Mental Health Disorder Detection Beyond Social Media: A Systematic Review of Available Datasets

Sadiya Sayara Chowdhury Puspo[1], Ana-Maria Bucur[2], Stevie Chancellor[3], Ozlem Uzuner[1], Marcos Zampieri[1]

[1]George Mason University, USA
[2]Università della Svizzera italiana, Switzerland
[3]University of Minnesota, USA

Abstract

Detecting mental health disorders in a timely manner is an important societal challenge. NLP and machine learning (ML) methods used to assist with detection rely on data collected primarily from social media. However, such datasets often have sampling biases and inherent ethical and privacy issues. One avenue to overcome these limitations is non-social media data. We present the first comprehensive review of non-social media, free-text datasets for mental health research. We use the PRISMA methodology to conduct our survey and we review datasets available in multiple languages. We find that non-social media free-text based datasets are predominantly focused on English and on detecting depression. These datasets also vary in demographics, platforms, data types, annotation techniques, and methodologies. This systematic review also reveals key gaps and highlights opportunities to develop more diverse, reliable and clinically-relevant resources.

Keywords: language resources, mental health disorders, clinical NLP

1. Introduction

The prevalence of mental health disorders is a global concern. In the USA, for example, one in every four adults experiences a diagnosable mental health disorder each year[1]. Furthermore, research shows that the majority of individuals who die by suicide have an identifiable mental health condition such as depression (Bailey et al., 2011) or substance use disorder (Lynch et al., 2020).

The limited access to mental health care services has become an urgent societal issue (Cummings et al., 2013; Hester, 2017). This motivated research on applying NLP and ML models to social media data for early identification of mental disorders and supporting individuals at risk. Social media platforms such as Twitter (Chatterjee et al., 2022; Khafaga et al., 2023), Reddit (Boettcher, 2021; Adil Jaafar and Abdul-Salam Jasim, 2022) and Facebook (Calvo et al., 2017; Islam et al., 2018) offer a rich, unobtrusive stream of user-generated content that captures real-life expressions with reduced reporting bias, aiding in the detection and understanding of mental disorders (Sametoğlu et al., 2024; Raihan et al., 2024, 2026).

Using social media data for mental health screening raises ethical and privacy concerns regarding user consent, construct validity, and the potential for algorithmic misuse (Chancellor et al., 2019a; Nicholas et al., 2020; Chancellor and De Choudhury, 2020). Collecting data from social media can lead to unintended biases as the data may only reflect the experiences of individuals who are willing to openly discuss their mental health online, which primarily includes those who are active on social media platforms (Chancellor et al., 2019b). Finally, demographics vary across platforms and they may not be representative of the general population. For instance, X users are primarily male, while TikTok and Instagram users are mostly female; likewise, many platforms tend to be used by teens and young adults (Olteanu et al., 2019; Zhao et al., 2022). Pragmatically, many social media sites have limited access to APIs for research.

The widespread use of social media data in this domain is partly due to the significant shortage of non-social media free-text based datasets related to mental health disorders derived from clinical and other reliable sources. Clinical data, such as electronic medical records (EMRs) or electronic health records (EHRs) with discharge summaries or clinical notes, psychiatric interview transcripts, and responses to standardized open-ended questionnaires, offer rich, detailed insights into patients' mental health. Unlike social media data, these sources contain carefully documented information from healthcare professionals, including diagnostic details, symptom descriptions, and treatment histories, often supported by validated clinical scales.

Despite their great potential, non-social media free-text based datasets remain underexplored primarily due to privacy concerns, data access challenges, and annotation complexities. This scarcity presents a barrier to advancing robust, generaliz-

[1] https://www.hopkinsmedicine.org/health/wellness-and-prevention/mental-health-disorder-statistics

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 6235–6250
11–16 May 2026.

able NLP models that can be effectively integrated into clinical practice. We aim to address this gap by systematically reviewing non-social media free-text based datasets for mental health research. We explore their diversity in terms of source, structure, clinical annotation, types, and research adoption to guide future efforts in dataset development and application. Previous related surveys have primarily focused on datasets from social media platforms (Harrigian et al., 2021; Ríssola et al., 2020; Skaik and Inkpen, 2020; Abdulsalam and Alhothali, 2024; Bucur et al., 2025a,b). To the best of our knowledge, this is the first systematic review[2] (Grant and Booth, 2009) of free-text based mental health data sources beyond social media.

We address the following research questions:

RQ_1: What non-social media free-text based datasets are available for mental health research, and how do they vary by source, structure, and population?

RQ_2: How are mental health conditions defined and labeled in these datasets, and what are the clinical implications of these labeling methods?

RQ_3: What factors contribute to the popularity and adoption of non-social media free-text based mental health datasets in research?

2. Methods

This systematic review adopts the PRISMA[3] reporting guidelines to comprehensively map and systematically analyze the landscape of the free-text based datasets in mental health research beyond social media, building on prior frameworks (e.g. Templier and Paré (2015); Zarate et al. (2022); Pazdur et al. (2025)). PRISMA is a standardized guideline aimed at ensuring clear and thorough reporting of systematic reviews and meta-analyses. It features a 27-item checklist that helps authors include all essential components of their review, from initial identification to final conclusions. A key part of PRISMA is its flow diagram (Figure 1), which visually maps the process of selecting studies, making the review process more transparent and easy to follow (Page et al., 2021).

To follow the checklists of PRISMA, we began with a systematic literature search that continued through June 2025. We used Publish or Perish[4] to query multiple academic databases, including Google Scholar, and PubMed. The searches were carried out separately for key mental health related terms: 'depression', 'anxiety', 'suicidal ideation',

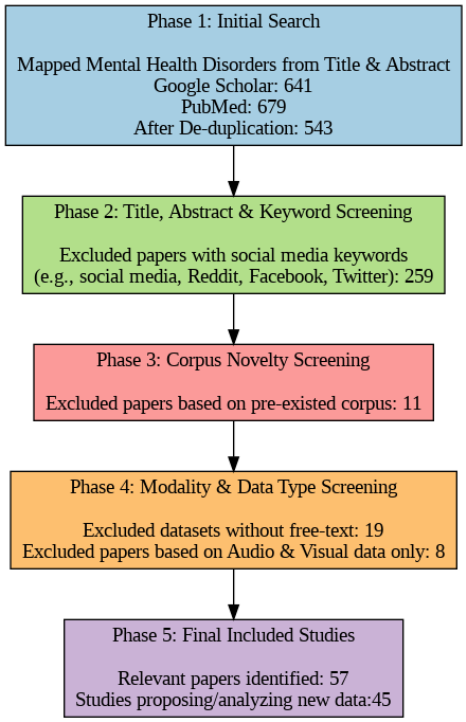

Figure 1: PRISMA flowchart illustrating the systematic literature search and screening process.

'suicide', 'mental disorder', 'mental health crisis', and 'mental health disorder', combined with targeted keywords 'identification', 'detection', 'prediction', and 'analysis' to capture each relevant disorder type of mental health and associated dataset corpus. We then manually screened the abstracts and included studies that explicitly referenced the use of any datasets or the application of any computational models or techniques after deduplication, resulting in 543 papers within our initial scope. The rationale for selecting depression and anxiety as key terms is that they are among the most prevalent mental health disorders worldwide (Institute for Health Metrics and Evaluation, 2024). Additionally, given that suicide remains a leading cause of death globally among adults (Centers for Disease Control and Prevention, 2025), we prioritize the inclusion of these mental health conditions in our study.

To ensure focus on non-social media sources, we excluded papers whose titles or abstracts included keywords related to social media platforms or mentions of datasets scraped from them, such as 'social media', 'Reddit', 'Facebook', or 'Twitter'. This filtering was essential to isolate papers based on clinical notes, semi-structured interviews, and discharge summaries from other non-social media platforms, and we ended up with 259 total papers after this filtering. Subsequently, we conducted corpus novelty screening to remove studies based on pre-existing datasets, followed by modality screen-

[2] https://github.com/SadiyaPuspo/MHD-Beyond-Social-Media-Datasets-Review

[3] https://www.prisma-statement.org/prisma-2020

[4] https://harzing.com/resources/publish-or-perish

ing to exclude papers that focused solely on audio or visual data, and data type screening to filter out datasets containing only numerical or categorical values, such as scale-based survey responses. Through this multi-phase screening, we identified 57 relevant papers, of which 40 proposed or analyzed newly collected free-text based mental health datasets.

In addition to keyword-based search, we followed a backtracking strategy; whenever a paper used an existing dataset, we traced it back to the original dataset publication, even if it did not contain our predefined keywords. We also examined all papers cited in the literature review or background sections of included studies and explored any dataset-related papers mentioned in relevant survey papers found from the initial search. As a result, our final collection may include important dataset papers whose titles or abstracts do not directly match the initial search terms, but were directly aligned with our objectives. This process yielded 5 additional papers, for a total of 45.

Several datasets were initially annotated for mental disorders, but their final task focused on emotion classification, including loneliness, fear, anger, hopelessness, and self-identification. Since emotions like loneliness and hopelessness can serve as indicators of depression (Rholes et al., 1985) or suicidal ideation (Baryshnikov et al., 2020), we decided to include these studies in our survey.

