A Survey on Mental Health Datasets and Resources

Ziwei Gong[♠] **Chengtao Dai**[♠] **Bolei Ma**[♡] **Melody Ma**[♠] **Manasi Sharma**[♠]
Thai Quoc On[♠] **Medha Morparia**[♠] **Benjamin Enos**[◇] **Yuwen Yu**[†]
Philip Resnik[♣] **Sarah Ita Levitan**[†] **Julia Hirschberg**[♠]

[♠]Columbia University [♡]LMU & MCML [◇]BU [†]CUNY [♣]UMD

{sara.ziweigong, julia}@cs.columbia.edu

Abstract

Computational approaches to mental health have become an increasingly important area of AI research, supported by a growing number of datasets. This survey presents a dataset-centric review of mental health resources from 2001 to 2025, focusing on how mental-health states are defined, represented, and evaluated in NLP. We analyze datasets across modalities, organizing them by condition, data source, labeling strategy, task formulation, evaluation practice, and finally identify recurring challenges and opportunities in existing resources, aiming to inform the development of more clinically meaningful and responsible datasets. We aim to provide a holistically review to clarify the current states and guide the development of future mental health resources and applications.

1 Introduction

In recent years, the field has witnessed a significant surge of interest in applying artificial intelligence to mental health, accompanied by rapid advances in both dataset creation and model development in this high-stakes domain (Shatte et al., 2019). AI-based systems have been explored for a wide range of applications, including mental health screening, symptom monitor, diagnostic support, and even early-stage therapeutic interaction (Calvo et al., 2017; Malgaroli et al., 2023). Given the sensitive and safety-critical nature of the applications, the availability of high-quality datasets and reliable evaluation protocols is essential for developing models that are trustworthy, robust, and well-aligned with clinical and ethical standards (Topol, 2019; Wiens et al., 2019).

Recent efforts have provided valuable surveys of AI and NLP for mental health, including broad reviews of machine learning and natural language processing for mental-illness detection (Zhang et al., 2022; Le Glaz et al., 2021), social-media based mental state assessment (Wongkoblap et al., 2017; Harrigian et al., 2021; Bucur et al., 2025a,b), and disorder- or population-specific work such as *bipolar disorder* and youth-focused prevention (Harvey et al., 2022; Ibrahim et al., 2025). Other reviews take a more explicitly dataset-oriented perspective in particular domains, for example clinical mental-health AI datasets (Mandal et al., 2025), directories of public datasets for youth mental health (Min et al., 2025), and physiological datasets for stress and anxiety research (Cobá Juárez Pegueros and Rodríguez-Arce, 2025). Despite these advances, most existing surveys are organized around modeling, a specific task or data source, and typically treat datasets at a high level. To our knowledge, none provide a unified, dataset-centric view that systematically organizes mental-health resources across modalities and settings, and across a broad set of disorders.

To advance the field and clarify the research space, a comprehensive, dataset-centered review of mental health resources is much needed: what resources exist across text, speech, and multimodal signals; how they have been constructed and used; and where important gaps remain.

Hence, we present a comprehensive and structured review of mental health datasets and resources since 2001[1], organizing prior work by disorder and key dataset design dimensions, and identifying recurring opportunities and challenges. To support transparency and reuse, we additionally release an accompanying online resource that consolidates all 229 reviewed datasets into structured tables and interactive visualizations. Available at `https://ziweig.github.io/mental-health-datasets-resources-review/` and Appendix G for details.

The remainder of the paper is organized as follows: We present our structured review framework

[1] We survey 200+ papers from major *CL venues, as well as COLING, NeurIPS, ICML, ICLR, and influential arXiv preprints. Appendix A details the paper selection process.

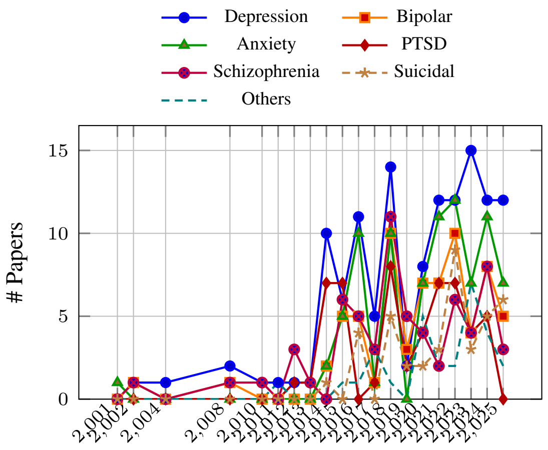


Figure 1: Annual paper counts by disorder since 2001.

and take a closer look at datasets created for different types of mental health disorders (Sec 2), task formulations (Sec 3), data construction methods (Sec 4), and evaluation and application settings (Sec 5). Finally, in our discussion (Sec 6), we summarize challenges and opportunities, with a focus on clinically-grounded dataset curation, representation of populations and disorders, and open questions about what is represented and what remains missing in current resources.

