 controls	
	3	Depression and Self-Harm Risk Assessment in Online Forums [33]	2017	Posts published between January 2006 and October 2016 of 9210 Reddit users with a self-stated diagnosis of depression and 107274 controls	Development and evaluation of a method for recognizing depressed users from their language use alone
	4	RSDD-Time: Temporal Annotation of Self-Reported Mental Health Diagnoses [34]	2018	Self-reported depression diagnosis posts of 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
	5	SMHD: A Large-Scale Resource for Exploring Online	2018	Posts published between January 2006 and December 2017 of 385476	Development of methods for recognizing self-reported mental health condition diagnoses and obtaining high-quality labeled data

		Language Usage for Multiple Mental Health Conditions [35]		Reddit users with a self-stated diagnosis of a mental health disorder: 10098 users with attention deficit hyperactivity disorder (ADHD) 8783 users with anxiety 2911 users with autism 6434 users with bipolar disorder 14139 users with depression 598 users with eating disorders 2336 with Obsessive Compulsive Disorder (OCD) 2894 with PTSD 1331 with Schizophrenia 335952 controls	automatically, rather than manually
	6	Mental Health Surveillance over Social Media with Digital Cohorts [36]	2019	Randomly selected posts belonging to 48000 United States Twitter users	Development of methods for automatically inferring characteristics including gender, ethnicity, and location of randomly collected Twitter users

	7	Overview of eRisk at CLEF 2021: Early Risk Prediction on the Internet (Extended Overview [37]	2021	Posts published between November 2009 and October 2020 of 80 Reddit users who completed a Beck Depression Inventory-II (BDI-II) questionnaire	Development of methods for determining the severity of depression in Reddit users
Subject area: Methods for Predicting Mental Health Status					
	8	Social Media as a Measurement Tool of Depression in Populations [38]	2013	Posts of 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
	9	Beyond LDA: Exploring Supervised Topic Modeling for Depression-Related	2015	Posts of approximately 2000 Twitter users, of whom approximately 600 self-identify as having been	Investigation into the use of topic models to analyze linguistic signals for detecting depression

		Language in Twitter [39]		clinically diagnosed with depression	
	10	Quantifying the Language of Schizophrenia in Social Media [40]	2015	Posts published between 2008 and 2015 of 174 Twitter users who self-report a diagnosis of schizophrenia	Development of methods for analyzing how the language of schizophrenia can aid in identifying and getting help to people suffering from schizophrenia
	11	Recognizing Depression from Twitter Activity [41]	2015	(Center for Epidemiologic Studies Depression Scale) CES-D Questionnaire responses and posts of 209 Twitter users	Development of methods for extracting and using features from the activity histories of Twitter users to estimate the presence of depression
	12	A Collaborative Approach to Identifying Social Media Markers of Schizophrenia by Employing Machine Learning and Clinical Appraisals [42]	2017	Posts published between 2012 and 2016 of 146 Twitter users who self-disclose 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
	13	Detecting depression and mental illness on social media: an integrative review [43]	2017	43 peer-reviewed articles	Literature review of methods for predicting mental illness using social media

	14	Forecasting the onset and course of mental illness with Twitter data [44]	2017	Posts of 105 Twitter users who have a diagnosis of depression and 99 controls. Also, posts of 174 Twitter users who have a diagnosis of PTSD.	Development of models to predict the emergence of depression in Twitter users
	15	A text classification framework for simple and effective early depression detection over social media streams [45]	2019	Posts of 135 Reddit users who have depression and 752 controls	Development of a text classification approach for early risk detection with respect to social media users with depression, with an emphasis on explainable AI
	16	Towards Preemptive Detection of Depression and Anxiety in Twitter [46]	2020	Posts of 548 Twitter users who self-disclose having either depression or anxiety and 4650 controls	Development of a Language Model-based (LM) approach for early detection of depression in Twitter users
	17	A Transformers Approach to Detect Depression in Social Media [47]	2021	Posts of 4000 Reddit users who self-disclose having depression and 4000 controls	Development of transformer-based models for detecting depression in social media users
	18	Characterisation of Mental Health Conditions in Social Media	2022	77 peer-reviewed articles	Literature review of research concerning deep learning (DL) techniques for identifying various mental health

		Using Deep Learning Techniques [48]			conditions from social media data
	19	Utilizing ChatGPT Generated Data to Retrieve Depression Symptoms from Social Media [49]	2023	Posts from 3107 Reddit users	Development of methods for generating synthetic social media data for subsequent use in transformer-based language model depression detection
	20	Prompt-based mental health screening from social media text [50]	2024	Posts of 1684 Twitter users who self-report a diagnosis of depression and 11788 controls	Development of methods that use Large Language Model (LLM) prompting as an aid to mental health screening in social media text
Subject area: Longitudinal Analyses on Mental Health					
	21	Feeling bad on Facebook: Depression disclosures by college students on a social networking site [51]	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
	22	Towards Assessing Changes in Degree of Depression	2014	Status updates and survey responses of 28749 Facebook users collected	Development of a regression model to predict users' degrees of depression based on their Facebook status updates

		through Facebook [52]		between June 2009 and March 2011	
	23	Small but Mighty: Affective Micropatterns for Quantifying Mental Health from Social Media Language [53]	2017	Posts of 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 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
	24	Monitoring Online Discussions About Suicide Among Twitter Users With Schizophrenia: Exploratory Study [54]	2018	Posts of 203 Twitter users who self-identify as having schizophrenia and 173 controls	An exploration of the feasibility of monitoring online discussions about suicide among Twitter users who self-identify as having schizophrenia
	25	Predicting Depression From Language-Based Emotion Dynamics: Longitudinal Analysis of	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

		Facebook and Twitter Status Updates [55]			
	26	What about Mood Swings: Identifying Depression on Twitter with Temporal Measures of Emotions [56]	2018	Posts of 585 Twitter users who self-report a diagnosis of depression and 6596 controls	Development of a method for identifying users with or at risk of depression by incorporating measures of eight emotions as features from Twitter posts over time, including a temporal analysis of these features
	27	Monitoring Depression Trends on Twitter During the COVID-19 Pandemic: Observational Study [57]	2021	Posts of 2575 Twitter users who self-disclose a diagnosis of depression and 2575 controls	Development of transformer-based DL language models to identify depression users from their everyday language and to monitor the fluctuation of their depression levels
	28	Using language in social media posts to study the network dynamics of depression longitudinally [58]	2022	Posts of 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
	29	Enabling Early Health Care Intervention by Detecting Depression in Users of Web-Based Forums	2023	Posts of 56 Reddit users who self-report a diagnosis of depression and 168 controls	An investigation to determine the time points during a depressed person's posting history that are most indicative of their depression

		using Language Models: Longitudinal Analysis and Evaluation [59]			
Subject area: Ethical Aspects on the Data and Analysis of Mental Health					
	30	Effectiveness of Social Media Interventions for People With Schizophrenia: A Systematic Review and Meta-Analysis [60]	2016	2 peer-reviewed publications	Literature review of the effectiveness of social media interventions for supporting people with schizophrenia
	31	Ethical issues in using Twitter for population-level depression monitoring: a qualitative study [61]	2016	16 Twitter users with a self-reported diagnosis of depression participating in a series of focus groups and 10 controls	Cross-sectional survey study of public attitudes towards using Twitter data for mental health monitoring
	32	Social media, big data, and mental	2016	62 peer-reviewed articles	Literature review of work that uses social media "big data", Natural Language Processing

		health: current advances and ethical implications [62]			(NLP), and Machine Learning (ML) for mental health surveillance and the ethical considerations therein
	33	Who is the "Human" in Human-Centered Machine Learning: The Case of Predicting Mental Health from Social Media [63]	2019	55 peer-reviewed articles	Literature review of how scientific articles represent human research subjects in human-centered machine learning
	34	Ethics and Privacy in Social Media Research for Mental Health [64]	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 usage of such data
	35	Understanding the Role of Social Media–Based Mental Health Support Among College Students: Survey and Semistructured Interviews [65]	2021	101 US university students aged 18 to 24	Web-based survey followed by semi-structured 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

Data sets on Social Media and Mental Health

To develop methods for predicting mental health status, access to high quality data sets 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 timestamped insights to their psychological state. To this end, they used crowdsourcing to compile a data set of tweets belonging to 476 Twitter users who self-reported a diagnosis of depression. The data was subsequently used to analyze linguistic and behavioral patterns, such as symptom mentions and diurnal activity, respectively. While the data was deemed high quality by Coppersmith et al [32], they pointed to its limited size and scope 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. They therefore proposed an automated method for labeled data set construction, which sought self-reports of mental illness diagnoses in Twitter such as "I was diagnosed with depression.". Their yield of over 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. Firstly, only Twitter users were captured, a sample not likely representative of the general population but in this sense, similar to other social media data sets. Secondly, it was not possible to verify that the self-stated diagnoses were genuine or capture the same psychopathology as clinical diagnoses. For example, population biobank data has 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 [66]. Nevertheless, this approach has ostensibly provided the foundation for several publicly available and widely used mental health data sets.