3. Distribution Analysis

In this section, we provide information on the 45 free-text based datasets included in this systematic review. Table 1 provides a comprehensive summary of the final set of datasets. The datasets were published between 2004 and 2025 in multiple languages, including English, Chinese, Polish, Korean, Japanese, Arabic, and some code-mixed. These datasets cover a broad spectrum of mental health conditions, such as depression, postnatal depression, anxiety, schizophrenia, suicidal ideation, Post-Traumatic Stress Disorder, and bipolar disorder, collected from platforms like clinics, colleges, mobile apps, and therapy sessions. Data types include interview transcripts, essays, discharge summaries, clinical records, forum posts, and suicide notes. While some datasets are publicly available, others are restricted or require agreements, with a few lacking availability details. We show their distribution below across disorders, languages, platforms, data types, demographics, and availability types.

Temporal Distribution Figure 2 illustrates the number of free-text based datasets proposed and analyzed in published studies each year between 2004 and 2025. The introduction of such datasets

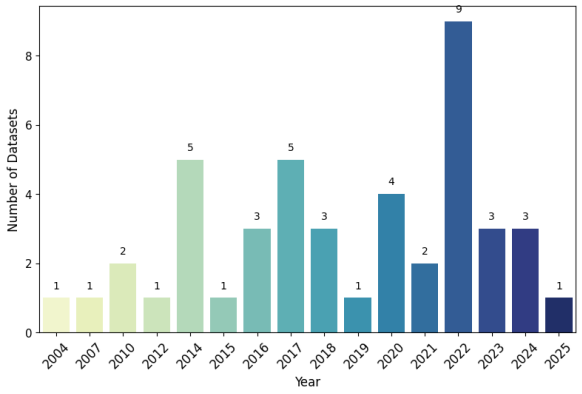

Figure 2: Distribution of the datasets by year

remained minimal and steady up to 2012. In 2014 and 2017, there was a noticeable rise in the number of proposed datasets, indicating intensified research efforts in this domain during that period. The proposal of such datasets peaked in 2022, which can be attributed to greater attention to mental health concerns after the COVID-19 pandemic (Bucur et al., 2025b). The lower count in 2025 is probably a result of the systematic search being conducted until June 2025. Despite this overall growth, the process of collecting such datasets often involves complex procedures, including ethics reviews, annotator training, and repeated permission requests, which may explain the limited number of datasets over the years.

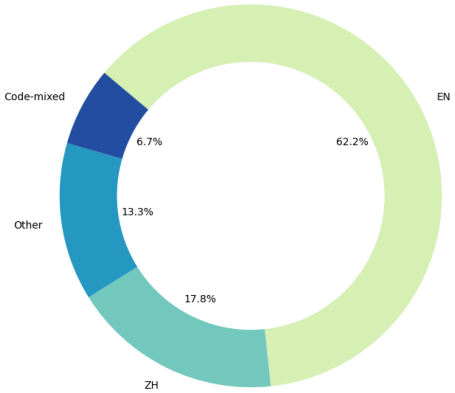

Figure 3: Distribution of datasets into by language groups.

Language Distribution Figure 3 illustrates the distribution of the datasets by language group. English (EN) accounts for the majority, comprising 62.2% of the datasets, highlighting its predominant role in this research area. Chinese (ZH) represents 17.8% of the total. The 'Other' group includes a variety of less-represented languages, such as Korean (KO), Polish (PL), Greek (EL), Japanese

Dataset	Language	Mental Disorder	Platform	Data Type	Annotation Procedure	Annotation Instrument	Label	Size	Availability	Citation
Pestian et al. (2012)	EN	SD	CLINIC	SD Notes	Manual	Krippendorff's α	15	1004	DUA	236
Uzuner et al. (2017)	EN	DP, PTSD, OCD, BD	CLINIC	Clinical Notes	Manual	-	4	816	DUA	45
Jackson et al. (2017)	EN	SMI	CLINIC	EHR(DS)	Manual	Text Hunter	50	37,211	RSTR	233
Low et al. (2010)	EN	MDD	CLINIC	Interview	Manual	LIFE	2	139	RSTR	313
Pestian et al. (2010)	EN	SD	CLINIC	SD Notes	Manual	Ontology	2	66	RSTR	322
Geraci et al. (2017)	EN	MDD & DD	CLINIC	EMR	Manual	DSM-IV	2	861	RSTR	88
Zhou et al. (2015)	EN	DP	CLINIC	EHR (DS)	Manual	-	3	1200	UNK	74
Meng et al. (2021)	EN	DP	CLINIC	EHR	Manual	ICD-9 & AD	2	10,148	UNK	53
Marrie et al. (2018)	EN	DP & AX	CLINIC	Interview	Manual	DSM-IV	-	308	UNK	51
Diederich et al. (2007)	EN	SZ	CLINIC	Essay	Manual	-	2	56	UNK	71
Poulin et al. (2014)	EN	SD	CLINIC	Clinical Notes	Manual	-	3	210	UNK	233
Saini et al. (2014)	EN	SZ, DP, BD, SD & Other	CLINIC	Interview & Clinical Notes	Manual	-	2	198	UNK	26
Milne et al. (2016)	EN	MHD	FORUM	Post	Manual	Fleiss's Kappa & Cohen's Kappa	4	1227	PUB	142
He et al. (2017)	EN	PTSD	FORUM	Essay	Manual	DSM-IV & CAP Scale	2	300	UNK	95
Tyshchenko (2018)	EN	DP & AX	FORUM	Post	Self-disclosure	-	2	16,975	UNK	55
Nguyen et al. (2014)	EN	DP, BD & SD	FORUM	Post	Self-disclosure	-	2	267,964	UNK	303
Aich et al. (2022)	EN	SZ & BD	COLLEGE	Interview	Manual	DSM-V & DSM-IV	3	644	DUA	11
Rude et al. (2004)	EN	DP	COLLEGE	Essay	Manual	BDI & IDD-L	3	124	UNK	1543
Rheault (2016)	EN	AX	DEBATE	Political Speech	Manual	-	2	4000	PUB	17
Hull et al. (2020)	EN	DP & AX	Telemedicine Platform	Message	Manual	PHQ-9 GAD-7	PHQ-9 GAD-7	10,718	DUA	43
Krishnamurti et al. (2022)	EN	PD	APP	Essay	Manual	EPDS	2	1,091	UNK	14
Nobles et al. (2018)	EN	DP & SD	Phone	SMS	Manual	Self-Disclosure	2	94	UNK	127
Howes et al. (2014)	EN	DP & AX	Online Therapy Chat	Dialogue	Mixed	MALLET & LIWC	2	882	UNK	78
Ringeval et al. (2019)	EN	DP & PTSD	Virtual Agent	Interview	Manual	PHQ-8	5	275	DUA	402
Schoene and Dethlefs (2016)	EN	DP & SD	MIXED	SD Notes & Articles	Manual	-	12	426	UNK	64
Gratch et al. (2014)	EN	DP, AX & PTSD	MIXED	Interview	Manual	-	-	621	DUA	759
Ghosh et al. (2020)	EN	DP & SDI	MIXED	SD Notes & Book	Manual	Cohen's Kappa	15	2393	PUB	45
Tasnim et al. (2022)	EN	DP & AX	mTurk	Interview	Manual	PHQ-9 GAD-7	PHQ-9 GAD-7	2674	UNK	29
Caicedo et al. (2020)	EN-ES	DP & SDI	Social Network & Forum	Post	Manual	Cohen's Kappa	4	102	PUB	24
Wu et al. (2020)	EN-ZH	MDD, MinDP, SZ, BD, AjD, DEM & Other	CLINIC	EHR(DS)	Manual	DSM-IV Sheehan Disability Scale	9	4,836	RSTR	122
Zou et al. (2023)	ZH	MDD	CLINIC	Interview	Manual	HAMD & PHQ-9	2	78	PUB	48
Jiang et al. (2022)	ZH	DP & AX	CLINIC	Interview	Manual	HAMD & HAMA	3	1025	PUB	9
Xu et al. (2025)	ZH	DP & AX	CLINIC	EMR	Manual	DSM-V & ICD-10	4	1,160	PUB	1
Mao et al. (2023)	ZH	DP	CLINIC	Interview	Manual	MADRS	2	113	DUA	5
Ive et al. (2024)	ZH	AX	CLINIC	EHR	Manual	ICD-9 & ICD-10	2	84,426	RSTR	0
Li et al. (2023)	ZH	DP & SD	CLINIC	Interview	Manual	HAMD	3	305	UNK	10
Shen et al. (2022)	ZH	DP	APP	Interview	Manual	SDS	2	162	PUB	146
Mo et al. (2024)	ZH	AX	Phn. Recording	Essay	Manual	GAD-7	3	227	UNK	11
Shin et al. (2022a)	KO	DP, AX & SD	CLINIC	Interview	Manual	PHQ-9, HDRS BAI & BSS	2	166	DUA	22
Figueroa-Barra et al. (2022)	KO	SZ & FEP	CLINIC	Interview	Manual	DSM-V, PANSS	3	133	RSTR	33
Wawer et al. (2022)	PL	SZ	CLINIC	Interview	Manual	ICD-10	2	94	UNK	20
Oh et al. (2024)	ES-CL	DP	CLINIC	Interview	Manual	DSM-V	8	451	UNK	2
Hämäläinen et al.	TH	DP	FORUM	Post	Manual	Key-word Search	2	944	PUB	18
Hiraga (2017)	JA	DP	FORUM	Post	Self-disclosure	-	2	108	UNK	29
Alghamdi et al. (2020)	AR	DP	FORUM	Post	Self-disclosure & Manual	DSM-5, PHQ-9 QIDS-SR	2	20,000	UNK	75
Mulholland and Quinn	EN	SD	Online DB	Song Lyrics	Manual	-	2	810	UNK	37
Zervopoulos et al. (2019)	EL	SDI	-	POEM	Manual	-	2	90	UNK	12
Reynolds and Wilson (2013)	CF	DP, AX & PA	Tablets	-	-	-	-	-	UNK	50