2 Mental Health Conditions Represented in Existing Datasets

We organize our review first by mental health conditions, focusing on *depression, bipolar, anxiety, PTSD, suicidal ideation, schizophrenia*, and a set of less-studied conditions grouped as *other* disorders. As shown in Figure 1, we see steady growth in dataset-related publications since 2001, with particularly rapid expansion since 2015, alongside persistent imbalance across disorders.

This organization follows common practice in the NLP literature and reflects how datasets are named and retrieved, rather than a clinical taxonomy: many conditions encompass diverse subtypes and symptom profiles that are often collapsed in dataset design; Appendix B provides brief clinical background to contextualize this abstraction.

2.1 Depression Disorders

Depression disorders are characterized by persistent low mood and loss of interest or pleasure, typically accompanied by changes in sleep or appetite, fatigue, and difficulty concentrating that interfere with daily functioning (American Psychiatric Association, 2022). Depression is one of the most extensively studied mental health conditions in NLP. Large-scale shared tasks, most notably the CLEF eRisk[2] challenges and the CLPsych workshop[3] series, have played a central role in standardizing task formulations and benchmarking user- and timeline-level depression detection on social media (Yates et al., 2017).

Early work primarily leveraged Twitter and Reddit data to identify linguistic correlates of depression, consistently observing increased self-focus and negative affect across English and Spanish forums (Ramirez-Esparza et al., 2021). As these corpora scaled, research expanded to examine robustness and equity, including gender and racial fairness (Aguirre et al., 2021) and cross-cultural generalization gaps (Abdelkadir et al., 2024a). While English dominates the landscape, depression datasets have also been developed for German (Valstar et al., 2013, 2014), Mandarin (Shen et al., 2022; Yao et al., 2022), and Italian (Tao et al., 2023).

In parallel, speech and multimodal datasets have become canonical benchmarks for severity modeling, often collected via structured or semi-structured interviews. Resources such as AVEC and DAIC-WOZ pair audio–video recordings and transcripts with validated clinical instruments (e.g., PHQ-8/9, BDI, MADRS) (Mundt et al., 2007; Gratch et al., 2014). Recent datasets increasingly incorporate clinician involvement (Dumpala et al., 2024), longitudinal sampling (Lewis et al., 2025), self-supervised speech representations (Maji et al., 2024; Dumpala et al., 2024, 2025), and advanced multimodal fusion methods (Campbell et al., 2023; Fara et al., 2023a), extending earlier cross-sectional and prompted-speech paradigms.

2.2 Bipolar Disorder

Bipolar disorder involves recurrent shifts between manic or hypomanic and depressive episodes that disrupt mood, energy, and functioning (Staff, 2024), which is often under-diagnosed and is managed as a lifelong condition (of Mental Health, 2025).

Early bipolar disorder datasets were primarily constructed from large-scale social media corpora. Coppersmith et al. (2014) introduced a scalable self-disclosure-based collection paradigm on Twitter, identifying users who explicitly reported a bipolar diagnosis and matching them with control users on linguistic and temporal factors. This framework laid the foundation for subsequent bipolar

[2] https://erisk.irlab.org/
[3] https://clpsych.org/

datasets and analyses (Coppersmith et al., 2015b,a; Amir et al., 2017; De Choudhury et al., 2021), as well as multi-condition shared-task and benchmark datasets such as CLPsych and SMHD (Milne et al., 2016; Cohan et al., 2018a). Across these resources, studies consistently report linguistic, affective, and behavioral signals associated with bipolar disorder.

Despite this progress, bipolar research remains text-dominated: 58% of datasets rely on social media text, only 9% include any non-text modality, and none are bipolar-specific multimodal resources (Figures 5 and 6). A small number of studies incorporate audio, video, or neuroimaging data (such as DAIC-WOZ (Gratch et al., 2014)), clinical speech interviews (Wang et al., 2020), audiovisual perception studies (Erdener et al., 2019), facial and motor analyses (Bersani et al., 2013; Kang et al., 2018), and multimodal neuroimaging fusion (Wang et al., 2024), nearly all appearing after 2020.

Overall, bipolar disorder datasets emphasize text-based social media data, despite clinical evidence that mood episodes manifest through vocal, facial, and motor channels. The absence of bipolar-specific multimodal resources remains a key limitation for developing clinically grounded and diagnostically robust models.

2.3 Anxiety Disorders

Anxiety disorders involve persistent and excessive fear or worry that interferes with daily functioning and affect roughly one-third of U.S. individuals over their lifetime (National Institute of Mental Health (NIMH), 2025).