Yates et al developed the Reddit Self-reported Depression Diagnosis (RSDD) data set [33], which contains the posting histories of 9210 users with a diagnosis of depression revealed by self-report statements, like that described above. Further populated with 107274 non-depressed users for control purposes, RSDD has become an oft-used resource in the development of methods for predicting depression [67-72]. It has also propagated the development of two sister data sets, Temporal Annotation of Self-Reported Mental Health Diagnoses (RSDD-Time) [34] and Large-Scale Resource for Exploring Online Language Usage for Multiple Mental Health Conditions (SMHD) [35]. The former was conceived by MacAvaney et al after recognition that research had, largely, not examined the temporality of mental health diagnoses. They randomly selected 598 posts from RSDD that contained the self-reported diagnosis statement of a depressed user and manually annotated them to denote when the diagnosis occurred. Owen et al 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 [59]. However, because many of the annotations in RSDD-Time denote that the diagnosis dates of many of the users cannot not 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 72 depressed users. This highlights a need for much larger data sets where the dates of depression diagnoses are denoted to a high degree of accuracy.

SMHD, meanwhile, was born out of a desire for data sets covering a broad range of mental health disorders. It provided a platform for development of methods concerning not only depression [73,74], but also suicidal ideation [75], schizophrenia [76], and even multi-class experimental setups involving combinations of anxiety, eating disorders, ADHD, bipolar disorder, and PTSD [77-80]. It was also intended that a wider range of higher positive predictive value patterns be used to collect a greater volume of diagnosed users. Such patterns detect diagnosis keywords relevant to each disorder, drawn from the Diagnostic and Statistical Manual of Mental Disorders (DSM-5) [81]. As a result, SMHD contains 20406 diagnosed users and 335952 matched controls. Despite these strengths, RSDD and SMHD are limited 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 [82]. 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 [83].

Other biases also exist in social media data. For example, most social media platforms, including Facebook, Twitter, and Instagram, have more male users [84]. There is also evidence to suggest that people with higher levels of education and household income are more frequent social media users [85]. To address such biases and improve the representativeness of social media data sets, Amir et al considered a cohort-based approach to data set construction [36]. 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 [86] to ascertain the prevalence of depression and PTSD across the 48000 users collected. This is as opposed 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, though they also acknowledged that proper evaluation of these patterns would require disentangling the ways in which social media data sets differ from representative samples of the underlying population. In any case, further development and adoption of surveillance-based methods is 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, health status). We explore these matters in more depth in the Ethical Aspects on the Data and Analysis of Mental Health section.

Finally in this section, we mention the work of the eRisk Lab [37], which touches upon another important dimension in the support of methods for predicting mental health status. Their 2021 data set, which comprises Reddit posting histories belonging to 80 users, is accompanied by ground truth data that can aid in the development of methods for gauging the severity of depression. Recorded against each user is a completed BDI-II

questionnaire, which categorizes the severity of their depression (ranging from minimal to severe). While the data set 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. Since the state of one's mental health is somewhat fluid [34] the data set may contain users whose depression may have long passed. This is plausible given that one user in the data set has a posting history spanning more than ten years, although it should be noted that this is an anomaly, with the median posting history in the data set being just over one year. Furthermore, the data set's small size in terms of number of users is a major constraint [87,88]. This highlights the difficulty in obtaining copious ground truth data that is traditionally collected via confidential questionnaires.

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

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

Data set	Platform	Contents	Compiled	Acquisition Enquiries	Article
Reddit Self-reported Depression Diagnosis (RSDD)	Reddit	116484 users: 9210 with depression 107274 controls	2014	Reddit Self-reported Depression Diagnosis (RSDD) data set [89]	Depression and Self-Harm Risk Assessment in Online Forums [33]
Reported Mental Health Diagnoses (RSDD-Time)	Reddit	598 users with depression	2018	ir@Georgetown – Resources [90]	RSDD-Time: Temporal Annotation of Self-Reported Mental Health Diagnoses [34]
Large-Scale Resource for Exploring Online Language Usage for Multiple Mental Health Conditions (SMHD)	Reddit	385476 users: 10098 with ADHD 8783 with anxiety	2018	ir@Georgetown – Resources – SMHD [91]	SMHD: A Large-Scale Resource for Exploring Online

		2911 with autism 6434 with bipolar disorder 14139 with depression 598 with eating disorders 2336 with OCD 2894 with PTSD 1331 with Schizophrenia 335952 controls			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 873 controls	2015	CLPsych 2015 Shared Task Evaluation [92]	Mental Health Surveillance over Social Media with Digital Cohorts [36]
eRisk 2021 Text Research Collection	Reddit	80 users who completed a BDI-II questionnaire	2021	eRisk 2021 Text Research Collection [93]	Overview of eRisk at CLEF 2021: Early Risk Prediction on the Internet (Extended Overview) [37]

Methods for Predicting Mental Health Status

Background

The methods covered in this review are supported by ML. Since there is broad terminology concerning ML, we introduce the relevant terms in Table 3.

Table 3. ML terms used in this review.

	Term	Description

Data representation		
	Latent Dirichlet Allocation (LDA) [94]	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.
	Linguistic Inquiry and Word Count (LIWC) [95]	A text analysis technique that can infer the emotion conveyed in text (eg, positive or negative).
	Ontology [96]	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 [97]	Methods used to increase the size of a data set by adding slightly modified copies of existing items in the data set.
Algorithms		
	Supervised learning [98]	A type of ML algorithm analogous to human learning from past experiences to gain new knowledge to improve our ability to perform real-world tasks.
	Support Vector Machine (SVM) [99]	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 [100]	A supervised ML algorithm that combines the output of multiple decision trees to reach a single result.
	Deep learning (DL) [101]	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.
Pre-trained models		
	Language Model (LM) [102]	An LM is a probability distribution over words or word sequences. They learn to predict text that might come before and after other text, thus are used in tasks such as predicting text when writing an email.
	Bidirectional Encoder Representations from Transformers (BERT) [103]	An LM that examines words within text by considering both left-to-right and right-to-left contexts.
	A Lite BERT (ALBERT) [104]	A lightweight alternative to BERT that is suitable for use where less computing power is available.
	MentalBERT [105]	An LM designed specifically to aid NLP tasks in the mental health care research community.
	MentalRoBERTa [106]	An alternative to MentalBERT that can perform predictions in longer left-to-right and right-to-left contexts.
	Large Language Model (LLM) [107]	Large-scale LM designed for NLP tasks such as producing complex text.
	Generative Pre-trained Transformers (GPT) [108]	A family of neural network (in that they mimic the workings of the human brain) models that support AI-driven applications for creating content such as text, images, or sound.
	ChatGPT [109]	A chatting robot that can provide a detailed response to a question or instruction.
Performance metrics		
	Positive Predictive Value [101]	Of the instances in a data set 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 [110]	Of the instances in a data set with a particular label, sensitivity denotes how many of them were predicted correctly by an ML algorithm. Sensitivity is also known as recall.
	F1 [111]	The harmonic mean of positive predictive value and sensitivity.
	Area under the receiver operating characteristics (AUROC) [112]	Denotes an ML algorithm's performance in terms of distinguishing between labels.

Traditional Machine Learning Approaches

In 2013, methods for predicting mental health status from social media data began to emerge [12] and have often involved inter-disciplinary teams of computer scientists and clinical psychologists. De Choudhury et al [38] were proponents of supervised learning methods for predicting depression amongst populations. Exploiting post-level and user-level features from a crowdsourced Twitter data set, they developed the social media depression index. To do this they used an 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 in social media than men, which marginally exceeded findings from epidemiological surveys on formal diagnoses that suggest the figure is 1.3 [113]. The overestimation was linked to the greater emotional expressivity of women [114], positing that methods more sensitive to language use could help develop more robust models. Such methods include topic modeling via LDA. While this approach has also been used for predicting depression in Twitter users [32], its results have to be taken cautiously as its data set, in terms of depressed and control users, was not deemed a representative sample of the population [39]. Later work used LDA-derived features as input to an SVM classifier to discern between depressed and control users in Twitter [41]. 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 [44], another traditional ML algorithm, Random Forests, was deployed using LIWC features derived from post text. A commendable AUROC score of 87% was achieved and the method 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 [41] acknowledged that emerging DL algorithms could well advance methods in this area and were likely to inform future work. We explore these in

the next subsection. 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 [43]. It also cited two studies that were influential in data set design methods [31,32] that we discussed in the previous section as being likely to help realize this.

By 2019, interest in methods for early prediction of depression had developed due to recognition that they could help people receive the health care and social support they need sooner than they otherwise might [45]. Burdisso et al 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 is to provide explainability [115] for its classification decisions. It can 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 cannot be gleaned from traditional "black box" algorithms like SVM. SS3 was also hailed as a low-resource method, since, unlike SVM, it does not necessarily need to process the entire input text before returning its classification decision. However, it was acknowledged that since it examines each word of the input text in a singleton fashion, it would not consider potentially crucial two-word phrases such as "kill myself" in a classification decision.