Table 1: Overview of non-social media, text-based mental health datasets. The table summarizes each dataset's language, mental disorder(s), platform, data form, annotation method, labeling instrument, availability, and citation count (till June 2025). *Abbreviations:* SD - Suicidal, DP - Depression, MinDP - Minor Depression, PD - Postnatal Depression, PTSD - Post Traumatic Stress Disorder, OCD - Obsessive Compulsive Disorder, SMI - Severe Mental Illness, MDD - Major Depressive Disorder, MHD - Mental Health Distress, AjD - Adjustment Disorder, BD - Bipolar Disorder, SZ - Schizophrenia, AX - Anxiety, SDI - Suicidal Ideation, FEP - First Episode Psychosis, DEM - Dementia, PA - Panic Attack, PUB - Public, DUA- Data Use Agreement, RSTR - Restricted, UNK - Unknown, DS - Discharge Summaries

(JA), Thai (TH), and Arabic (AR); collectively representing another 13.3%. Additionally, 6.4% of the datasets fall under the "Code-mixed" category, which includes multilingual combinations like EN-ES (English-Spanish), EN-ZH (English-Chinese), and ES-CL (Spanish-Chilean). While there is some linguistic variety, English datasets dominate in mental health research, indicating a need for more inclusive and multilingual dataset development.

Mental Health Disorder Distribution We consider the distribution of datasets for different mental health disorders and language diversity. In Figure 4, we present a heatmap with the distribution of mental health datasets across different disorders and languages. Since some datasets cover more than one disorder, the total count of disorders is higher than the number of datasets. To make things clearer, we group similar disorders; depression related conditions, grouped under 'DP' (including Depression (DP), Major Depressive Disorder (MDD), Minor Depression (MinDP), and Postnatal Depression (PD)) are the most studied, appearing in 33 datasets. Anxiety (AX) and suicidal ideation/-suicide(SD) follow with 12 and 11 datasets, respectively. Other less common disorders, such as Adjustment Disorder (AjD), Obsessive–compulsive disorder (OCD), Dysthymic Disorder (DD), and Dementia (DEM) are grouped under 'Others'; each accounts for 8 datasets. Schizophrenia (SZ) and First-episode Psychosis (FEP), grouped together as 'SZ', are analyzed in 5 datasets, while Post-Traumatic Stress Disorder (PTSD), and Bipolar Disorder (BD) are the least represented ones.

English datasets dominate the landscape, especially for depression (DP) and suicidal ideation (SD). Code-mixed and non-English data remain underrepresented across most disorders. This highlights a concentration on depression and a linguistic imbalance, emphasizing the need for more diverse mental health datasets.

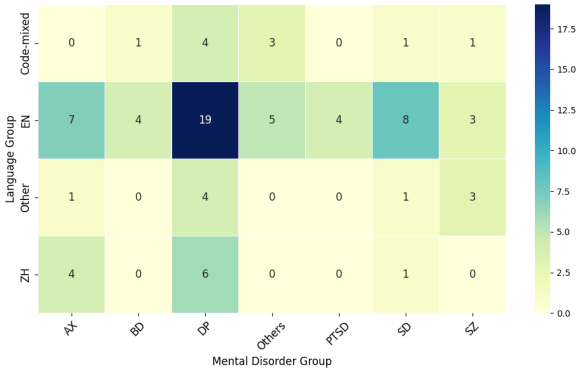

Figure 4: Heatmap showing the distribution of datasets across language groups and disorders.

Platform & Data Type Distribution Figure 5 shows the distribution of data types across four platform types: CLINIC, FORUM, MIXED (from multiple sources), and OTHER (including apps, virtual agents, online therapy chat, telemedicine platform and suicide notes). For simplicity, we group similar data types, like Electronic Health Records (EHR), Electronic Medical Records (EMR), clinical records (CR), and Discharge Summaries (DS) are grouped under 'EHR'; essays and questionnaires fall under 'questionnaires' category; while types like interviews and posts are kept distinct.

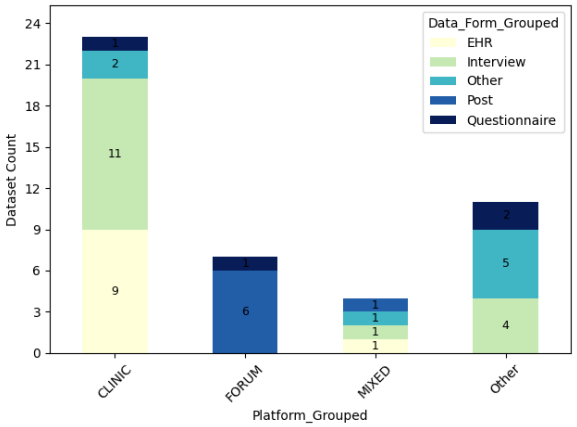

Figure 5: Distribution of data types across platforms.

As shown in Figure 5, clinical datasets mainly stem from interviews (48%) and EHRs (39%), reflecting structured clinical data sources. Forum datasets are heavily composed of user-generated posts (86%). Mixed platform datasets are more balanced, with one dataset each from interviews, posts, questionnaires, and other forms (each 25%). In the "Other" category, data forms are diverse: interviews and unconventional sources (like phone recordings, SMS, app data) both account for above 80%, and questionnaires make up the remaining.

Availability & Impact Figure 6 displays the distribution of citation counts (as per Google Scholar[5]) for mental health datasets categorized by their availability: datasets that are publicly available (PUB), datasets that require a Data Use Agreement (DUA), datasets labeled as Restricted (RSTR), and datasets with unspecified availability (UNK). Citation counts are normalized by year to account for differences in publication age. Datasets categorized as DUA and RSTR generally have higher citation counts compared to those that are publicly accessible or have unclear availability. Notably, DUA datasets show a wider range and a higher median citation rate, while RSTR datasets demonstrate a more consistent citation pattern. In contrast, PUB and UNK datasets have lower medians and are more tightly clustered around fewer citations, though some outliers exist.
The citation trends can also be partly explained by language distribution (Figure 7).

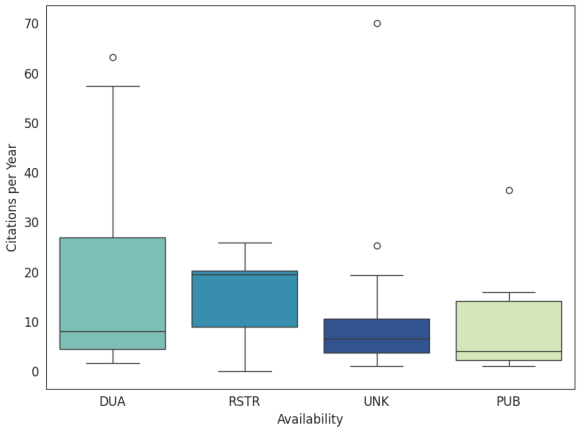

Figure 6: Distribution of citation counts by dataset availability type.

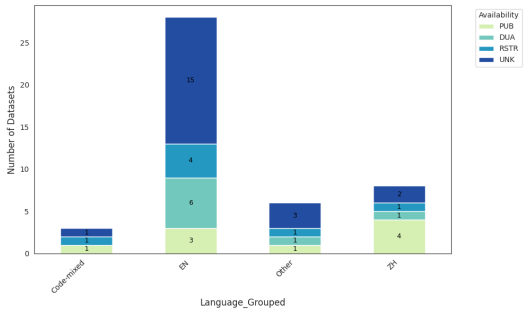

Figure 7: Availability of datasets across language groups.

[5] https://scholar.google.com/

Many English datasets are in the UNK category, while Chinese datasets are more publicly available (PUB). However, English-language PUB datasets rarely include depression-related data, and none originate from clinical settings. In contrast, English-language DUA datasets predominantly focus on depression and are collected in clinical contexts – this contributes to their higher credibility and research utility. Then again, most academic research is conducted in English; the higher citation rates of DUA and RSTR datasets reflect language dominance. Meanwhile, the lower citation rates for PUB datasets may be influenced by their association with less widely used languages like Chinese.

Demographic Representation Figure 8 shows how frequently different demographic attributes appear in the datasets shown in this survey.

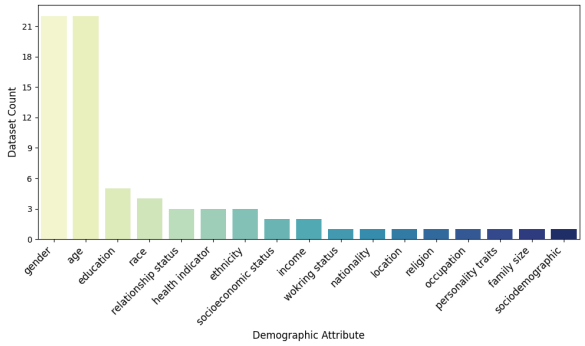

Figure 8: Frequency of demographic attributes across datasets.