Most anxiety datasets in NLP are text-based and derived from social media, with a clear shift from Twitter to Reddit over time due to platform access constraints and differences in post length during a period of time (App.Fig 7 and 8). Earlier work primarily leveraged Twitter, whereas recent datasets increasingly rely on Reddit, exemplified by DREADDit, a large-scale corpus with over 190k posts and comments annotated for stress and anxiety signals (Turcan and McKeown, 2019a). Additional anxiety-related corpora include Twitter-STMHD, which uses self-disclosure to label anxiety and other mental health conditions at scale (Suhavi et al., 2022), as well as smaller, targeted datasets such as the SIR-Lab student corpus, which collects structured self-reports outside of public social media contexts (Sahu et al., 2025).

Beyond text, only a limited number of multimodal resources exist. Therapy video corpora such as AlexStreet support analysis of therapist–client interactions (Lin et al., 2024a), while physiological datasets like WESAD combine wearable biosignals and self-reports across affective states (Schmidt et al., 2018). Overall, anxiety research remains dominated by text-based social media data, with comparatively sparse multimodal coverage.

2.4 Post-Traumatic Stress Disorder (PTSD)

PTSD develops following exposure to traumatic events and is characterized by intrusive memories, avoidance, negative alterations in cognition and mood, and heightened arousal (American Psychiatric Association, 2022). It frequently co-occurs with depression, anxiety, and substance use disorders (for PTSD, 2025).

PTSD datasets in NLP are relatively recent, with all reviewed resources published after 2012 (App.Fig 9), and are dominated by text-based social media data. The CLPsych 2015 shared task established a widely adopted benchmark using Twitter timelines from users self-reporting depression or PTSD, enabling user-level classification and catalyzing subsequent work (Coppersmith et al., 2015b). This paradigm was extended by later Twitter-based PTSD datasets using diagnosis statements and lexicon-driven filtering (Coppersmith et al., 2018), as well as Reddit-based corpora that support longitudinal symptom trajectories (Son et al., 2021) and stress- or trauma-related signals for PTSD and comorbid conditions (Murarka et al., 2021; Turcan and McKeown, 2019b; Raihan et al., 2024a; Kulkarni et al., 2021).

A small number of text-only datasets focus on high-risk or trauma-exposed populations, including studies of psychological distress in online forums (Saleem et al., 2012) and care-seeking behavior among women veterans (Kelly et al., 2020), underscoring ethical considerations when working with vulnerable groups.

Multimodal PTSD datasets remain rare. Existing resources include unconstrained audiovisual data (PTSD in the Wild (Sawadogo et al., 2024)) and clinically oriented interview corpora with synchronized audio–video signals (SimSensei, DAIC (Gratch et al., 2014; DeVault et al., 2014)), but these are limited in scale and population diversity. Overall, among the 37 PTSD datasets reviewed, only two are multimodal, with the remainder relying primarily on social media, restricting generalization and clinical relevance.

2.5 Suicidal Thoughts

Suicide is a leading cause of death among adolescents and young adults in the U.S., making suicidal ideation – a persistent pattern of suicidal thoughts or plans – a critical target for early detection and prevention (Centers for Disease Control and Prevention, 2021b,a).

Most datasets for suicidal ideation and self-harm detection are text-based, derived from social media, spanning diverse labeling strategies and granularities. Public resources include post-level annotations for empathic responses in peer-support forums (Behavioral Data Lab, 2020), user-level severity labels for Reddit users based on suicide risk rubrics (Shing et al., 2018; Chim et al., 2024; Gaur and Others, 2020), paraphrased post-level ideation labels for privacy preservation (Lab, 2022), and multi-platform self-harm corpora (Lab, 2021a). Population-specific resources such as the LGBTeenDataset further illustrate the heterogeneity of label scope and annotation practices (Lab, 2021b).

Multimodal resources remain limited but are emerging. Therapy-focused video corpora such as AlexStreet(Lin et al., 2024a) and Sahar Corpus (Izmaylov et al., 2023) integrate text with additional modalities to study multimodal suicide risk signals. Despite their promise, multimodal datasets remain sparse relative to large-scale text corpora, constraining clinically grounded modeling.

2.6 Schizophrenia

Schizophrenia is a chronic psychiatric disorder marked by disturbances in thought, perception, emotion, and behavior, commonly divided into positive and negative symptom clusters (Mayo Foundation for Medical Education and Research, 2024). Research activity has increased steadily, drawing on datasets spanning speech, clinician interviews, and online text (Fig 1 & 11).