Language Models and Transformers

The capabilities of language models had become well understood in NLP by the start of the 2020s. So, further to Burdisso et al's work, BERT and ALBERT were deployed in an early depression prediction task involving tweets that denoted that of a user with depression or anxiety, or with neither disorder [46]. Since 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. In an experimental setting where depressed and control users were balanced, F1 of 77% was achieved using BERT, compared to an SVM baseline of 65%. In an imbalanced data set however, which is a more accurate representation of real-world scenarios where these tools could be applied, BERT achieved 74% compared to SVM's 75%. Malviya et al performed a similar experiment where individual posts in a Reddit data set would be classified as depressed or non-depressed by BERT and traditional baseline algorithms [47]. 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 [116] and re-sampling [117]. A review of DL approaches to mental health prediction [48] that post-dates both studies [46,47] echoed

the need for further work involving much larger data sets while acknowledging the impact of existing data sets that we have already highlighted [33,35].

Some of the most recent methods have harnessed generative AI, principally using GPT [108]. The arrival of generative AI has enhanced opportunities in this domain. We have already noted that use of quality data is crucial in the pursuit of methods for predicting mental health status. Such data is often scarce and has given rise to data augmentation techniques [97,118]. A slightly different approach involves synthesizing data derived from existing data [119]. 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 [49]. 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 [106], 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 [50] of the SetembroBR Twitter corpus of depressed and control users [120]. The GPT was prompted to label each tweet as having either high, medium, or low relevance to mental health. The labeled data set was then used as input to a bag-of-words classifier and its prediction performance compared with that of a BERT-derived baseline produced by an earlier study [121]. While this approach was markedly low resource and bettered 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, 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, LDA was applied in a Twitter data set with the goal of distinguishing between users with this disorder and controls [40]. Key findings were that irrealis mood (denoted by use of uncertain terms such as "think" or "believe") [122] and flat affect (due to lack of emoticon use) [123], were prevalent in the postings of people with schizophrenia. A limitation of their data set 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 [124,125]. In any case, people with schizophrenia may be reluctant to disclose their diagnoses on social media since they are likely to receive stigmatized responses [126,127]. Birnbaum et al [42] attempted more

accurate identification in 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 electronic health records.

Longitudinal Analyses on 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 [128]. Therefore, research has also examined temporal profiles of mental health disorders and symptoms. A 2011 study considered US college student Facebook status updates and their potential for exhibiting content that may reveal symptoms of depression [51]. It was noted that opportunities for recognition and treatment of depression were being missed, particularly among college students [129,130]. Therefore Facebook, a social media platform that had become well-established amongst the student population [131], presented innovative opportunities to identify college students at risk. A manual exercise saw the collection of Facebook status updates of 200 students that spanned one year. Human annotators then scrutinized each post, denoting a depressive symptom if deemed present according to Diagnostic and Statistical Manual of Mental Disorders criteria [132]. A quarter of profiles exhibited at least one depressive symptom (as inferred through use of terms like "hopeless" or "giving up"). This evidence that Facebook may allow identification of at-risk students would be a precursor to future longitudinal analyses.

Schwartz et al [52] sought to gauge how depression changes among Facebook users during a calendar year. Their method involved extraction of 1-to-3-word terms, LDA-derived topics, and LIWC categories from the status updates of over 28000 users. A regression model was developed that indicated a significantly higher degree of depression among users during winter months than summer months, which is compatible with observations made in the psychiatry literature [133]. A baseline model that considered only average sentiment across each user's status updates was outperformed in terms of accuracy almost threefold, although the optimal model only exceeded 30% [134]. By comparison, Loveys et al conducted experiments predicting mental health statuses during much shorter time spans, hours in fact [53]. Tweets belonging to over 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 three subsequent tweets that occur within any three-hour window were observed. These observations were dubbed "micropatterns". It was noted that users with schizophrenia were less likely than control users to show emotional variability between tweets, which perhaps demonstrates a deficit in affective expression, a known

schizophrenia symptom [135]. Users with anxiety were less likely than controls to make consecutive positive tweets, again consistent with psychological findings [136]. However, the micropatterns did not contain sufficient detail 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 online postings have also been studied. Seabrook et al considered whether "emotion dynamics" in Twitter and Facebook may provide early indicators for depression risk [55]. The feasibility of using emotion variability and instability as an indicator of depression severity, measured by the Patient Health Questionnaire-9 (PHQ-9) [137], 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 [138] would gauge the emotion of their status updates and the severity of depression (via PHQ-9) over a one-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, and greater variability of negative emotion expression on Twitter may in fact be protective for mental health. These observations were constrained, however, 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 in Twitter for their use in predicting depression [56]. Eight basic emotions (anger, disgust, fear, happiness, sadness, surprise, shame, and confusion) were sought in the tweets of 585 depressed users across a four-month period. The average intensity of each emotion was calculated via the EMOTIVE ontology [139] 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 of an individual's emotions over time show potential in identifying users with depression. Fine-grained consideration of language used in tweets, such as tentative (eg, "maybe") and temporal related terms, may not only predict its presence, but its severity too [58].

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 3rd and May 22nd, 2020, were collected regarding users who self-disclosed having depression [57]. 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 [140] and a geographical aggregation of users in the data set, they demonstrated how depression levels fluctuated between the above dates in New York,

California, Florida, and the US as a whole. It was observed that depression levels of all four geographical areas were similar during the pandemic, with a steady increase after announcement of the US National Emergency on March 13th, a modest decrease after April 23rd, followed by a steep increase after May 10th. The overall depression score of Florida was substantially lower than the US average and the other two 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 [59] aimed to determine how far in advance of a Reddit user's depression diagnosis that their postings were most indicative of their condition. 56 depressed users and 168 controls were acquired from an intersection of the RSDD [33] and RSDD-Time data sets [34]. BERT and a specialist LM, MentalBERT [105], considered all user posts in increasingly large temporal bands up to 24 weeks (approximately six months) before the depressed users' diagnosis dates. The LMs achieved F1-scores of 0.726 and 0.715, respectively, when 12 weeks of postings were considered, suggesting therefore that the most poignant language used by depressed users occurs in the final 3 months before their eventual diagnosis. The reason for the specialist LM performing less well 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 Data sets 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 investigated the language employed by Twitter users with schizophrenia to observe if it would help assess suicide risk [54]. 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 [141,142]. Crucially, the time-of-day of each tweet was recorded. A logistic regression model predicted that 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). Taking the times of tweets into consideration, the frequency of conversations about suicide on Twitter correlated significantly with discussions about depression and anxiety, another trend that is consistent with established data [143,144]. However, like studies discussed

previously [40,42], the inability to be able to verify the diagnoses of the schizophrenia users was cited as a main limitation.

Ethical Aspects on the Data and Analysis of Mental Health

When constructing data sets, developing methods, and performing longitudinal analyses to aid mental health prediction, people's privacy ought to be considered. In 2016, Mikal et al [61] sought to determine attitudes of Twitter users towards 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 were 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 user identities be concealed. Also noted were concerns about the reliability of methods that use crude keyword searches and the misleading findings they could yield. An incorrect labeling of depression for a user whose identity is revealed would be considered stigmatizing, according to participants. The study was only indicative since 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 gleaned further evidence of fears regarding such stigmatization [62].

Nicholas et al [64] address similar privacy matters. They note that the introduction of 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 fear research findings may affect credit card applications [145], employment prospects, and attract stigma [146]. Fears are compounded by evidence that deidentified data can be re-identified using materials published alongside research articles [147]. Indeed, the desire for anonymity appears particularly widely held, which echoes Mikal et al and is reinforced by Vornholt and De Choudhury [65]. Therefore, obtaining explicit user consent for use of their data is considered crucial. A possible route is via acceptance of social media platform terms and conditions. However, since these may not be read and understood [148] this may not constitute informed consent. One solution is to explicitly invite participants to donate their social media data for research purposes [149]. Another proposal is a feature that enables users to opt in or out of their data being used as they post it [150].

A matter has also been identified regarding the terminology used in this area of mental health research. Chancellor et al reviewed how human subjects are referred to in literature for predicting mental health status using social media data [63]. Common traits were seen across 55 articles. For example, introductions often refer to human subjects 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 [151]. This may cause individuals and communities to become stigmatized, echoing the findings of studies discussed above. To alleviate this, it is suggested that more human-centered methods like participatory design should be considered where interviews and field studies are conducted. However, this is at odds with the challenges highlighted in the Data sets on Social Media and Mental Health section where acquiring sizable data sets through such methods is largely intractable [32].

With respect to schizophrenia, Välimäki et al determined via review that the perceptions and risks of social media interventions are largely unexplored [60]. There are however suggestions that some clinicians fear that use of online peer support without professional moderation may cause anxiety in the bearer of the disorder [152]. Cognitive deficits in people with schizophrenia can inhibit the development of digital skills [153], 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 remotely monitoring individual and population level mental states grew. Indeed, the search strategy followed for this review yielded more articles in the years 2020-2021 than in the previous 20 years; 917 and 903, respectively (see Multimedia Appendix 2). 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 subjects involved.