Demographic attributes like gender and age dominate, appearing in over 24 datasets, while others like education, race, and ethnicity are moderately included. Fewer datasets report attributes like income, relationship status, or health indicators (e.g., height, weight, blood pressure). Some authors (Hiraga, 2017; Low et al., 2010) use demographic information to select negative samples that match the same profile type as the positive ones, which is rarely done in social media-based data, resulting in poor performance on the minority class (Rai et al., 2024; Cao et al., 2024). Some datasets focus on specific populations such as students (Shen et al., 2022; Rude et al., 2004; Nobles et al., 2018), veterans (Gratch et al., 2014; Poulin et al., 2014), seafarers (Mo et al., 2024), new mothers (Krishnamurti et al., 2022), or politicians (Rheault, 2016). Some have used demographic features in classification tasks, often reporting improved model performance (Shin et al., 2022b; Meng et al., 2021; Tasnim et al., 2022).

4. Tools & Techniques for Data Annotation & Labeling

Most datasets in our review are manually annotated by annotators or expert clinicians, following widely used tools and techniques. This section describes these datasets and the tools employed. Some datasets, such as those by Tyshchenko (2018), Nguyen et al. (2014), Hiraga (2017), and Alghamdi et al. (2020), are based on self-reported user content, typically collected from forums where individuals voluntarily share their thoughts under relevant discussion threads.

4.1. Clinical Diagnostic Instruments

Clinical diagnostic tools are assessments carried out by qualified healthcare professionals using standardized frameworks. They typically involve face-to-face evaluations, structured interviews, and expert judgment to ensure reliable and consistent mental health diagnoses. This systematic review highlights multiple studies that utilized clinical diagnostic tools for labeling and assessment.

The DSM-IV (Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition) (APA, 1994) and its updated version, the DSM-V (APA, 2013), offer standardized criteria for diagnosing a wide range of mental health conditions. Diagnoses based on DSM are often recorded during clinical visits and stored in EHRs, making them a key source of labeled data. Several works in this survey rely on DSM-IV for annotation, including Geraci et al. (2017), He et al. (2017), Marrie et al. (2018) and Wu et al. (2020), while studies such as Xu et al. (2025), Alghamdi et al. (2020), and Oh et al. (2024) utilize the updated DSM-V guidelines. Notably, Aich et al. (2022) incorporated both frameworks to account for changes in diagnostic criteria across versions.

The ICD (International Classification of Diseases), particularly ICD-9 and ICD-10 (World Health Organization, 1992), is used for coding clinical diagnoses, including mental and behavioral disorders. While the DSM focuses on mental health, the ICD covers all diseases. For instance, the code F32.9 in ICD-10 represents an unspecified single episode of major depressive disorder. In this survey, ICD-9 is used by Meng et al. (2021), while Xu et al. (2025) and Wawer et al. (2022) utilize ICD-10 for data annotation, and Ive et al. (2024) uses both editions to label anxiety.

The LIFE (Longitudinal Interval Follow-Up Evaluation) is a structured method used to monitor the long-term progression of psychiatric conditions, typically every six months, particularly within longitudinal studies (Keller et al., 1987). Porter et al. (2022) underlines that LIFE is used in both clinical

practice and research to assess the long-term impact of psychiatric disorders. In this survey, only Low et al. (2010) uses LIFE to classify the interviews.

The CAPS (Clinician-Administered PTSD Scale) is a structured interview for diagnosing PTSD. It assesses 20 core PTSD symptoms, onset, duration, impairment, and dissociative features related to a specific traumatic event and is considered the gold standard for PTSD evaluation (Blake et al., 1995). Clinicians in He et al. (2017) use both the DSM-IV PTSD module and the CAPS scale to annotate the dataset with binary labels for PTSD presence or absence.

The HAMD (Hamilton Depression Rating Scale) (Renemane and Vrublevska, 2021) is a clinician-administered scale used to assess depression severity through patient interviews, including 17 core items that rate depression from mild to severe. Similarly, HAMA (Hamilton Anxiety Rating Scale) (Matza et al., 2010) evaluates anxiety symptoms. In this survey, Zou et al. (2023), Li et al. (2023), Shin et al. (2022a) use HAMD to annotate the dataset for MDD or suicidal ideation, while Jiang et al. (2022) use both HAMD and HAMA for depression and anxiety labeling.

The MADRS (Montgomery-Åsberg Depression Rating Scale) (Müller et al., 2003) is a clinician-administered tool for assessing depression severity, emphasizing core symptoms. It is often used in clinical trials and is sensitive to changes in symptoms. Mao et al. (2023) uses MADRS to identify depression from transcribed interviews.

4.2. Screening Questionnaires and Self-Report Scales

Screening questionnaires and self-report scales are tools where individuals assess their own mental health by answering standardized questions. These are often based on validated clinical criteria for quickly screening for symptoms. Several studies included in this review employ screening questionnaires and self-report scales for mental health disorder identification.

Some studies in this review, Tasnim et al. (2022), Zou et al. (2023), Alghamdi et al. (2020), Hull et al. (2020) and Shin et al. (2022a) use the Patient Health Questionnaire (PHQ-9) (Kroenke et al., 2001) for dataset annotations and defining gold standard labels for depression. Ringeval et al. (2019) uses the older version of PHQ-9, PHQ-8, which consists of the same questions, excluding the suicidal ideation item (Kroenke et al., 2009). Rude et al. (2004) employs the Inventory to Diagnose Depression–Lifetime (IDD-L) (Zimmerman and Coryell, 1987) to analyze college students' essays, identifying depression through their reflections on past experiences. The Beck Depression Inventory (BDI-II) (Beck et al., 1996), a self-reported tool used to assess the intensity of depressive symptoms, Rude et al. (2004) employs both BDI-II and IDD-L for dataset annotation. In addition, the Beck Anxiety Inventory (BAI) (Beck et al., 1993) and the Beck Scale for Suicide Ideation (BSS) (Beck and Steer, 1991) are used to measure anxiety and suicidal thoughts, respectively, with Shin et al. (2022a) applying both to label transcribed interviews. In a study by Krishnamurti et al. (2022), the Edinburgh Postnatal Depression Scale (EPDS) (Cox et al., 1987) is used to identify depression in essays collected from pregnant women through an app. Wu et al. (2020) combines Sheehan Disability Scale (Sheehan et al., 1996) with DSM-IV guidelines to identify major depressive disorder (MDD), schizophrenia (SZ), bipolar disorder (BD), and other psychiatric conditions from discharge summaries, extracted from patients' EHRs; whereas Shen et al. (2022) uses Zung Self-Rating Depression Scale (SDS) (Tenti et al., 2022) to label depression from transcribed interviews collected through an app. Tasnim et al. (2022), Hull et al. (2020) and Mo et al. (2024) use the GAD-7 (Generalized Anxiety Disorder-7) (Spitzer et al., 2006) to classify anxiety from their respective datasets. Alghamdi et al. (2020) uses the Quick Inventory of Depressive Symptomatology- Self-Report (QIDS-SR) (Rush et al., 2003) tool to label depression in posts from a psychological forum in Arabic.

4.3. Manual or Contextual Labeling

Several datasets in this review do not explicitly mention the annotation tools used; however, they do indicate that trained annotators or clinicians conducted the labeling (Uzuner et al., 2017). In some instances, studies report inter-rater agreement measures like Cohen's Kappa (e.g. Milne et al. (2016), Ghosh et al. (2020), Caicedo et al. (2020)), Fleiss's Kappa (Milne et al., 2016), or Krippendorff's α (Pestian et al., 2012) to justify the labeling decisions, and occasionally a third annotator was used to resolve conflicts. A few datasets generate gold-standard labels using NLP-based methods such as Text Hunter (Jackson et al., 2017), custom ontologies (Pestian et al., 2010), or frameworks like MALLET and LIWC (Howes et al., 2014). Hämäläinen et al. uses keyword search (i.e., "depression", "anxiety") to label the datasets.

Few miscellaneous datasets contain song lyrics (Mulholland and Quinn), poems (Zervopoulos et al., 2019) from artists who have committed suicide, even ancient texts scraped from Babylonian tablets (Reynolds and Wilson, 2013); the authors do not mention using any labeling techniques.

4.4. Clinical Implications

Different labeling strategies in mental health datasets carry distinct clinical implications. Clinical diagnostic tools are widely considered the most reliable or gold standard (Arrow et al., 2023), offering standardized and consistent evaluations that support accuracy across clinicians[6]. However, these tools have limitations, including time constraints (e.g., typical sessions lasting 15–20 minutes), limited accessibility (shortage of clinicians), and missed opportunities for monitoring between appointments (often referred to as "clinical whitespace") (Stene-Larsen and Reneflot, 2019).

To address these gaps, self-report tools have become increasingly valuable. They capture patients' first-hand perspectives and are ideal for digital use (Arrow et al., 2023), especially when clinicians are unavailable, such as during clinical whitespace periods (Coppersmith et al., 2017) or while patients await intake appointments (Cruz et al., 2013). While self-reports offer efficiency and scalability, they should not replace in-person evaluations, as they are susceptible to social desirability effects, recall bias (Althubaiti, 2016), and trust issues. Instead, they should serve as supplementary tools (Arrow et al., 2023). In fact, research supports combining clinician ratings and self-reports for a more comprehensive understanding of patient conditions (Uher et al., 2012).

When clinical labels are unavailable, manual annotation, together with annotator agreement techniques, is often used to scale data labeling. While useful, these methods can suffer from reduced accuracy if not guided by trained professionals (Sylolypavan et al., 2023). Overall, each labeling method presents trade-offs between scalability, clinical rigor, and data quality.