Early computational resources focused on identifying linguistic irregularities distinguishing schizophrenic from healthy or other disordered speech (Kuperberg, 2010). Studies using conversational data, writing samples, and social media consistently reported reduced lexical diversity, increased self-focus and pronoun use, greater concreteness, and elevated affective markers such as anger (Aich et al., 2022; Coppersmith et al., 2015a; Sarioglu Kayi et al., 2017). Speech-based datasets and analyses further revealed syntactic and semantic disruptions, including ambiguous pronoun usage, incoherence, repetition, and reduced semantic coherence (Iter et al., 2018; Bar et al., 2019; Hong et al., 2012; Li et al., 2021; Tang et al., 2021; Espy-Wilson, 1992). Social-media–based datasets extend these findings at scale, showing increased use of function words, pronouns, and cognitive-, health-, and death-related categories among self-reported schizophrenic users (Mitchell et al., 2015; Zomick et al., 2019). Recent work emphasizes interpretability and clinical relevance, including masking-based analyses of diagnostically salient lexical features (Shriki et al., 2022) and automated assessment of speech and social functioning using clinical assessment (Aich et al., 2025).

Overall, schizophrenia datasets primarily capture linguistic manifestations of positive symptoms; while spontaneous and spoken data are increasingly used to improve ecological validity, reliance on self-reported social media data remains common, limiting diagnostic certainty and generalizability.

2.7 Other Disorders

Beyond the major disorders discussed above, computational mental health research increasingly covers additional conditions. This expansion is enabled primarily by multi-condition social media corpora and clinically grounded datasets.

Attention-Deficit/Hyperactivity Disorder (ADHD) involves persistent inattention and/or hyperactivity–impulsivity that impairs functioning(American Psychiatric Association, 2022). Neuroimaging is central in public benchmarks such as ADHD-200, which aggregates resting-state fMRI and structural MRI across sites(Bellec et al., 2017). More recent work expands to multimodal behavioral and physiological signals (e.g., EEG, eye tracking, actigraphy, VR-based experiments)(Wiebe et al., 2023, 2024). Transdiagnostic datasets such as the Healthy Brain Network further enable cross-condition modeling(Alexander et al., 2017), and ADHD is also represented in multi-disorder social media corpora such as SMHD (Cohan et al., 2018b).

Obsessive-Compulsive Disorder (OCD) is characterized by intrusive, unwelcome thoughts and repetitive behaviors or cognitive actions executed to alleviate anxiety (American Psychiatric Association, 2022). Large-scale neuroimaging resources from the ENIGMA-OCD Working Group provide structural/diffusion benchmarks, including medication-related effects and white-matter findings across international cohorts(Bruin et al., 2020;

Piras et al., 2021). In contrast, linguistic analyses of OCD largely rely on multi-diagnosis social media datasets rather than disorder-specific corpora (Cohan et al., 2018b).

Eating disorders (ED), such as anorexia nervosa, bulimia nervosa, and binge-eating disorder, are characterized by disordered eating patterns and body image issues (World Health Organization, 2022). ED research leverages online communities to study both risk and recovery. Analyses of pro-anorexia and recovery-oriented discourse track linguistic trajectories and community dynamics (Chancellor et al., 2016), while orthographic variation studies reveal adaptation to platform moderation practices (Stewart et al., 2017). Broader subreddit-level analyses contextualize eating-disorder language within health and crisis discussions (Low et al., 2020).

Borderline Personality Disorder (BPD) is marked by pervasive instability in mood, self-perception, and interpersonal relationships, frequently accompanied by impulsive behavior and a fear of abandonment (American Psychiatric Association, 2022). BPD remains underrepresented in standalone datasets, but appears in comparative analyses across disorder-specific Reddit communities, enabling joint modeling with anxiety, bipolar disorder, and schizophrenia (Kim et al., 2023).

Seasonal Affective Disorder (SAD) is a recurrent depressive disorder characterized by a winter-predominant seasonal pattern (American Psychiatric Association, 2022). Similarly, SAD is primarily studied through early Twitter-based self-disclosure corpora, highlighting seasonal linguistic and temporal patterns relevant for risk monitoring (Coppersmith et al., 2015a).

Insomnia involves chronic difficulty maintaining sleep with daytime impairment (American Academy of Sleep Medicine, 2023). Existing datasets emphasize digital phenotyping from social media, including inference of sleep timing from Reddit activity and surveys (Meyerson et al., 2023), longitudinal analysis of treatment-related discourse (Cummins et al., 2025), and deep learning approaches for detecting irregular sleep patterns (Islam et al., 2026).

Dementia includes neurocognitive disorders with progressive decline in memory, reasoning, and communication. Alzheimer's disease is the most common subtype (World Health Organization, 2022). In contrast to social-media-centered conditions, dementia research is anchored in clinically curated speech corpora. Resources such as DementiaBank (Lanzi et al., 2023) and the ADReSS/ADReSSo shared tasks establish standardized benchmarks for spontaneous speech analysis (Luz et al., 2021b,a), with recent surveys summarizing open challenges in dementia-focused NLP (Peled-Cohen and Reichart, 2025).