Our review has shown depression to be the most common condition reported in publicly available data sets, but 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 [59]. Obtaining such ground truth data via traditional confidential questionnaires is time-consuming and intrusive from the subject's point of view [138]. A solution may involve obtaining consented access to EHRs to accompany the users' social media postings, as piloted by Eichstaedt et al [154]. Indeed, this means of verification is in fact crucial in studies that consider schizophrenia because diagnosis self-disclosure statements, although having high sensitivity [155], may lack specificity [124,125]. In any case, social media data obtained also needs to be broadened to better support NLP methods. For example, Reddit data sets should routinely include postings from mental health subreddits in addition to other subreddits [82]. This would help ensure that LMs

pre-trained on such data are less prone to biases that may dampen the effectiveness of methods developed thereon [83]. 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 [49]. It was suggested that improved prompt manipulation is needed to produce synthesized data that is less stilted. Another role for LLMs may be in the automatic labeling of mental health data set instances. Santos and Paraboni produced evidence that an LLM (GPT-3.5) can perform promisingly (72% sensitivity) when distinguishing between tweets of users that may have depression and users those that likely do not [50]. LLMs may eventually offer a far less costly alternative to data set 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 [156,157]. Instruction fine-tuning is one such proposal for improving LLM performance. LLMs including GPTs are trained on very large, non-domain specific data sets such as Wikipedia. But further training an LLM on smaller, domain-specific data sets may enhance its performance in that domain. For example, when comparing the performance of a non-finetuned LLM and its finetuned counterpart, Xu et al measured a 23.4% increase in accuracy across six different mental health prediction tasks involving Reddit data [158]. However, finetuning 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 emotion conveyed in social media posts to be an underrepresented topic of research in this area. Consideration of finer-grained language features may also help better predict depression severity over time [58]. In fact, the most promising approaches will probably involve those that augment NLP; multimodal methods that consider non-text 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 [159] attempted to distinguish between depressed and non-depressed users by considering their profile images and the text of their profile description. A classifier that used features from the profile image outperformed a baseline classifier that used only features from the profile description by almost 10% in terms of F1. While profile images may therefore 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. 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 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 [160]. 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 F1. The conclusion therefore was that images seem to contain significant information regarding a user's mental health status, thus motivating further study in mental health status prediction. In Reddit meanwhile, multimodal methods may involve time-aware consideration of user posts. One study considered the relative time between posts as a feature for distinguishing between depressed and non-depressed Reddit and Twitter users [161]. Obtaining F1 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 with the often much lengthier Reddit posts lends to users posting more frequently on Twitter. A further study considered a multimodal approach with emphasis on emoji, again in the task of distinguishing between depressed and non-depressed users in both Twitter and Reddit [162]. With F1 scores of 0.80 and 0.95 being achieved for Twitter and Reddit, respectively, it could be concluded, given the two 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 have 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 [61,62]. Multimodal approaches that more accurately capture people's real-life behaviors are thus being pursued [163]. 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 needs to be more explicit and have user consent. Inviting participants 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 [164]. Such invitations must be accessible to a wide demographic, however. 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 [165].

Finally, our literature search returned many articles that consider the effects of social media use on a user's depression and anxiety (see Multimedia Appendix 2). A primary hypothesis, greatly debated in the specialist literature [166], 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 [167]. 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 amongst certain demographics (for example, young people accessing accounts linked to pro-anorexia or encouraging self-harm). As noted above, 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 [168]. 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 and data usage and privacy, as noted in the Ethical Aspects on 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 [169]. 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 proven a useful tool to identify relevant individuals for research, including delivering interventions to young people living with eating disorders [170] and who have been exposed to suicide [169], and using Facebook data to detect relapse in patients with schizophrenia [171]. As an example, the latter study 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 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 note, 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 does not represent diverse populations [172,173]. 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 since the field of mental healthcare 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 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 [174]. Mental health care therefore desires AI-driven methods that are transparent, explainable, and able to provide guidance to clinicians [26,173,175].

Limitations

We have reviewed the literature in what we deemed four 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 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] (see Multimedia Appendix 1).

We also only considered articles where the subjects of the studies self-reported a diagnosis of depression, anxiety, or schizophrenia; but 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 data sets that have information about mental health statuses, this approach also has weaknesses that have been explored in the psychopathological literature [176]. For example, compared to a manually compiled and curated data set, there are likely to be more false positive instances of nearly any common diagnosis, though false negatives 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 data sets 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 article may well not be applicable to non-English language platforms such as Weibo [177] and VK [178], 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 in this section, 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 in platforms such as Facebook, Twitter, and Instagram, male users are in the majority [84] and that social media users are generally well educated and affluent [85]. Cohort-based strategies for data set construction have been trialed to account for these biases [36]. There are also user-oriented biases that may distort data sets. A user's posting habits may change over time and convey a distorted view of their life and experiences [52]. This behavior may be influenced by reports published in traditional print media on the negative consequences of social media use [179]. It may also be influenced by the proliferation of usage-limiting tools, which encourage users to choose carefully the personal information they choose to share on social media platforms [180]. On a collective scale, certain users may post content significantly more frequently than others, creating imbalances in data sets and subsequent models. This is evident in two of the data sets we have covered [89,93]. Data augmentation is one approach that may alleviate this problem [97,118] while another includes data synthesis via LLMs [49]. Lastly, we should mention confirmation bias, which involves people's tendency to seek data that supports their beliefs and ignore or distort data contradicting them [181]. Where possible, a selection of appropriate data sets 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 reliable enough 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, 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 usage 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 usage agreements. Then, when data is subsequently compiled to 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 Welsh Government through Health and Care Research Wales. SES and AFP were supported by the European Union's Horizon 2020 research and innovation programme under grant agreement No 964874. AFP was also supported by a Medical Research Council (MRC) Programme Grant (MR/Y004094/1). JCC was supported by a UK Research and Innovation (UKRI) Future Leaders Fellowship.

Conflicts of Interest

None declared.

Abbreviations

ADHD: attention-deficit/hyperactivity disorder

ALBERT: A Lite Bidirectional Encoder Representations from Transformers

AI: artificial intelligence

BDI-II: Beck Depression Inventory-II

BERT: Bidirectional Encoder Representations from Transformers

DL: deep learning

EHR: electronic health record

GPT: Generative Pre-trained Transformer

LDA: Latent Dirichlet Allocation

LIWC: Linguistic Inquiry and Word Count

LLM: large language model

ML: machine learning

MDD: major depressive disorder

NLP: natural language processing

OCD: obsessive compulsive disorder

PTSD: post-traumatic stress disorder

RSDD: Reddit Self-reported Depression Diagnosis

SMHD: Self-reported Mental Health Diagnoses

SVM: support vector machine

Multimedia Appendix 1

Literature Search and Selection Strategy.

Multimedia Appendix 2

Verbose exposition of the four-stage literature selection process.

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. 2022 Feb;9(2):137-150. doi: 10.1016/S2215-0366(21)00395-3. Epub 2022 Jan 10. PMID: 35026139; PMCID: PMC8776563.
2. Picardi A, Lega I, Tarsitani L, Caredda M, Matteucci G, Zerella MP, Miglio R, Gigantesco A, Cerbo M, Gaddini A, Spandonaro F, Biondi M; 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. 2016 Jul 1;198:96-101. doi: 10.1016/j.jad.2016.03.025. Epub 2016 Mar 15. PMID: 27015158.
3. De Choudhury M, 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: Web Science 2013; May 2 - 4, 2013; Paris France
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 Apr 18 (pp. 2805-2814).
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
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 computing 2017 Feb 25 (pp. 1485-1500).
7. Dinu A, Moldovan AC. Automatic detection and classification of mental illnesses from general social media texts. In: Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021) 2021 Sep (pp. 358-366).
8. Singh A, Singh J. Automation of detection of social network mental disorders–A review. In: IOP Conference Series: Materials Science and Engineering 2021 (Vol. 1022, No. 1, p. 012008). IOP Publishing.
9. Muhammad KA. Unveiling the Emotional and Psychological States of Instagram Users: A Deep Learning Approach to Mental Health Analysis. Information Sciences Letters 2023: Vol. 12: Iss. 5
10. Ganguly C, Nayak S, Gupta AK. Mental health impact of COVID-19 and machine learning applications in combating mental disorders: a review. Artificial Intelligence, Machine Learning, and Mental Health in Pandemics. 2022 Jan 1:1-51.
11. Holmes EA, O'Connor RC, Perry VH, Tracey I, Wessely S, Arseneault L, Ballard C, Christensen H, Cohen Silver R, Everall I, Ford T, John A, Kabir T, King K, Madan I, Michie S, Przybylski AK, Shafran R, Sweeney A, Worthman CM, Yardley L, Cowan K, Cope C, Hotopf M, Bullmore E. Multidisciplinary research priorities for the COVID-19