5. Computational Modeling

The studies reviewed in this systematic review have been used to develop mental health classification and prediction systems using a broad spectrum of approaches (Harrigian et al., 2021; Ríssola et al., 2020; Bucur et al., 2025b). While a thorough review of computational approaches is outside the scope of this survey, in this section, we provide the reader with a brief summary of the computational models applied to the task. Exploring computational models allows us to better understand prevailing methodological trends, evaluate benchmarks, and identify potential limitations or biases inherent in different approaches. This insight is vital for ensuring replicability and shaping future research directions.

[6] https://www.verywellmind.com/what-is-reliability-2795786

Early works use manual or qualitative techniques while later studies applied traditional machine learning (ML) models like support vector machine (SVM) (Jackson et al., 2017; Pestian et al., 2010, 2012; He et al., 2017; Uzuner et al., 2017; Tyshchenko, 2018; Nobles et al., 2018; Diederich et al., 2007), Logistic Regression (Pestian et al., 2012; Hiraga, 2017; Li et al., 2023; Aich et al., 2022; Meng et al., 2021; Howes et al., 2014; Mo et al., 2024; Tasnim et al., 2022), Decision Trees (He et al., 2017; Zhou et al., 2015; Shen et al., 2022), Adaboost (Pestian et al., 2010; Uzuner et al., 2017; Jiang et al., 2022; Mo et al., 2024), XGboost (Wawer et al., 2022; Oh et al., 2024; Mao et al., 2023), and Naive Bayes (Schoene and Dethlefs, 2016; Ghosh et al., 2020; He et al., 2017; Hiraga, 2017). Traditional ML classifiers are often supported by feature engineering techniques such as Linguistic Inquiry and Word Count (LIWC) (Boyd et al., 2022) analysis (Rude et al., 2004; Li et al., 2023; Howes et al., 2014; Nobles et al., 2018), Term Frequency-Inverse Document Frequency (tf-idf) (Ive et al., 2024; Geraci et al., 2017; Nobles et al., 2018), or keyword extraction. Deep learning models, particularly Convolutional Neural Networks (CNN) (Ghosh et al., 2020; Uzuner et al., 2017; Tyshchenko, 2018), Long Short-Term Memory (LSTM) (Hämäläinen et al.; Uzuner et al., 2017; Mao et al., 2023; Krishnamurti et al., 2022), and Gated Recurrent Unit (GRU) (Jackson et al., 2017; Ghosh et al., 2021) are also used to capture nuanced patterns in unstructured text.

Some datasets have been used to develop interpretable models or hybrid systems that combine rule-based methods with ML (e.g., Conditional Random Field (CRF) (Pestian et al., 2012; Wu et al., 2020), TextHunter + ConText (Jackson et al., 2017)). More recent work adopts transformer-based models like Clinical-BigBird (Ive et al., 2024), MentalBERT (Aich et al., 2022), MentalRoBERTa (Aich et al., 2022), and large language models (LLMs) like Qwen2-72B (Xu et al., 2025), highlighting a shift from fine-tuning pre-trained language models to using instruction-tuned LLMs for domain-specific tasks with little or no fine-tuning.

6. Identified Trends & Research Gaps

This section summarizes the main findings of this review along with suggestions for future directions. Unsurprisingly, most available datasets are in English. Chinese follows in prevalence, while there are very few datasets in other languages, such as Korean, Arabic, or Polish. There are no datasets in low-resource languages. This gap may be partly due to the complexity of creating these datasets, particularly in clinical contexts. Collecting data often requires patient consent, approval from ethics

boards, and substantial manual effort to annotate datasets.

Another important finding is that depression is the most studied mental health disorder within these datasets. Studies either focus exclusively on depression or include related conditions such as anxiety and bipolar disorder. Other conditions, such as PTSD, schizophrenia, and eating disorders, are not well represented. This highlights the need to develop datasets that encompass a broader range of mental health issues. On the other hand, compared to the number of social media datasets (Garg, 2023; Bucur et al., 2026), clinical datasets are scarce due to their sensitive nature. Releasing more public or even agreement-governed clinical datasets could expand research opportunities and improve reproducibility.

Moreover, the annotation procedures and labeling techniques vary across the datasets. Some use clinical diagnostic tools such as DSM and ICD, while others rely on self-reported questionnaires like PHQ and BDI. However, most papers do not explain why a specific tool was chosen, indicating a lack of reporting standards (Chancellor and De Choudhury, 2020). Additionally, some datasets are labeled manually or through keyword searches, and a few do not clarify their labeling methods at all. This lack of transparency can make it difficult to trust or replicate the research findings.

Advancements in NLP offer promising solutions to the challenges posed by inconsistent and opaque labeling practices in mental health datasets. With instruction-tuned language models and structured prompting techniques, NLP can support (semi-)automation and standardization of labeling based on formal diagnostic criteria such as DSM or ICD. Approaches like chain-of-thought (CoT) prompting (Wei et al., 2022) can emulate clinical reasoning and even improve labeling accuracy (Shi et al., 2024), while chain-of-empathy (CoE) (Lee et al., 2023) frameworks are well-suited for understanding emotionally nuanced texts, such as therapy transcripts or suicide notes. These strategies can enhance both the consistency and interpretability of labels, enabling the creation of scalable and clinically relevant datasets, even when direct clinician involvement is limited.

7. Conclusion & Future Directions

In this paper, we presented the first systematic review of mental health free-text datasets beyond social media. We analyze these datasets with respect to different dimensions such as their distribution in terms of languages and mental health disorders, the data types included in them, and their availability.

We revisit the research questions (RQ) posed in the introduction (Section 1) and present the main findings of our review below:

RQ_1: What non-social media free-text based datasets are available for mental health research, and how do they vary by source, structure, and population?

> **RQ_1 Findings:** *Most datasets are in English and primarily focus on depression, while other languages and mental health conditions remain underrepresented. More Chinese datasets are publicly available than English ones. In terms of data collection, some actively balance samples by selecting equal numbers from the control group matching the same profile type. The nature of text-based data also varies widely, ranging from phone recordings and messages to essays, interview transcripts, and suicide notes, mostly collected in clinical settings.*

RQ_2: How are mental health conditions defined and labeled in these datasets, and what are the clinical implications of these labeling methods?

> **RQ_2 Findings:** *Clinical diagnostic tools, screening questionnaires, self-report scales, and manual contextual labeling have been used to annotate mental health disorders. These labeling techniques can complement one another and collectively serve as gold standards.*

RQ_3: What factors contribute to the popularity and adoption of non-social media free-text based mental health datasets in research?

> **RQ_3 Findings:** *Along with dataset availability, factors such as the data collection setting (clinical vs. informal), language, data type (interviews, EHRs, essays), mental health disorders covered, and the credibility and reliability of the labeling and annotation methods all contribute to the popularity and adoption of non-social media free-text based mental health datasets in research.*

There is significant room for improvement in the development and use of these types of datasets. Expanding dataset creation to include more languages and geographic regions can help address current imbalances. Additionally, increasing the representation of a broader range of mental health disorders, not just depression, would make research findings more comprehensive. Establishing clear and consistent labeling methods is essential for ensuring reproducibility and building trust in research outcomes, and advances in NLP, particularly in explainable AI, can help bridge these gaps. Lastly, improving access to high-quality datasets, while maintaining ethical and privacy standards, can facilitate collaborative research.

Limitations

In this systematic review, we used the PRISMA methodology and conducted a comprehensive literature search using the Publish or Perish software. This approach ensured thorough coverage of datasets on mental health disorders across both the NLP and clinical domains. Although we used mental health-related terms for our keyword searches, it is possible that studies may use alternative terms or less common phrases to describe mental health disorders, which may not have been included in our search strategy. As a result, some relevant works might have been unintentionally overlooked. Additionally, the search queries were in English, which may have excluded relevant non-English publications. Finally, Publish or Perish does not index certain databases, such as CINAHL, potentially limiting coverage of some clinically oriented studies.

Ethics Statement

This systematic review draws from previously published studies to guide future research in identifying mental health disorders beyond social media. While we present the available datasets used for this purpose in our review, we did not make any attempts to build prediction models using this data. We acknowledge that detecting early signs of mental health disorders requires adherence to ethical protocols. Moreover, misuse of these sensitive data or of the models trained on the data can lead to stigmatization or harm to individuals with mental health disorders (Chancellor et al., 2019b).

Acknowledgements

We would like to thank the anonymous reviewers for their constructive feedback and thoughtful suggestions, which helped improve the clarity and quality of this work. We are also grateful to our collaborators for their valuable contributions and insightful discussions throughout the project. Finally, we acknowledge the creators and maintainers of the datasets studied in our review for developing and curating these resources, thus enabling and advancing research in this domain.

References

Asma Abdulsalam and Areej Alhothali. 2024. Suicidal ideation detection on social media: A review of machine learning methods. *Social Network Analysis and Mining*, 14(1):188.

Eman Adil Jaafar and Haya Abdul-Salam Jasim. 2022. A corpus-based stylistic analysis of online suicide notes retrieved from reddit. *Cogent Arts & Humanities*, 9(1):2047434.

Ankit Aich, Avery Quynh, Varsha Badal, Amy Pinkham, Philip Harvey, Colin Depp, and Natalie Parde. 2022. Towards intelligent clinically-informed language analyses of people with bipolar disorder and schizophrenia. In *Findings of EMNLP*.