Overall, datasets for these conditions reflect a broader trajectory from small, clinic-based collections toward larger, socially embedded and trans-diagnostic resources (Alexander et al., 2017; Bellec et al., 2017; Cohan et al., 2018b; Coppersmith et al., 2015a). Since 2019, multimodal fusion and transformer-based modeling have increasingly integrated physiological, behavioral, and linguistic signals, pointing toward more ecologically valid and data-diverse computational mental health systems (Devlin et al., 2019a; Wiebe et al., 2023, 2024).

3 Task Settings

Mental-health datasets are not only defined by which disorder they target, but also by how they are operationalized for modeling. We start with summarize common task settings from three perspectives: task formulation, data modality and format, language and cultural context. These perspectives shape downstream decisions about labeling, data access, and dataset construction, and they also determine which evaluation protocols are appropriate.

3.1 Task Types

We begin by outlining the dominant task formulations used across disorders. Across disorders, three task families dominate: **(i) binary detection** at the post- or user-level on social-media timelines and forum histories; **(ii) severity/symptom prediction** using clinician-rated or self-report instruments; and **(iii) early-risk or temporal modeling** where observations arrive sequentially and time-aware metrics are used. For depression, these are instantiated in the eRisk and CLPsych shared tasks for user/post-level detection (Yates et al., 2017; Coppersmith et al., 2015b; Losada and Crestani, 2016), and in speech/multimodal settings via DAIC-WOZ and AVEC for severity regression/classification against PHQ/BDI/MADRS (Gratch et al., 2014; Valstar et al., 2013, 2014). PTSD work commonly adopts CLPsych'15 Twitter-based detection with user/post labels (Coppersmith et al., 2015b). Schizophrenia studies frequently frame *case-vs.-control* classification over conversations/interviews or social media, with feature- or transformer-based models identifying linguistic markers (Iter et al., 2018; Bar et al., 2019; Mitchell et al., 2015; Zomick et al., 2019; Tang et al., 2021; Li et al., 2021).

While these task families recur across disorders, depression provides one of the clearest examples of how task formulation aligns with evaluation protocols and benchmark evolution; we provide a detailed case study in Appendix D. These task formulations directly constrain what signals a dataset must provide (e.g., text-only vs. multimodal, snapshot vs longitudinal) and whether the intended use case requires real-time monitoring or retrospective classification. We therefore next summarize the modalities and structural forms in which mental-health data are collected and released.

3.2 Data Modality and Format

Format refers to both the *modality* and the *structural form* of the data collected, including whether the data are textual, audio, visual, multimodal, or longitudinal. The dominant modality in computational mental health remains **text**. A large portion of datasets rely on Reddit/Twitter timelines and forum posts, often with self-disclosure labels or weak supervision (Yates et al., 2017; Turcan and McKeown, 2019a). For example, the bipolar subset shows strong social-media predominance (Fig 5: 58% of datasets), and similarly for shared tasks resources (Coppersmith et al., 2015b). Text dominates because mental health signals manifest in written communication, and text is easier to collect, share, and annotate; even speech datasets are often released as transcripts (Gratch et al., 2014).

Beyond text, a second major category consists of **speech and interview-based data**. Depression severity datasets such as DAIC-WOZ and AVEC pair audio, video, and transcripts with standardized instruments including PHQ, BDI, and MADRS (Gratch et al., 2014; Valstar et al., 2013, 2014). Schizophrenia studies analyze patient speech in interviews and narratives, highlighting coherence and derailment, pronoun ambiguity, and filler patterns (Iter et al., 2018; Bar et al., 2019; Hong et al., 2012; Tang et al., 2021; Li et al., 2021). It's noteworthy dementia resources are heavily speech-centric (Luz et al., 2021b,a; Lanzi et al., 2023).

There has been a clear trend toward **multimodal and physiological data** in recent years, motivated by the recognition that nonverbal behavior carries clinically important signals. Depression research increasingly adopts multimodal fusion and self-supervised speech representations (Maji et al., 2024; Dumpala et al., 2024, 2025; Campbell et al., 2023; Fara et al., 2023a). Anxiety and affective computing literature includes wearable-sensor datasets and therapy video corpora (Schmidt et al., 2018; Lin et al., 2024a). However, multimodal resources remain unevenly distributed across disorders: for example, bipolar-specific multimodal datasets are rare, as shown in Fig 6.