pandemic: a call for action for mental health science. Lancet Psychiatry. 2020 Jun;7(6):547-560. doi: 10.1016/S2215-0366(20)30168-1. Epub 2020 Apr 15. PMID: 32304649; PMCID: PMC7159850.
12. Chancellor S, Birnbaum ML, Caine ED, Silenzio VM, De Choudhury M. A taxonomy of ethical tensions in inferring mental health states from social media. In: Proceedings of the conference on fairness, accountability, and transparency 2019 Jan 29 (pp. 79-88).
13. Hirschberg J, Manning CD. Advances in natural language processing. Science. 2015 Jul 17;349(6245):261-6. doi: 10.1126/science.aaa8685. PMID: 26185244.
14. Weizenbaum J. ELIZA—a computer program for the study of natural language communication between man and machine. Communications of the ACM. 1966 Jan 1;9(1):36-45.
15. Rajaraman V. From ELIZA to ChatGPT: history of human-computer conversation. Resonance. 2023 Jun;28(6):889-905.
16. Gilman ES, Kot S, Engineer M, Dixon E. Training Adults with Mild to Moderate Dementia in ChatGPT: Exploring Best Practices. In: Companion Proceedings of the 29th International Conference on Intelligent User Interfaces 2024 Mar 18 (pp. 101-106).
17. Whissell CM. The dictionary of affect in language. In: The measurement of emotions 1989 Jan 1 (pp. 113-131). Academic Press.
18. Whissell Dictionary of Affect in Language – Freeware. https://www.god-helmet.com/wp/whissel-dictionary-of-affect/index.htm [accessed 2024-08-02]
19. Doktoronline. URL: https://www.klikk.no/helse/doktoronline/ [accessed 2024-02-29]
20. Johnsen JA, Rosenvinge JH, Gammon D. Online group interaction and mental health: an analysis of three online discussion forums. Scand J Psychol. 2002 Dec;43(5):445-9. doi: 10.1111/1467-9450.00313. PMID: 12500784.
21. McKenna KY, Bargh JA. Coming out in the age of the Internet: Identity" demarginalization" through virtual group participation. Journal of personality and social psychology. 1998 Sep;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 May 23;9:20552076231176689. doi: 10.1177/20552076231176689. PMID: 37252260; PMCID: PMC10214072.
23. Haker H, Lauber C, Rössler W. Internet forums: a self-help approach for individuals with schizophrenia? Acta Psychiatr Scand. 2005 Dec;112(6):474-7. doi: 10.1111/j.1600-0447.2005.00662.x. PMID: 16279878.
24. Coppersmith G, Dredze M, Harman C. Quantifying mental health signals in Twitter. In: Proceedings of the workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality 2014 Jun (pp. 51-60).

25. Winn, J.G., Hao, T., Hardaway, J.W. and Oh, H., 2024. Redefining searching in non-medical sciences systematic reviews: The ascendance of Google Scholar as the primary database. Journal of Librarianship and Information Science, p.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. doi: 10.1007/s11192-020-03690-4. Epub 2020 Sep 21. PMID: 32981987; PMCID: PMC7505221.
27. Gusenbauer, M., 2019. Google Scholar to overshadow them all? Comparing the sizes of 12 academic search engines and bibliographic databases. Scientometrics, 118(1), pp.177-214.
28. Martín-Martín, A., Orduna-Malea, E., Thelwall, M. and López-Cózar, E.D., 2018. Google Scholar, Web of Science, and Scopus: A systematic comparison of citations in 252 subject categories. Journal of informetrics, 12(4), pp.1160-1177.
29. Ferrari R. Writing narrative style literature reviews. Medical writing. 2015 Dec 1;24(4):230-5.
30. Szeto MD, Barber C, Ranpariya VK, Anderson J, Hatch J, Ward J, Aguilera MN, Hassan S, Hamp A, Coolman T, Dellavalle RP. Emojis and Emoticons in Health Care and Dermatology Communication: Narrative Review. JMIR Dermatol. 2022 Aug 1;5(3):e33851. doi: 10.2196/33851. PMID: 36405493; PMCID: PMC9642845.
31. De Choudhury M, Gamon M, Counts S, Horvitz E. Predicting depression via social media. In: Proceedings of the international AAAI conference on web and social media 2013 (Vol. 7, No. 1, pp. 128-137).
32. Coppersmith G, Dredze M, Harman C. Quantifying mental health signals in Twitter. In: Proceedings of the workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality 2014 Jun (pp. 51-60).
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 Sep (pp. 2968-2978).
34. MacAvaney S, Desmet B, Cohan A, Soldaini L, Yates A, Zirikly A, Goharian N. RSDD-Time: Temporal Annotation of Self-Reported Mental Health Diagnoses. In: Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic 2018 Jun (pp. 168-173).
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 Aug (pp. 1485-1497).
36. Amir S, Dredze M, Ayers JW. Mental health surveillance over social media with digital cohorts. In: Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology 2019 Jun (pp. 114-120).

37. Parapar J, Martín-Rodilla P, Losada DE, Crestani F. Overview of eRisk at CLEF 2021: Early Risk Prediction on the Internet (Extended Overview). CLEF (Working Notes). 2021:864-87.
38. De Choudhury M, 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 May 2 (pp. 47-56).
39. Resnik P, Armstrong W, Claudino L, Nguyen T, Nguyen VA, 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: from linguistic signal to clinical reality 2015 (pp. 99-107).
40. 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: From linguistic signal to clinical reality 2015 (pp. 11-20).
41. 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 Apr 18 (pp. 3187-3196).
42. 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. 2017 Aug 14;19(8):e289. doi: 10.2196/jmir.7956. PMID: 28807891; PMCID: PMC5575421.
43. Guntuku SC, Yaden DB, Kern ML, Ungar LH, Eichstaedt JC. Detecting depression and mental illness on social media: an integrative review. Current Opinion in Behavioral Sciences. 2017 Dec 1;18:43-9.
44. Reece AG, Reagan AJ, Lix KLM, Dodds PS, Danforth CM, Langer EJ. Forecasting the onset and course of mental illness with Twitter data. Sci Rep. 2017 Oct 11;7(1):13006. doi: 10.1038/s41598-017-12961-9. PMID: 29021528; PMCID: PMC5636873.
45. Burdisso SG, Errecalde M, Montes-y-Gómez M. A text classification framework for simple and effective early depression detection over social media streams. Expert Systems with Applications. 2019 Nov 1;133:182-97.
46. Owen D, Camacho-Collados J, Anke LE. Towards Preemptive Detection of Depression and Anxiety in Twitter. In: Proceedings of the Fifth Social Media Mining for Health Applications Workshop & Shared Task 2020 Dec (pp. 82-89).
47. Malviya K, Roy B, Saritha SK. A transformers approach to detect depression in social media. In: 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS) 2021 Mar 25 (pp. 718-723). IEEE.
48. Sharma T, Panchendrarajan R, Saxena A. Characterisation of Mental Health Conditions in Social Media Using Deep Learning Techniques. Deep Learning for Social Media Data Analytics. 2022 Sep 19:157-76.
49. Bucur AM. Utilizing ChatGPT Generated Data to Retrieve Depression Symptoms from Social Media. arXiv e-prints. 2023 Jul:arXiv-2307.

50. Ramos dos Santos W, Paraboni I. Prompt-based mental health screening from social media text. arXiv e-prints. 2024 Jan:arXiv-2401.
51. Moreno MA, Jelenchick LA, Egan KG, Cox E, Young H, Gannon KE, Becker T. Feeling bad on Facebook: depression disclosures by college students on a social networking site. Depress Anxiety. 2011 Jun;28(6):447-55. doi: 10.1002/da.20805. Epub 2011 Mar 11. PMID: 21400639; PMCID: PMC3110617.
52. Schwartz HA, Eichstaedt J, Kern M, Park G, Sap M, Stillwell D, Kosinski M, Ungar L. Towards assessing changes in degree of depression through facebook. In: Proceedings of the workshop on computational linguistics and clinical psychology: from linguistic signal to clinical reality 2014 Jun (pp. 118-125).
53. 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 fourth workshop on computational linguistics and clinical Psychology—From linguistic signal to clinical reality 2017 Aug (pp. 85-95).
54. Hswen Y, Naslund JA, Brownstein JS, Hawkins JB. Monitoring Online Discussions About Suicide Among Twitter Users With Schizophrenia: Exploratory Study. JMIR Ment Health. 2018 Dec 13;5(4):e11483. doi: 10.2196/11483. PMID: 30545811; PMCID: PMC6315229.
55. 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. 2018 May 8;20(5):e168. doi: 10.2196/jmir.9267. PMID: 29739736; PMCID: PMC5964306.
56. Chen X, Sykora MD, Jackson TW, Elayan S. What about mood swings: Identifying depression on twitter with temporal measures of emotions. In: Companion proceedings of the web conference 2018 2018 Apr 23 (pp. 1653-1660).
57. 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 Jul 18;1(1):e26769. doi: 10.2196/26769. PMID: 34458682; PMCID: PMC8330892.
58. Kelley SW, Gillan CM. Using language in social media posts to study the network dynamics of depression longitudinally. Nat Commun. 2022 Feb 15;13(1):870. doi: 10.1038/s41467-022-28513-3. PMID: 35169166; PMCID: PMC8847554.
59. 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. 2023 Mar 24;2:e41205. doi: 10.2196/41205. PMID: 37525646; PMCID: PMC7614849.
60. 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. 2016 Apr 22;18(4):e92. doi: 10.2196/jmir.5385. PMID: 27105939; PMCID: PMC4859871.