Norah Saleh Alghamdi, Hanan A Hosni Mahmoud, Ajith Abraham, Samar Awadh Alanazi, and Laura García-Hernández. 2020. Predicting depression symptoms in an arabic psychological forum. *IEEE access*, 8:57317–57334.

Alaa Althubaiti. 2016. Information bias in health research: definition, pitfalls, and adjustment methods. *Journal of multidisciplinary healthcare*, pages 211–217.

APA. 1994. *Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (DSM-IV)*. American Psychiatric Association.

APA. 2013. *Diagnostic and Statistical Manual of Mental Disorders*, 5th ed. edition. American Psychiatric Publishing.

Kaitlyn Arrow, Philip Resnik, Hanna Michel, Christopher Kitchen, Chen Mo, Shuo Chen, Carol Espy-Wilson, Glen Coppersmith, Colin Frazier, and Deanna L Kelly. 2023. Evaluating the use of online self-report questionnaires as clinically valid mental health monitoring tools in the clinical whitespace. *Psychiatric Quarterly*, 94(2):221–231.

Rahn K Bailey, Tejas C Patel, Jaymie Avenido, Milapkumar Patel, Mohammad Jaleel, Narviar C Barker, Jahanzeb Ali Khan, Shahid All, and Shagufta Jabeen. 2011. Suicide: current trends. *Journal of the National Medical Association*, 103(7):614–617.

Ilya Baryshnikov, Tom Rosenström, Pekka Jylhä, Maria Vuorilehto, Mikael Holma, Irina Holma, Kirsi Riihimäki, Gregory K Brown, Maria A Oquendo, and Erkki T Isometsä. 2020. Role of hopelessness in suicidal ideation among patients with depressive disorders. *The Journal of clinical psychiatry*, 81(2):8339.

Aaron T Beck, Norman Epstein, Gary Brown, and Robert Steer. 1993. Beck anxiety inventory. *Journal of consulting and clinical psychology*.

Aaron T Beck and Robert A Steer. 1991. Manual for the beck scale for suicide ideation. *San Antonio, TX: Psychological Corporation*, 63.

Aaron T Beck, Robert A Steer, and Gregory Brown. 1996. Beck depression inventory–ii. *Psychological assessment*.

Dudley David Blake, Frank W Weathers, Linda M Nagy, Danny G Kaloupek, Fred D Gusman, Dennis S Charney, and Terence M Keane. 1995. The development of a clinician-administered ptsd scale. *Journal of traumatic stress*, 8:75–90.

Nick Boettcher. 2021. Studies of depression and anxiety using reddit as a data source: scoping review. *JMIR mental health*, 8(11):e29487.

Ryan L Boyd, Ashwini Ashokkumar, Sarah Seraj, and James W Pennebaker. 2022. The development and psychometric properties of liwc-22. *Austin, TX: University of Texas at Austin*, 10(1-47):6.

Ana-Maria Bucur, Andreea Moldovan, Krutika Parvatikar, Marcos Zampieri, Ashiqur Khudabukhsh, and Liviu P Dinu. 2025a. Datasets for depression modeling in social media: An overview. In *Proceedings of CLPsych*.

Ana-Maria Bucur, Andreea-Codrina Moldovan, Krutika Parvatikar, Marcos Zampieri, Ashiqur R. KhudaBukhsh, and Liviu P. Dinu. 2025b. On the state of nlp approaches to modeling depression in social media: A post-covid-19 outlook. *IEEE Journal of Biomedical and Health Informatics*, 29(6):4439–4451.

Ana-Maria Bucur, Marcos Zampieri, Tharindu Ranasinghe, and Fabio Crestani. 2026. A survey on multilingual mental disorders detection from social media data. In *Proceedings of EACL*.

Roberto Wellington Acuña Caicedo, José Manuel Gómez Soriano, and Héctor Andrés Melgar Sasieta. 2020. Assessment of supervised classifiers for the task of detecting messages with suicidal ideation. *Heliyon*, 6(8).

Rafael A Calvo, David N Milne, M Sazzad Hussain, and Helen Christensen. 2017. Natural language processing in mental health applications using non-clinical texts. *Natural Language Engineering*, 23(5):649–685.

Yuchen Cao, Jianglai Dai, Zhongyan Wang, Yeyubei Zhang, Xiaorui Shen, Yunchong Liu, and Yexin Tian. 2024. Machine learning approaches for mental illness detection on social media: A systematic review of biases and methodological challenges. *arXiv preprint arXiv:2410.16204*.

Centers for Disease Control and Prevention. 2025. Suicidal thoughts and behavior. https://www.cdc.gov/mental-health/about-data/suicidal-thoughts-and-behavior.html. Accessed 22 February 2026.

Stevie Chancellor, Eric PS Baumer, and Munmun De Choudhury. 2019a. 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*, 3(CSCW):1–32.

Stevie Chancellor, Michael L Birnbaum, Eric D Caine, Vincent MB Silenzio, and Munmun De Choudhury. 2019b. A taxonomy of ethical tensions in inferring mental health states from social media. In *Proceedings of FACCT*.

Stevie Chancellor and Munmun De Choudhury. 2020. Methods in predictive techniques for mental health status on social media: a critical review. *NPJ digital medicine*, 3(1):43.

Moumita Chatterjee, Poulomi Samanta, Piyush Kumar, and Dhrubasish Sarkar. 2022. Suicide ideation detection using multiple feature analysis from twitter data. In *IEEE DELCON*.

Glen Coppersmith, Casey Hilland, Ophir Frieder, and Ryan Leary. 2017. Scalable mental health analysis in the clinical whitespace via natural language processing. In *IEEE BHI*.

John L Cox, Jeni M Holden, and Ruth Sagovsky. 1987. Detection of postnatal depression: development of the 10-item edinburgh postnatal depression scale. *The British journal of psychiatry*, 150(6):782–786.

Mario Cruz, Debra L Roter, Robyn F Cruz, Melissa Wieland, Susan Larson, Lisa A Cooper, and Harold Alan Pincus. 2013. Appointment length, psychiatrists' communication behaviors, and medication management appointment adherence. *Psychiatric Services*, 64(9):886–892.

Janet R Cummings, Hefei Wen, and Benjamin G Druss. 2013. Improving access to mental health services for youth in the united states. *Jama*, 309(6):553–554.

Joachim Diederich, Aqeel Al-Ajmi, and Peter Yellowlees. 2007. Ex-ray: Data mining and mental health. *Applied Soft Computing*, 7(3):923–928.

Alicia Figueroa-Barra, Daniel Del Aguila, Mauricio Cerda, Pablo A Gaspar, Lucas D Terissi, Manuel Durán, and Camila Valderrama. 2022. Automatic language analysis identifies and predicts schizophrenia in first-episode of psychosis. *Schizophrenia*, 8(1):53.

Muskan Garg. 2023. Mental health analysis in social media posts: a survey. *Archives of Computational Methods in Engineering*, 30(3):1819.

Joseph Geraci, Pamela Wilansky, Vincenzo de Luca, Anvesh Roy, James L Kennedy, and John Strauss. 2017. Applying deep neural networks to unstructured text notes in electronic medical records for phenotyping youth depression. *BMJ Ment Health*, 20(3):83–87.

Sayan Ghosh, Dylan Baker, David Jurgens, and Vinodkumar Prabhakaran. 2021. Detecting cross-geographic biases in toxicity modeling on social media. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 313–328, Online. Association for Computational Linguistics.

Soumitra Ghosh, Asif Ekbal, and Pushpak Bhattacharyya. 2020. Cease, a corpus of emotion annotated suicide notes in english. In *Proceedings of LREC*.

Maria J Grant and Andrew Booth. 2009. A typology of reviews: an analysis of 14 review types and associated methodologies. *Health information & libraries journal*, 26(2):91–108.

Jonathan Gratch, Ron Artstein, Gale Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood, Jill Boberg, David DeVault, Stacy Marsella, David Traum, Skip Rizzo, and Louis-Philippe Morency. 2014. The distress analysis interview corpus of human and computer interviews. In *Proceedings of LREC*.

Mika Hämäläinen, Pattama Patpong, Khalid Alnajjar, Niko Partanen, and Jack Rueter. Detecting depression in Thai blog posts: a dataset and a baseline. In *Proceedings of W-NUT*.

Keith Harrigian, Carlos Aguirre, and Mark Dredze. 2021. On the state of social media data for mental health research. In *Proceedings of CLPsych*.

Qiwei He, Bernard P Veldkamp, Cees AW Glas, and Theo de Vries. 2017. Automated assessment of patients' self-narratives for posttraumatic stress disorder screening using natural language processing and text mining. *Assessment*, 24(2):157–172.

Ronald D Hester. 2017. Lack of access to mental health services contributing to the high suicide rates among veterans. *International journal of mental health systems*, 11(1):47.

Misato Hiraga. 2017. Predicting depression for japanese blog text. In *Proceedings of ACL*.

Christine Howes, Matthew Purver, and Rose McCabe. 2014. Linguistic indicators of severity and progress in online text-based therapy for depression. In *Proceedings of CLPsych*.

Thomas D Hull, Matteo Malgaroli, Philippa S Connolly, Seth Feuerstein, and Naomi M Simon. 2020. Two-way messaging therapy for depression and anxiety: longitudinal response trajectories. *BMC psychiatry*, 20(1):297.