In addition to modality, datasets differ in their **temporal structure**. Some corpora consist of independent samples, whereas others are organized as longitudinal timelines. In early-risk detection challenges, data are explicitly revealed over time to simulate continuous monitoring (Losada and Crestani, 2016). Similarly, Reddit-based datasets often include complete user posting histories spanning months or years (Yates et al., 2017; Tsakalidis et al., 2022; Tseriotou et al., 2025), enabling temporal analysis of symptom progression. In contrast, most clinical interview datasets represent one-time snapshots. When available, longitudinal structure allows researchers to study symptom trajectories and assess the feasibility of predicting disorder onset from early signals.

In summary, while text remains the primary data format, the field is steadily expanding toward speech, audiovisual, wearable, and multi-stream data, alongside increasing use of longitudinal data to better reflect real-world mental health dynamics (Bufano et al., 2023; Khoo et al., 2024; Garcia-Ceja et al., 2018). Beyond tasks and signals, however, dataset coverage also determines what claims can be made about generalization and fairness. We therefore review the linguistic and cultural representation next.

3.3 Languages and Culture Representation

Resources are English-centric across task families, particularly for social-media datasets and speech corpora (eRisk/CLPsych/DAIC/AVEC) (Yates et al., 2017; Gratch et al., 2014; Valstar et al., 2013, 2014). Nonetheless, non-English datasets exist: German, Mandarin, and Italian (Valstar et al., 2013, 2014; Shen et al., 2022; Yao et al., 2022; Tao et al., 2023). Cross-cultural generalization and regional fairness have been explicitly raised in large-scale social-media analyses (Abdelkadir et al., 2024a; Aguirre et al., 2021).

4 Dataset Construction

Mental health datasets differ not only in the disorders they target, but also in how psychological states are operationalized through data collection and labeling. Dataset construction reflects recurring trade-offs between scale and diagnostic validity, accessibility and privacy, and ecological realism and experimental control.

Sources of Data. Mental health datasets vary in where and how raw data are collected, shaping both the scale and reliability of downstream analyses. A large proportion draw on user-generated content from online platforms, particularly social media forums, due to their accessibility and the candid discussions from personal disclosures. Reddit and Twitter are the dominant sources, typically consists of posts from users with self-disclosed diagnoses matched with controls (Yates et al., 2017). Continuing efforts similarly curated Twitter data for depression and PTSD research (Coppersmith et al., 2014, 2015b; Turcan and McKeown, 2019a). These sources are valued for their large scale and real-life language, though they come with noise, since users may not be formally diagnosed.

Meanwhile, a smaller but influential set of datasets originates from clinical or controlled settings. Corpora such as DAIC-WOZ and the AVEC challenge datasets consist of structured interviews with aligned audio, video, and transcripts paired with standardized mood assessments (Gratch et al., 2014; Valstar et al., 2014). Although involving smaller samples, these datasets offer stronger diagnostic grounding and more reliable labels.

Data sources have also expanded beyond English and Western platforms. Recent studies have constructed datasets from non-English online communities, including Chinese (Shen et al., 2022; Yao et al., 2022) and Italian health forums (Tao et al., 2023), broadening linguistic and cultural coverage. A small number of projects further incorporate population-level data such as surveys or electronic health records, though these remain uncommon in NLP due to privacy constraints. Overall, mental health datasets span a continuum from in vivo online expressions to in vitro clinical recordings, reflecting a fundamental trade-off between scale and diagnostic clarity.

Sources of Label. Labeling choices in mental health datasets determine how psychological states are operationalized and measured in NLP models. Across the literature, labeling methodologies broadly fall into three categories: self-reported labels, clinically assessed labels, and hybrid or proxy-based labels. Many large-scale social media datasets rely on self-disclosure, using explicit statements of diagnosis as supervision signals (Yates et al., 2017). This approach enables scalable annotation but introduces noise due to mis-reporting, ambiguity, or lack of formal diagnosis. Clinically grounded datasets instead derive labels from professional diagnoses or validated assessment instruments, including structured clinical evaluations and standardized questionnaires (Gratch et al., 2014; Valstar et al., 2014), which offer higher diagnostic validity but are typically limited in scale. Hybrid approaches use proxy signals to approximate mental health status without direct diagnosis, such as user behavior patterns, community membership, triage categories (Turcan and McKeown, 2019a). Some further integrate clinically informed risk frameworks into annotation or evaluation pipelines, particularly for suicide-related tasks (Gaur et al., 2021; Ji et al., 2021).

Overall, labeling methodologies reflect a trade-off between supervision scale and diagnostic rigor. While self-reported and proxy labels support large-scale modeling, clinically assessed labels provide stronger validity, motivating hybrid strategies that combine multiple sources of supervision to balance coverage and reliability.