61. Mikal J, Hurst S, Conway M. Ethical issues in using Twitter for population-level depression monitoring: a qualitative study. BMC Med Ethics. 2016 Apr 14;17:22. doi: 10.1186/s12910-016-0105-5. PMID: 27080238; PMCID: PMC4832544.
62. Conway M, O'Connor D. Social Media, Big Data, and Mental Health: Current Advances and Ethical Implications. Curr Opin Psychol. 2016 Jun;9:77-82. doi: 10.1016/j.copsyc.2016.01.004. PMID: 27042689; PMCID: PMC4815031.
63. 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. Proceedings of the ACM on Human-Computer Interaction. 2019 Nov 7;3(CSCW):1-32.
64. Nicholas J, Onie S, Larsen ME. Ethics and Privacy in Social Media Research for Mental Health. Curr Psychiatry Rep. 2020 Nov 23;22(12):84. doi: 10.1007/s11920-020-01205-9. PMID: 33225404.
65. 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. 2021 Jul 12;8(7):e24512. doi: 10.2196/24512. PMID: 34255701; PMCID: PMC8314152.
66. Cai N, Revez JA, Adams MJ, Andlauer TFM, Breen G, Byrne EM, Clarke TK, Forstner AJ, Grabe HJ, Hamilton SP, Levinson DF, Lewis CM, Lewis G, Martin NG, Milaneschi Y, Mors O, Müller-Myhsok B, Penninx BWJH, Perlis RH, Pistis G, Potash JB, Preisig M, Shi J, Smoller JW, Streit F, Tiemeier H, Uher R, Van der Auwera S, Viktorin A, Weissman MM; MDD Working Group of the Psychiatric Genomics Consortium; Kendler KS, Flint J. Minimal phenotyping yields genome-wide association signals of low specificity for major depression. Nat Genet. 2020 Apr;52(4):437-447. doi: 10.1038/s41588-020-0594-5. Epub 2020 Mar 30. PMID: 32231276; PMCID: PMC7906795.
67. Cong Q, Feng Z, Li F, Xiang Y, Rao G, Tao C. XA-BiLSTM: a deep learning approach for depression detection in imbalanced data. In: 2018 IEEE international conference on bioinformatics and biomedicine (BIBM) 2018 Dec 3 (pp. 1624-1627). IEEE.
68. Marnauzs S, Kalita J. A Domain Independent Social Media Depression Detection Model. Machine Learning in Computer Vision and Natural Language Processing. 2019 Aug 9:50. 23.
69. Trifan A, Semeraro D, Drake J, Bukowski R, Oliveira JL. Social Media Mining for Postpartum Depression Prediction. Stud Health Technol Inform. 2020 Jun 16;270:1391-1392. doi: 10.3233/SHTI200457. PMID: 32570674.
70. Bucur AM, Cosma A, Dinu LP. Early Risk Detection of Pathological Gambling, Self-Harm and Depression Using BERT. 25.
71. 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: Digital Humanism: A Human-Centric Approach to Digital Technologies 2022 Jun 29 (pp. 81-103). Cham: Springer International Publishing. 26.

72. Kulkarni H, MacAvaney S, Goharian N, Frieder O. Knowledge augmentation for early depression detection. In: International Workshop on Health Intelligence 2023 Feb 13 (pp. 175-191). Cham: Springer Nature Switzerland.
73. Souza VB, Nobre J, Becker K. Characterization of anxiety, depression, and their comorbidity from texts of social networks. Anais do XXXV Simpósio Brasileiro de Bancos de Dados. 2020:121-32.
74. 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: 2023 IEEE 11th International Conference on Healthcare Informatics (ICHI) 2023 Jun 26 (pp. 193-199). IEEE.
75. Buddhitha P, Inkpen D. Multi-task learning to detect suicide ideation and mental disorders among social media users. Front Res Metr Anal. 2023 Apr 17;8:1152535. doi: 10.3389/frma.2023.1152535. PMID: 37138946; PMCID: PMC10149941.
76. Khan MR, Sakib S, Habib AB, Hossain MI. A Machine Learning and Deep Learning Approach to Classify Mental Illness with the Collaboration of Natural Language Processing. In: Proceedings of International Conference on Information and Communication Technology for Development: ICICTD 2022 2023 Jan 26 (pp. 83-94). Singapore: Springer Nature Singapore.
77. 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 (LOUHI) 2022 Dec (pp. 184-196).
78. Sekulić I, Strube M. Adapting Deep Learning Methods for Mental Health Prediction on Social Mediage. In: Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019) 2019 Nov (pp. 322-327).
79. Dinu A, Moldovan AC. Automatic detection and classification of mental illnesses from general social media texts. In: Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021) 2021 Sep (pp. 358-366).
80. 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. 2022 Jul;26(7):3303-3311. doi: 10.1109/JBHI.2022.3151589. Epub 2022 Jul 1. PMID: 35230959.
81. American Psychiatric Association DS, American Psychiatric Association. Diagnostic and statistical manual of mental disorders: DSM-5. Washington, DC: American psychiatric association; 2013 May 22.
82. Ireland M, Iserman M. Within and between-person differences in language used across anxiety support and neutral reddit communities. In: Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic 2018 Jun (pp. 182-193).

83. Harrigian K, Aguirre C, Dredze M. Do models of mental health based on social media data generalize?. In: Findings of the association for computational linguistics: EMNLP 2020 2020 Nov (pp. 3774-3788).
84. The 2024 Social Media Demographics Guide. URL: https://khoros.com/resources/social-media-demographics-guide [accessed: 2024-04-13]
85. Hruska J, Maresova P. Use of social media platforms among adults in the United States—behavior on social media. Societies. 2020 Mar 23;10(1):27.
86. Amir S, Coppersmith G, Carvalho P, Silva MJ, Wallace BC. Quantifying mental health from social media with neural user embeddings. In: Machine Learning for Healthcare Conference 2017 Nov 6 (pp. 306-321). PMLR.
87. Wu SH, Qiu ZJ. A RoBERTa-based model on measuring the severity of the signs of depression. In: CLEF (Working Notes) 2021 (pp. 1071-1080).
88. Inkpen D, Skaik R, Buddhitha P, Angelov D, Fredenburgh MT. uOttawa at eRisk 2021: Automatic Filling of the Beck's Depression Inventory Questionnaire using Deep Learning. In: CLEF (Working Notes) 2021 (pp. 966-980).
89. Reddit Self-reported Depression Diagnosis (RSDD) dataset. URL: https://georgetown-ir-lab.github.io/emnlp17-depression/ [accessed 2024-04-05]
90. ir@Georgetown – Resources. URL: https://ir.cs.georgetown.edu/resources/ [accessed 2024-04-05]
91. ir@Georgetown – Resources - SMHD. URL: https://ir.cs.georgetown.edu/resources/smhd.html [accessed 2024-04-05]
92. CLPsych 2015 Shared Task Evaluation. URL: https://www.cs.jhu.edu/~mdredze/clpsych-2015-shared-task-evaluation/ [accessed 2024-04-05]
93. eRisk 2021 Text Research Collection. URL: https://tec.citius.usc.es/ir/code/eRisk2021.html [accessed 2024-04-05]
94. What Is Topic Modeling? A Beginner's Guide. URL: https://levity.ai/blog/what-is-topic-modeling [accessed: 2024-04-12]
95. Pennebaker JW, Boyd RL, Jordan K, Blackburn K. The development and psychometric properties of LIWC2015.
96. Hier DB, Brint SU. A Neuro-ontology for the neurological examination. BMC Med Inform Decis Mak. 2020 Mar 4;20(1):47. doi: 10.1186/s12911-020-1066-7. PMID: 32131804; PMCID: PMC7057564.
97. Chen J, Tam D, Raffel C, Bansal M, Yang D. An empirical survey of data augmentation for limited data learning in nlp. Transactions of the Association for Computational Linguistics. 2023 Mar 14;11:191-211.
98. Liu B, Liu B. Supervised learning. Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data. 2011:63-132.
99. Noble WS. What is a support vector machine? Nat Biotechnol. 2006 Dec;24(12):1565-7. doi: 10.1038/nbt1206-1565. PMID: 17160063.
100. Breiman L. Random forests. Machine learning. 2001 Oct;45:5-32.