Institute for Health Metrics and Evaluation. 2024. Global burden of disease study 2021 (gbd 2021) results. https://vizhub.healthdata.org/gbd-results/. Online database. Seattle, WA. Accessed 13 August 2025.

Md Rafiqul Islam, Muhammad Ashad Kabir, Ashir Ahmed, Abu Raihan M Kamal, Hua Wang, and Anwaar Ulhaq. 2018. Depression detection from social network data using machine learning techniques. *Health information science and systems*, 6:1–12.

Julia Ive, Paulina Bondaronek, Vishal Yadav, Daniel Santel, Tracy Glauser, Tina Cheng, Jeffrey R Strawn, Greeshma Agasthya, Jordan Tschida, Sanghyun Choo, et al. 2024. A data-centric approach to detecting and mitigating demographic bias in pediatric mental health text: A case study in anxiety detection. *arXiv preprint arXiv:2501.00129*.

Richard G Jackson, Rashmi Patel, Nishamali Jayatilleke, Anna Kolliakou, Michael Ball, Genevieve Gorrell, Angus Roberts, Richard J Dobson, and Robert Stewart. 2017. Natural language processing to extract symptoms of severe mental illness from clinical text: the clinical record interactive search comprehensive data extraction (cris-code) project. *BMJ open*, 7(1):e012012.

Yueqi Jiang, Ziyang Zhang, and Xiao Sun. 2022. Mmda: A multimodal dataset for depression and anxiety detection. In *Proceedings of ICPR*.

Martin B Keller, Philip W Lavori, Barbara Friedman, Eileen Nielsen, Jean Endicott, Pat McDonald-Scott, and Nancy C Andreasen. 1987. The longitudinal interval follow-up evaluation: A comprehensive method for assessing outcome in prospective longitudinal studies. *Archives of general psychiatry*, 44(6):540–548.

D Sami Khafaga, Maheshwari Auvdaiappan, K Deepa, Mohamed Abouhawwash, and F Khalid Karim. 2023. Deep learning for depression detection using twitter data. *Intelligent Automation & Soft Computing*, 36(2):1301–1313.

Tamar Krishnamurti, Kristen Allen, Laila Hayani, Samantha Rodriguez, and Alexander L Davis. 2022. Identification of maternal depression risk from natural language collected in a mobile health app. *Procedia computer science*, 206:132–140.

Kurt Kroenke, Robert L Spitzer, and Janet BW Williams. 2001. The phq-9: validity of a brief depression severity measure. *Journal of general internal medicine*, 16(9):606–613.

Kurt Kroenke, Tara W Strine, Robert L Spitzer, Janet BW Williams, Joyce T Berry, and Ali H Mokdad. 2009. The phq-8 as a measure of current depression in the general population. *Journal of Affective Disorders*, 114(1-3):163–173.

Yoon Kyung Lee, Inju Lee, Minjung Shin, Seoyeon Bae, and Sowon Hahn. 2023. Chain of empathy: Enhancing empathetic response of large language models based on psychotherapy models. *arXiv preprint arXiv:2311.04915*.

Tim MH Li, Jie Chen, Framenia OC Law, Chun-Tung Li, Ngan Yin Chan, Joey WY Chan, Steven WH Chau, Yaping Liu, Shirley Xin Li, Jihui Zhang, et al. 2023. Detection of suicidal ideation in clinical interviews for depression using natural language processing and machine learning: cross-sectional study. *JMIR medical informatics*, 11(1):e50221.

Lu-Shih Alex Low, Namunu C Maddage, Margaret Lech, Lisa B Sheeber, and Nicholas B Allen. 2010. Detection of clinical depression in adolescents' speech during family interactions. *IEEE transactions on biomedical engineering*, 58(3):574–586.

Frances L Lynch, Edward L Peterson, Christine Y Lu, Yong Hu, Rebecca C Rossom, Beth E Waitzfelder, Ashli A Owen-Smith, Samuel Hubley, Deepak Prabhakar, L Keoki Williams, et al. 2020. Substance use disorders and risk of suicide in a general us population: a case control study. *Addiction science & clinical practice*, 15(1):14.

Kaining Mao, Deborah Baofeng Wang, Tiansheng Zheng, Rongqi Jiao, Yanhui Zhu, Bin Wu, Lei Qian, Wei Lyu, Jie Chen, and Minjie Ye. 2023. Analysis of automated clinical depression diagnosis in a chinese corpus. *IEEE Transactions on Biomedical Circuits and Systems*, 17(5):1135–1152.

Ruth Ann Marrie, Lesley Graff, John R Walker, John D Fisk, Scott B Patten, Carol A Hitchon, Lisa M Lix, James Bolton, Jitender Sareen, Alan Katz, et al. 2018. Effects of psychiatric comorbidity in immune-mediated inflammatory disease: protocol for a prospective study. *JMIR Research Protocols*, 7(1):e8794.

Louis S Matza, Robert Morlock, Chris Sexton, Karen Malley, and Douglas Feltner. 2010. Identifying ham-a cutoffs for mild, moderate, and severe generalized anxiety disorder. *International journal of methods in psychiatric research*, 19(4):223–232.

Yiwen Meng, William Speier, Michael Ong, and Corey W. Arnold. 2021. Hcet: Hierarchical clinical embedding with topic modeling on electronic health records for predicting future depression. *IEEE Journal of Biomedical and Health Informatics*, 25(4):1265–1272.

David N Milne, Glen Pink, Ben Hachey, and Rafael A Calvo. 2016. Clpsych 2016 shared task: Triaging content in online peer-support forums. In *Proceedings of the third workshop on computational linguistics and clinical psychology*, pages 118–127.

Haimiao Mo, Siu Cheung Hui, Xiao Liao, Yuchen Li, Wei Zhang, and Shuai Ding. 2024. A multimodal data-driven framework for anxiety screening. *IEEE Transactions on Instrumentation and Measurement*, 73:1–13.

Matthew Mulholland and Joanne Quinn. Suicidal tendencies: The automatic classification of suicidal and non-suicidal lyricists using nlp. In *Proceedings of IJCNLP*.

Matthias J. Müller, Hubertus Himmerich, Barbara Kienzle, and Armin Szegedi. 2003. Differentiating moderate and severe depression using the montgomery–Åsberg depression rating scale (madrs). *Journal of Affective Disorders*, 77(3):255–260.

Thin Nguyen, Dinh Phung, Bo Dao, Svetha Venkatesh, and Michael Berk. 2014. Affective and content analysis of online depression communities. *IEEE Transactions on Affective Computing*, 5(3):217–226.

Jennifer Nicholas, Sandersan Onie, and Mark E Larsen. 2020. Ethics and privacy in social media research for mental health. *Current psychiatry reports*, 22:1–7.

Alicia L Nobles, Jeffrey J Glenn, Kamran Kowsari, Bethany A Teachman, and Laura E Barnes. 2018. Identification of imminent suicide risk among young adults using text messages. In *Proceedings of CHI*.

Jihoon Oh, Taekgyu Lee, Eun Su Chung, Hyonsoo Kim, Kyongchul Cho, Hyunkyu Kim, Jihye Choi, Hyeon-Hee Sim, Jongseo Lee, In Young Choi, et al. 2024. Development of depression detection algorithm using text scripts of routine psychiatric interview. *Frontiers in psychiatry*, 14:1256571.

Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kıcıman. 2019. Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in big data*, 2:13.

Matthew J Page, Joanne E McKenzie, Patrick M Bossuyt, Isabelle Boutron, Tammy C Hoffmann, Cynthia D Mulrow, Larissa Shamseer, Jennifer M Tetzlaff, Elie A Akl, Sue E Brennan, et al. 2021. The prisma 2020 statement: an updated guideline for reporting systematic reviews. *bmj*, 372.

Michelle Pazdur, Dunja Tutus, and Ann-Christin Haag. 2025. Risk factors for problematic social media use in youth: A systematic review of longitudinal studies. *Adolescent Research Review*, pages 1–17.

John Pestian, Henry Nasrallah, Pawel Matykiewicz, Aurora Bennett, and Antoon Leenaars. 2010. Suicide note classification using natural language processing: A content analysis. *Biomedical informatics insights*, 3:BII–S4706.

John P Pestian, Pawel Matykiewicz, Michelle Linn-Gust, Brett South, Ozlem Uzuner, Jan Wiebe, K Bretonnel Cohen, John Hurdle, and Christopher Brew. 2012. Sentiment analysis of suicide notes: A shared task. *Biomedical informatics insights*, 5:BII–S9042.

Richard J. Porter, Will Moot, Maree L. Inder, Marie T. Crowe, Katie M. Douglas, Janet D. Carter, and Christopher Frampton. 2022. Validation of the longitudinal interval follow-up evaluation for the long-term measurement of mood symptoms in bipolar disorder. *Brain Sciences*, 12(12).

Chris Poulin, Brian Shiner, Paul Thompson, Linas Vepstas, Yinong Young-Xu, Benjamin Goertzel, Bradley Watts, Laura Flashman, and Thomas McAllister. 2014. Predicting the risk of suicide by analyzing the text of clinical notes. *PloS one*, 9(1):e85733.

Sunny Rai, Elizabeth C. Stade, Salvatore Giorgi, Ashley Francisco, Lyle H. Ungar, Brenda Curtis, and Sharath C. Guntuku. 2024. Key language markers of depression on social media depend on race. *Proceedings of the National Academy of Sciences*, 121(14):e2319837121.