Dataset Sharing. Dataset sharing practices vary widely due to privacy, ethical, and legal constraints. A subset of datasets are released publicly in anonymized form (Ren et al., 2025), including large-scale social media corpora (Yates et al., 2017) and multi-disorder dataset (Cohan et al., 2018b), which are openly hosted on online platforms like GitHub to support reproducible benchmarking. Physiological and affective datasets such as WESAD are also publicly accessible under research-only licenses (Schmidt et al., 2018). Anonymization methods detailed in Appendix F

In contrast, many datasets are distributed under restricted-access models. Shared-task resources such as eRisk are available only to registered participants under usage agreements (Losada and Crestani, 2016), while CLPsych-era datasets for depression and PTSD were primarily released within workshop contexts rather than as fully open corpora (MacAvaney et al., 2021). Clinical datasets are typically subject to the strictest controls: resources involving patient interviews or medical speech data require formal application processes and data-use agreements (Gratch et al., 2014; Lanzi et al., 2023; Luz et al., 2021a). Overall, while dataset availability has expanded, a substantial portion of clinically grounded resources remains limited-access.

5 Evaluation

The evaluation of computational models for mental health disorder detection and assessment, particularly those leveraging speech and social media text (Benton et al., 2017), requires a comprehensive set of quantitative metrics and methodological frameworks. These evaluation strategies are designed not only to measure predictive accuracy but also to ensure robustness, clinical validity, and generalizability across diverse populations and modalities.

Computational Metrics. Quantitative evaluation remains the foundation of performance benchmarking, tailored to the specific nature of the prediction task. From a computational perspective, the choice of metrics is strictly dictated by the nature of the predictive task. For classification-based symptom detection, standard metrics such as precision, recall, accuracy, and F1-score are ubiquitous (Young et al., 2025; Thamrin and Chen, 2024; Srivastava et al., 2025; Zhou et al., 2025; Dumpala et al., 2024; Campbell et al., 2023; Maji et al., 2023; Tao et al., 2023; Sampath et al., 2023; Shen et al., 2022). However, given the high stakes of missing at-risk individuals, metrics emphasizing recall, such as the F2-score, are crucial for minimizing false negatives (Sawhney et al., 2021), while the Area Under the Receiver Operating Characteristic Curve (ROC-AUC) is employed to assess robustness (Fara et al., 2023b; Campbell et al., 2023). To address the prevalent class imbalance in mental health datasets, studies increasingly rely on the Matthews Correlation Coefficient (MCC) (Agarwal et al., 2022) and macro-averaged scores (Dumpala et al., 2025; Raihan et al., 2024b). Conversely, for continuous severity estimation, regression metrics like Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) are the standard for measuring deviation from ground-truth ratings (Dumpala et al., 2024; White et al., 2025). With the rise of LLMs for therapeutic dialogue, evaluation has expanded to include generation-specific metrics such as BLEU, ROUGE, and Perplexity (PPL) (Chen et al., 2024b; Park et al., 2022) to assess linguistic fluency and diversity.

Human-centric Evaluation. In addition to computational evaluation, several studies integrate human assessments to ensure clinical interpretability and user relevance. These include evaluations conducted by clinical trainees (Callejas et al., 2014), licensed psychiatrists (Tao et al., 2023; Yin et al., 2025), professional clinicians (Lewis et al., 2025; Shen et al., 2022), trained annotators (Abdelkadir et al., 2024a; Lissak et al., 2024), and end users to assess satisfaction or perceived empathy. Recent work has also explored the use of large language models such as GPT-3.5, GPT-4 (Lissak et al., 2024), GPT-4 Turbo, and Gemini Pro 1.0 (Xu et al., 2025) as evaluation agents or "judges" for dialogue quality and therapeutic relevance.

Agent-based and Scalable Evaluation. End-user evaluations are critical for assessing perceived empathy and satisfaction in conversational agents. Recently, a novel mode of evaluation has emerged in the form of "LLM-as-a-Judge," where advanced models like GPT-4 and Gemini Pro are deployed to simulate human evaluation, scoring dialogue quality and therapeutic relevance (Lissak et al., 2024; Xu et al., 2025). Taken with appropriate caveats (Lawrence et al., 2024; Jung et al., 2025; Chen et al., 2024a), this represents a shift towards scalable, albeit proxy, qualitative assessment.

Despite these advancements, a critical gap remains between machine learning metrics and real-world clinical efficacy, as static benchmarks fail to capture the dynamic nature of mental health. Bridging this gap requires a framework grounded in **human-AI collaboration**. Ultimately, trustworthy deployment relies on standardized benchmarks that integrate scalable computational scoring with continuous clinician feedback.

6 Toward Clinically Aware and Responsible Dataset Design

Across the datasets surveyed, advances in scale and modeling sophistication have outpaced progress in how mental-health states are defined, represented, and evaluated. Rather than isolated limitations, the challenges identified in this review reflect recurring dataset design choices that systematically shape what models can learn, generalize, and credibly claim. We organize the discussion around three cross-cutting questions: how current labels align, who and what existing datasets represent, and how dataset design choices constrain evaluation and sustainable deployment.