101. Zhong G, Wang LN, Ling X, Dong J. An overview on data representation learning: From traditional feature learning to recent deep learning. The Journal of Finance and Data Science. 2016 Dec 1;2(4):265-78.
102. What is BERT and how is it used in AI? URL: https://h2o.ai/wiki/bert/ [accessed: 2024-04-11]
103. Devlin, J., Chang, M.W., Lee, K. and Toutanova, K., 2019, June. 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, Volume 1 (Long and Short Papers) (pp. 4171-4186).
104. 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 arXiv:1909.11942. 2019 Sep 26.
105. 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 Thirteenth Language Resources and Evaluation Conference 2022 Jun (pp. 7184-7190).
106. 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 Thirteenth Language Resources and Evaluation Conference 2022 Jun (pp. 7184-7190).
107. Jo A. The promise and peril of generative AI. Nature. 2023 Feb 9;614(1):214-6.
108. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, Agarwal S. Language models are few-shot learners. Advances in neural information processing systems. 2020;33:1877-901.
109. 5u T, He S, Liu J, Sun S, Liu K, Han QL, Tang Y. A brief overview of ChatGPT: The history, status quo and potential future development. IEEE/CAA Journal of Automatica Sinica. 2023 May 1;10(5):1122-36.
110. Monaghan TF, Rahman SN, Agudelo CW, Wein AJ, Lazar JM, Everaert K, Dmochowski RR. Foundational Statistical Principles in Medical Research: Sensitivity, Specificity, Positive Predictive Value, and Negative Predictive Value. Medicina (Kaunas). 2021 May 16;57(5):503. doi: 10.3390/medicina57050503. PMID: 34065637; PMCID: PMC8156826.
111. Accuracy, Precision, Recall or F1? URL: https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9 [accessed 2024-04-18]
112. Understanding AUC - ROC Curve. https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5 [accessed 2024-08-18]
113. Kessler RC, Berglund P, Demler O, Jin R, Koretz D, Merikangas KR, Rush AJ, Walters EE, Wang PS; National Comorbidity Survey Replication. The epidemiology of

major depressive disorder: results from the National Comorbidity Survey Replication (NCS-R). JAMA. 2003 Jun 18;289(23):3095-105. doi: 10.1001/jama.289.23.3095. PMID: 12813115.
114. Fujita F, Diener E, Sandvik E. Gender differences in negative affect and well-being: the case for emotional intensity. J Pers Soc Psychol. 1991 Sep;61(3):427-34. doi: 10.1037//0022-3514.61.3.427. PMID: 1941513.
115. Software and systems engineering — Software testing — Part 11: Guidelines on the testing of AI-based systems. 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]
116. Shaikh S, Daudpota SM, Imran AS, Kastrati Z. Towards improved classification accuracy on highly imbalanced text dataset using deep neural language models. Applied Sciences. 2021 Jan 19;11(2):869.
117. 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 (ICNLSP 2022) 2022 Dec (pp. 79-88).
118. 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. 2023;14(1):135-150. doi: 10.1007/s13042-022-01553-3. Epub 2022 Apr 12. PMID: 35432623; PMCID: PMC9001823.
119. Li B, Hou Y, Che W. Data augmentation approaches in natural language processing: A survey. Ai Open. 2022 Jan 1;3:71-90.
120. Santos WR, de Oliveira RL, Paraboni I. SetembroBR: a social media corpus for depression and anxiety disorder prediction. Language Resources and Evaluation. 2023 Jan 11:1-28.
121. Santos W, Yoon S, Paraboni I. Mental health prediction from social media text using mixture of experts. IEEE Latin America Transactions. 2023 Jun;21(6):723-9.
122. Andrews GJ. Co-creating health's lively, moving frontiers: brief observations on the facets and possibilities of non-representational theory. Health Place. 2014 Nov;30:165-70. doi: 10.1016/j.healthplace.2014.09.002. Epub 2014 Oct 1. PMID: 25282125.
123. Gur RE, Kohler CG, Ragland JD, Siegel SJ, Lesko K, Bilker WB, Gur RC. Flat affect in schizophrenia: relation to emotion processing and neurocognitive measures. Schizophr Bull. 2006 Apr;32(2):279-87. doi: 10.1093/schbul/sbj041. Epub 2006 Feb 1. PMID: 16452608; PMCID: PMC2632232.
124. Rickelman BL. Anosognosia in individuals with schizophrenia: toward recovery of insight. Issues Ment Health Nurs. 2004 Apr-May;25(3):227-42. doi: 10.1080/01612840490274741. PMID: 14965844.
125. National Alliance on Mental Illness. URL: https://www.nami.org/About-Mental-Illness/Mental-Health-Conditions/Schizophrenia [accessed 2024-03-19]

126. 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. 2020 Apr 21;22(4):e16470. doi: 10.2196/16470. PMID: 32314969; PMCID: PMC7201321.
127. Robinson P, Turk D, Jilka S, Cella M. Measuring attitudes towards mental health using social media: investigating stigma and trivialisation. Soc Psychiatry Psychiatr Epidemiol. 2019 Jan;54(1):51-58. doi: 10.1007/s00127-018-1571-5. Epub 2018 Aug 1. PMID: 30069754; PMCID: PMC6336755.
128. Meyer-Lindenberg A. The non-ergodic nature of mental health and psychiatric disorders: implications for biomarker and diagnostic research. World Psychiatry. 2023 Jun;22(2):272-274. doi: 10.1002/wps.21086. PMID: 37159352; PMCID: PMC10168159.
129. Zivin K, Eisenberg D, Gollust SE, Golberstein E. Persistence of mental health problems and needs in a college student population. J Affect Disord. 2009 Oct;117(3):180-5. doi: 10.1016/j.jad.2009.01.001. Epub 2009 Jan 28. PMID: 19178949.
130. Eisenberg D, Golberstein E, Gollust SE. Help-seeking and access to mental health care in a university student population. Med Care. 2007 Jul;45(7):594-601. doi: 10.1097/MLR.0b013e31803bb4c1. PMID: 17571007.
131. 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 May 7 (pp. 571-580).
132. Bell CC. DSM-IV: diagnostic and statistical manual of mental disorders. Jama. 1994 Sep 14;272(10):828-9.
133. Westrin A, Lam RW. Seasonal affective disorder: a clinical update. Ann Clin Psychiatry. 2007 Oct-Dec;19(4):239-46. doi: 10.1080/10401230701653476. PMID: 18058281.
134. Mohammad S, Kiritchenko S, Zhu X. NRC-Canada: Building the State-of-the-Art in Sentiment Analysis of Tweets. In: Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013) 2013 Jun (pp. 321-327).
135. Messinger JW, Trémeau F, Antonius D, Mendelsohn E, Prudent V, Stanford AD, Malaspina D. Avolition and expressive deficits capture negative symptom phenomenology: implications for DSM-5 and schizophrenia research. Clin Psychol Rev. 2011 Feb;31(1):161-8. doi: 10.1016/j.cpr.2010.09.002. Epub 2010 Sep 18. PMID: 20889248; PMCID: PMC2997909.
136. 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. 2007 Jan;133(1):1-24. doi: 10.1037/0033-2909.133.1.1. PMID: 17201568.

137. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med. 2001 Sep;16(9):606-13. doi: 10.1046/j.1525-1497.2001.016009606.x. PMID: 11556941; PMCID: PMC1495268.
138. Rickard N, Arjmand HA, 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. 2016 Nov 23;3(4):e49. doi: 10.2196/mental.6202. PMID: 27881358; PMCID: PMC5143469.
139. Sykora MD, Jackson TW, O'Brien A, Elayan S. Emotive Ontology: Extracting Fine-grained Emotions from Terse, Informal Messages. IADIS International Journal on Computer Science & Information Systems. 2013 Jul 1;8(2).
140. Yang Z, Dai Z, Yang Y, Carbonell J, Salakhutdinov RR, Le QV. Xlnet: Generalized autoregressive pretraining for language understanding. Advances in neural information processing systems. 2019;32.
141. Jashinsky J, Burton SH, Hanson CL, West J, Giraud-Carrier C, Barnes MD, Argyle T. Tracking suicide risk factors through Twitter in the US. Crisis. 2014;35(1):51-9. doi: 10.1027/0227-5910/a000234. PMID: 24121153.
142. Fodeh S, Goulet J, Brandt C, Hamada AT. Leveraging Twitter to better identify suicide risk. In: Medical Informatics and Healthcare 2017 Oct 18 (pp. 1-7). PMLR.
143. Hawton K, Sutton L, Haw C, Sinclair J, Deeks JJ. Schizophrenia and suicide: systematic review of risk factors. Br J Psychiatry. 2005 Jul;187:9-20. doi: 10.1192/bjp.187.1.9. PMID: 15994566.
144. Sareen J, Cox BJ, Afifi TO, de Graaf R, Asmundson GJ, ten Have M, Stein MB. Anxiety disorders and risk for suicidal ideation and suicide attempts: a population-based longitudinal study of adults. Arch Gen Psychiatry. 2005 Nov;62(11):1249-57. doi: 10.1001/archpsyc.62.11.1249. PMID: 16275812.
145. Golder S, Ahmed S, Norman G, Booth A. Attitudes Toward the Ethics of Research Using Social Media: A Systematic Review. J Med Internet Res. 2017 Jun 6;19(6):e195. doi: 10.2196/jmir.7082. PMID: 28588006; PMCID: PMC5478799.
146. 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. 2019 Nov 13;6(11):e12942. doi: 10.2196/12942. PMID: 31719022; PMCID: PMC6881781.
147. 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 Transactions on Computational Social Systems. 2019 Jun 3;6(5):907-21.
148. Reidenberg JR, Breaux T, Cranor LF, French B, Grannis A, Graves JT, Liu F, McDonald A, Norton TB, Ramanath R, Russell NC. Disagreeable privacy policies: Mismatches between meaning and users' understanding. Berkeley Tech. LJ. 2015;30:39.
149. Sleigh J. Experiences of Donating Personal Data to Mental Health Research: An Explorative Anthropological Study. Biomed Inform Insights. 2018 Jun