Nishat Raihan, Sadiya Sayara Chowdhury Puspo, Ana-Maria Bucur, Stevie Chancellor, and Marcos Zampieri. 2026. Large language models for mental health: A multilingual evaluation. In *Proceedings of LowResLM*.

Nishat Raihan, Sadiya Sayara Chowdhury Puspo, Shafkat Farabi, Ana-Maria Bucur, Tharindu Ranasinghe, and Marcos Zampieri. 2024. Mentalhelp: A multi-task dataset for mental health in social media. In *Proceedings of LREC-COLING*.

Lubova Renemane and Jelena Vrublevska. 2021. Chapter 17 - hamilton depression rating scale: Uses and applications. In Colin R. Martin, Lan-Anh Hunter, Vinood B. Patel, Victor R. Preedy, and Rajkumar Rajendram, editors, *The Neuroscience of Depression*, pages 175–183. Academic Press.

Edward H Reynolds and James V Kinnier Wilson. 2013. Depression and anxiety in babylon. *Journal of the Royal Society of Medicine*, 106(12):478–481.

Ludovic Rheault. 2016. Expressions of anxiety in political texts. In *Proceedings of NLP-CSS*.

William S Rholes, John H Riskind, and Brian Neville. 1985. The relationship of cognitions and hopelessness to depression and anxiety. *Social Cognition*, 3(1):36–50.

Fabien Ringeval, Björn Schuller, Michel Valstar, Nicholas Cummins, Roddy Cowie, Leili Tavabi, Maximilian Schmitt, Sina Alisamir, Shahin Amiriparian, Eva-Maria Messner, et al. 2019. Avec 2019 workshop and challenge: state-of-mind, detecting depression with ai, and cross-cultural affect recognition. In *Proceedings of AVEC*.

Stephanie Rude, Eva-Maria Gortner, and James Pennebaker. 2004. Language use of depressed and depression-vulnerable college students. *Cognition & Emotion*, 18(8):1121–1133.

A.John Rush, Madhukar H Trivedi, Hicham M Ibrahim, Thomas J Carmody, Bruce Arnow, Daniel N Klein, John C Markowitz, Philip T Ninan, Susan Kornstein, Rachel Manber, Michael E Thase, James H Kocsis, and Martin B Keller. 2003. The 16-item quick inventory of depressive symptomatology (qids), clinician rating (qids-c), and self-report (qids-sr): a psychometric evaluation in patients with chronic major depression. *Biological Psychiatry*, 54(5):573–583.

Esteban Andrés Ríssola, Mohammad Aliannejadi, and Fabio Crestani. 2020. Beyond modelling: understanding mental disorders in online social media. In *Proc. of ECIR*, pages 296–310. Springer.

Pooja Saini, David While, Khatidja Chantler, Kirsten Windfuhr, and Navneet Kapur. 2014. Assessment and management of suicide risk in primary care. *Crisis*.

Selim Sametoğlu, DHM Pelt, Johannes C Eichstaedt, Lyle H Ungar, and Meike Bartels. 2024.

The value of social media language for the assessment of wellbeing: A systematic review and meta-analysis. *The Journal of Positive Psychology*, 19(3):471–489.

Annika Marie Schoene and Nina Dethlefs. 2016. Automatic identification of suicide notes from linguistic and sentiment features. In *Proceedings of SIGHUM, LaTeCH*.

David V Sheehan, Kathy Harnett-Sheehan, and BA8923116 Raj. 1996. The measurement of disability. *International clinical psychopharmacology*, 11:89–95.

Ying Shen, Huiyu Yang, and Lin Lin. 2022. Automatic depression detection: An emotional audio-textual corpus and a gru/bilstm-based model. In *IEEE ICASSP*.

Elysia Shi, Adithri Manda, London Chowdhury, Runeema Arun, Kevin Zhu, and Michael Lam. 2024. Enhancing depression diagnosis with chain-of-thought prompting. *arXiv preprint arXiv:2408.14053*.

Daun Shin, Kyungdo Kim, Seung-Bo Lee, Changwoo Lee, Ye Seul Bae, Won Ik Cho, Min Ji Kim, C Hyung Keun Park, Eui Kyu Chie, Nam Soo Kim, et al. 2022a. Detection of depression and suicide risk based on text from clinical interviews using machine learning: possibility of a new objective diagnostic marker. *Frontiers in psychiatry*, 13:801301.

Jamin Shin, Hangyeol Yu, Hyeongdon Moon, Andrea Madotto, and Juneyoung Park. 2022b. Dialogue summaries as dialogue states (DS2), template-guided summarization for few-shot dialogue state tracking. In *Findings of ACL*.

Ruba Skaik and Diana Inkpen. 2020. Using social media for mental health surveillance: a review. *ACM Computing Surveys*, 53(6):1–31.

Robert L. Spitzer, Kurt Kroenke, Janet B. W. Williams, and Bernd Löwe. 2006. A brief measure for assessing generalized anxiety disorder: The gad-7. *Archives of Internal Medicine*, 166(10):1092–1097.

Kim Stene-Larsen and Anne Reneflot. 2019. Contact with primary and mental health care prior to suicide: a systematic review of the literature from 2000 to 2017. *Scandinavian journal of public health*, 47(1):9–17.

Aneeta Sylolypavan, Derek Sleeman, Honghan Wu, and Malcolm Sim. 2023. The impact of inconsistent human annotations on ai driven clinical decision making. *NPJ Digital Medicine*, 6(1):26.

Mashrura Tasnim, Malikeh Ehghaghi, Brian Diep, and Jekaterina Novikova. 2022. DEPAC: a corpus for depression and anxiety detection from speech. In *Proceedings of CLPsych*.

Mathieu Templier and Guy Paré. 2015. A framework for guiding and evaluating literature reviews. *Communications of the Association for Information Systems*, 37(1):6.

Michael Tenti, William Raffaeli, and Paola Gremigni. 2022. A narrative review of the assessment of depression in chronic pain. *Pain Management Nursing*, 23(2):158–167.

Yevhen Tyshchenko. 2018. Depression and anxiety detection from blog posts data. *Nature Precis. Sci., Inst. Comput. Sci., Univ. Tartu, Tartu, Estonia*, pages 6–46.

Rudolf Uher, Roy H Perlis, Anna Placentino, Mojca Zvezdana Dernovšek, Neven Henigsberg, Ole Mors, Wolfgang Maier, Peter McGuffin, and Anne Farmer. 2012. Self-report and clinician-rated measures of depression severity: can one replace the other? *Depression and anxiety*, 29(12):1043–1049.

Ozlem Uzuner, Amber Stubbs, and Michele Filannino. 2017. A natural language processing challenge for clinical records: Research domains criteria (rdoc) for psychiatry. *Journal of biomedical informatics*, 75:S1.

Aleksander Wawer, Izabela Chojnicka, Lukasz Okruszek, and Justyna Sarzynska-Wawer. 2022. Single and cross-disorder detection for autism and schizophrenia. *Cognitive Computation*, 14(1):461–473.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

World Health Organization. 1992. *The ICD-10 classification of mental and behavioural disorders: clinical descriptions and diagnostic guidelines*, volume 1. World Health Organization.

Chi-Shin Wu, Chian-Jue Kuo, Chu-Hsien Su, Shi-Heng Wang, and Hong-Jie Dai. 2020. Using text mining to extract depressive symptoms and to validate the diagnosis of major depressive disorder from electronic health records. *Journal of affective disorders*, 260:617–623.

Shihao Xu, Yiming Yan, Yanli Ding, Feng Li, Shu Zhang, Haoyun Tang, Chao Luo, Yan Li, Hao Liu,

Yu Mei, et al. 2025. Identifying psychiatric manifestations in outpatients with depression and anxiety: A large language model-based approach. *medRxiv*, pages 2025–01.

Daniel Zarate, Vasileios Stavropoulos, Michelle Ball, Gabriel de Sena Collier, and Nicholas C Jacobson. 2022. Exploring the digital footprint of depression: a prisma systematic literature review of the empirical evidence. *BMC psychiatry*, 22(1):421.

Alexandros Dimitrios Zervopoulos, Evangelos Geramanis, Alexandros Toulakis, Asterios Papamichail, Dimitrios Triantafylloy, Theofanis Tasoulas, and Katia Kermanidis. 2019. Language processing for predicting suicidal tendencies: a case study in greek poetry. In *IFIP AIAI*.

Yunpeng Zhao, Xing He, Zheng Feng, Sarah Bost, Mattia Prosperi, Yonghui Wu, Yi Guo, and Jiang Bian. 2022. Biases in using social media data for public health surveillance: A scoping review. *International Journal of Medical Informatics*, 164:104804.

Li Zhou, Amy W Baughman, Victor J Lei, Kenneth H Lai, Amol S Navathe, Frank Chang, Margarita Sordo, Maxim Topaz, Feiran Zhong, Madhavan Murrali, et al. 2015. Identifying patients with depression using free-text clinical documents. In *MEDINFO*.

Mark Zimmerman and William Coryell. 1987. The inventory to diagnose depression, lifetime version. *Acta Psychiatrica Scandinavica*, 75(5):495–499.

Bochao Zou, Jiali Han, Yingxue Wang, Rui Liu, Shenghui Zhao, Lei Feng, Xiangwen Lyu, and Huimin Ma. 2023. Semi-structural interview-based chinese multimodal depression corpus towards automatic preliminary screening of depressive disorders. *IEEE Transactions on Affective Computing*, 14(4):2823–2838.