27;10:1178222618785131. doi: 10.1177/1178222618785131. PMID: 30013355; PMCID: PMC6043936.
150. Beninger K, Fry A, Jago N, Lepps H, Nass L, Silvester H. Research using social media; users' views. NatCen Social Research. 2014 Feb 20;20.
151. Haslam N. Dehumanization: an integrative review. Pers Soc Psychol Rev. 2006;10(3):252-64. doi: 10.1207/s15327957pspr1003_4. PMID: 16859440.
152. 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. 2011 Jan;72(1):54-62. doi: 10.1016/j.socscimed.2010.09.037. Epub 2010 Oct 26. PMID: 21112682.
153. Spanakis P, Wadman R, Walker L, Heron P, Mathers A, Baker J, Johnston G, Gilbody S, Peckham E. Measuring the digital divide among people with severe mental ill health using the essential digital skills framework. Perspect Public Health. 2024 Jan;144(1):21-30. doi: 10.1177/17579139221106399. Epub 2022 Aug 5. PMID: 35929589; PMCID: PMC10757390.
154. Eichstaedt JC, Smith RJ, Merchant RM, Ungar LH, Crutchley P, Preoţiuc-Pietro D, Asch DA, Schwartz HA. Facebook language predicts depression in medical records. Proc Natl Acad Sci U S A. 2018 Oct 30;115(44):11203-11208. doi: 10.1073/pnas.1802331115. Epub 2018 Oct 15. PMID: 30322910; PMCID: PMC6217418.
155. Woolway GE, Legge SE, Lynham A, Smart SE, Hubbard L, Daniel ER, Pardiñas AF, Escott-Price V, O'Donovan MC, Owen MJ, Jones IR, Walters JT. Assessing the validity of a self-reported clinical diagnosis of schizophrenia. medRxiv [Preprint]. 2023 Dec 8:2023.12.06.23299622. doi: 10.1101/2023.12.06.23299622. PMID: 38106032; PMCID: PMC10723562.
156. Cheng SW, Chang CW, Chang WJ, Wang HW, Liang CS, Kishimoto T, Chang JP, Kuo JS, Su KP. The now and future of ChatGPT and GPT in psychiatry. Psychiatry Clin Neurosci. 2023 Nov;77(11):592-596. doi: 10.1111/pcn.13588. Epub 2023 Sep 11. PMID: 37612880; PMCID: PMC10952959.
157. Grabb D. The impact of prompt engineering in large language model performance: a psychiatric example. Journal of Medical Artificial Intelligence. 2023 Oct 30;6.
158. Xu X, Yao B, Dong Y, Gabriel S, Yu H, Hendler J, Ghassemi M, Dey AK, Wang D. Mental-llm: Leveraging large language models for mental health prediction via online text data. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies. 2024 Mar 6;8(1):1-32.
159. Ghosh S, Ekbal A, Bhattacharyya P. What does your bio say? Inferring Twitter users' depression status from multimodal profile information using deep learning. IEEE transactions on computational social systems. 2021 Oct 12;9(5):1484-94.
160. Semwal N, Suri M, Chaudhary D, Gorton I, Kumar B. Multimodal Analysis and Modality Fusion for Detection of Depression from Twitter Data. Association for the Advancement of Artificial Intelligence. 2023:1-5.

161. Bucur AM, Cosma A, Rosso P, Dinu LP. It's just a matter of time: Detecting depression with time-enriched multimodal transformers. In: European Conference on Information Retrieval 2023 Mar 17 (pp. 200-215). Cham: Springer Nature Switzerland.
162. Zhang H, Wang H, Han S, Li W, Zhuang L. Detecting depression tendency with multimodal features. Comput Methods Programs Biomed. 2023 Oct;240:107702. doi: 10.1016/j.cmpb.2023.107702. Epub 2023 Jul 6. PMID: 37531689.
163. Khoo LS, Lim MK, Chong CY, McNaney R. Machine Learning for Multimodal Mental Health Detection: A Systematic Review of Passive Sensing Approaches. Sensors (Basel). 2024 Jan 6;24(2):348. doi: 10.3390/s24020348. PMID: 38257440; PMCID: PMC10820860.
164. van Driel II, Giachanou A, Pouwels JL, Boeschoten L, Beyens I, Valkenburg PM. Promises and pitfalls of social media data donations. Communication Methods and Measures. 2022 Oct 2;16(4):266-82.
165. Choi S. Privacy literacy on social media: Its predictors and outcomes. International Journal of Human–Computer Interaction. 2023 Jan 2;39(1):217-32.
166. Orben A, Przybylski AK. The association between adolescent well-being and digital technology use. Nat Hum Behav. 2019 Feb;3(2):173-182. doi: 10.1038/s41562-018-0506-1. Epub 2019 Jan 14. PMID: 30944443.
167. Torous J, Bucci S, Bell IH, Kessing LV, Faurholt-Jepsen M, Whelan P, Carvalho AF, Keshavan M, Linardon J, Firth J. The growing field of digital psychiatry: current evidence and the future of apps, social media, chatbots, and virtual reality. World Psychiatry. 2021 Oct;20(3):318-335. doi: 10.1002/wps.20883. PMID: 34505369; PMCID: PMC8429349.
168. 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. 2020 Oct 12;22(10):e22635. doi: 10.2196/22635. PMID: 32936777; PMCID: PMC7575341.
169. La Sala L, Pirkis J, Cooper C, Hill NTM, Lamblin M, Rajaram G, Rice S, Teh Z, Thorn P, Zahan R, Robinson J. Acceptability and Potential Impact of the #chatsafe Suicide Postvention Response Among Young People Who Have Been Exposed to Suicide: Pilot Study. JMIR Hum Factors. 2023 May 19;10:e44535. doi: 10.2196/44535. PMID: 37204854; PMCID: PMC10238962.
170. Kasson E, Vázquez MM, Doroshenko C, Fitzsimmons-Craft EE, Wilfley DE, Taylor CB, Cavazos-Rehg PA. Exploring Social Media Recruitment Strategies and Preliminary Acceptability of an mHealth Tool for Teens with Eating Disorders. Int J Environ Res Public Health. 2021 Jul 28;18(15):7979. doi: 10.3390/ijerph18157979. PMID: 34360270; PMCID: PMC8345665.
171. Birnbaum ML, Ernala SK, Rizvi AF, Arenare E, R Van Meter A, De Choudhury M, Kane JM. Detecting relapse in youth with psychotic disorders utilizing patient-generated and patient-contributed digital data from Facebook. NPJ Schizophr. 2019

Oct 7;5(1):17. doi: 10.1038/s41537-019-0085-9. PMID: 31591400; PMCID: PMC6779748.
172. Alowais SA, Alghamdi SS, Alsuhebany N, Alqahtani T, Alshaya AI, Almohareb SN, Aldairem A, Alrashed M, Bin Saleh K, Badreldin HA, Al Yami MS, Al Harbi S, Albekairy AM. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. BMC Med Educ. 2023 Sep 22;23(1):689. doi: 10.1186/s12909-023-04698-z. PMID: 37740191; PMCID: PMC10517477.
173. 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. 2024 Jan;66(Suppl 2):S414-S419. doi: 10.4103/indianjpsychiatry.indianjpsychiatry_926_23. Epub 2024 Jan 24. PMID: 38445270; PMCID: PMC10911327.
174. Lee EE, Torous J, De Choudhury M, Depp CA, Graham SA, Kim HC, Paulus MP, Krystal JH, Jeste DV. Artificial Intelligence for Mental Health Care: Clinical Applications, Barriers, Facilitators, and Artificial Wisdom. Biol Psychiatry Cogn Neurosci Neuroimaging. 2021 Sep;6(9):856-864. doi: 10.1016/j.bpsc.2021.02.001. Epub 2021 Feb 8. PMID: 33571718; PMCID: PMC8349367.
175. Maddox TM, Rumsfeld JS, Payne PRO. Questions for Artificial Intelligence in Health Care. JAMA. 2019 Jan 1;321(1):31-32. doi: 10.1001/jama.2018.18932. PMID: 30535130.
176. Tiego J, Martin EA, DeYoung CG, Hagan K, Cooper SE, Pasion R, Satchell L, Shackman AJ, Bellgrove MA, Fornito A; HiTOP Neurobiological Foundations Work Group. Precision behavioral phenotyping as a strategy for uncovering the biological correlates of psychopathology. Nat Ment Health. 2023 May;1(5):304-315. doi: 10.1038/s44220-023-00057-5. Epub 2023 May 10. PMID: 37251494; PMCID: PMC10210256.
177. Weibo. URL: https://m.weibo.cn/ [accessed 2024-07-21]
178. VK. URL: https://vk.com/ [accessed 2024-07-21]
179. Just How Harmful Is Social Media? Our Experts Weigh-In. URL: https://www.publichealth.columbia.edu/news/just-how-harmful-social-media-our-experts-weigh [accessed 2024-08-31]
180. Kozyreva A, Lewandowsky S, Hertwig R. Citizens Versus the Internet: Confronting Digital Challenges With Cognitive Tools. Psychol Sci Public Interest. 2020 Dec;21(3):103-156. doi: 10.1177/1529100620946707. PMID: 33325331; PMCID: PMC7745618.
181. Peters U. What is the function of confirmation bias?. Erkenntnis. 2022 Jun;87(3):1351-76.