What do current labels represent, and are they aligned with clinical use? A central limitation to clinical validity in mental-health NLP lies not in model capacity, but in how psychological states are operationalized through dataset labels. As summarized in Sec 4, datasets across depression, anxiety, bipolar disorder, PTSD, suicidality, and related conditions predominantly rely on self-disclosure, community membership, or weak supervision, with only a minority incorporating manual verification or clinician-rated instruments (Wu et al., 2024; Yao et al., 2022). Importantly, label noise in these settings is not random: self-disclosure rates vary with stigma, access to care, demographic factors, and platform norms, complicating cross-dataset comparison and transfer. Moreover, self-reported statements are not equivalent to current clinical status and may reflect historical, provisional, or context-dependent conditions (Chancellor et al., 2023).

A closely related misalignment concerns target formulation. As shown in Sec 3, most datasets cast mental-health modeling as a binary detection problem, despite clinical practice emphasizing symptom severity, trajectory, and risk. This gap is particularly salient for suicidality, where intervention decisions depend on graded risk levels rather than the presence of ideation alone; datasets that preserve ordinal or severity-based risk annotations are therefore more compatible with application-facing modeling than binary labels (Gaur et al., 2021; Ji et al., 2021; Liu et al., 2025). Even when validated instruments are available, scores are frequently thresholded for classification.

Finally, most datasets abstract away comorbidity, despite it being common rather than exceptional in clinical settings (e.g., depression with anxiety, bipolar disorder with suicidality). Single-disorder framing and one-vs.-rest evaluation risk producing models that capture generic distress signals rather than disorder-specific patterns, or that fail under realistic multi-label conditions. While modeling techniques can partially address this, the absence of explicit co-morbidity labels limits supervision and evaluation. Together, these observations point to a clear opportunity: future datasets should preserve severity- and risk-aware labels, explicitly document label provenance, treat comorbidity as a first-order property rather than a confound, *all of which can benefit from actively involve domain experts in the dataset design process to ensure alignment with clinical realities.*

Who and what is represented in current datasets? What models learn about mental health is fundamentally constrained by who appears in the data and through which signals their experiences are captured. As shown in Fig 1 and Sec 2, depression and suicidality dominate the dataset landscape, supported by large social-media corpora and multiple shared tasks, whereas other conditions have far fewer public, well-documented resources and often appear only as secondary labels. This imbalance shapes benchmarks and research incentives, skewing methodological progress toward conditions with abundant data rather than those where modeling would be clinically impactful but data are harder to obtain.

Platform and population bias further narrow coverage. The dominance of Reddit and Twitter/X privileges users who are younger, English-speaking, and willing to self-disclose online, while underrepresenting older adults, non-Western populations, and individuals whose mental-health experiences are primarily offline or clinical. Linguistic patterns on social media also differ substantially from those in daily dialogues or clinical intake notes, limiting transfer (Chancellor et al., 2019). Although non-English datasets exist, cross-lingual and cross-cultural generalization is rarely evaluated (Abdelkadir et al., 2024b).

Modality gaps compound these issues. Despite clinical evidence that vocal, facial, motor, and physiological signals are central to the manifestation of several disorders, most datasets remain text-dominated, with multimodal resources concentrated on a small subset of conditions. This mismatch constrains models' ability to capture state changes (e.g., manic versus depressive phases) and nonverbal risk cues. Without broader disorder coverage, population diversity, and modality alignment, claims of "general" mental-health capability remain difficult to substantiate.

What constrains evaluation and sustainable deployment in mental health NLP? Evaluation practices and data governance reflect dataset design decisions and, in turn, constrain real-world applicability. In particular, mental health NLP now faces a *"data crisis"*: whereas general NLP progress has been accelerated by openly shared datasets and community-wide benchmarks, the sensitive nature of mental health data has stifled similar large-scale collaboration. Shared tasks and benchmarking efforts in this domain have indeed proven valuable in providing data for research and fostering novel modeling approaches, but such resources remain the exception rather than the norm. As a result, standard metrics continue to dominate current evaluations, even for tasks where calibration, uncertainty, and asymmetric error costs are critical (e.g., risk stratification). Aligning metrics with label type and intended use is therefore essential for clinically-awared and responsible evaluation.

A parallel and increasingly urgent challenge contributing to the *data crisis* is sustainable access to data. Open social-media datasets have enabled shared benchmarks and rapid iteration, while clinically grounded resources are often restricted due to consent, privacy, and regulatory constraints. This fragmentation contributes to what can be viewed as a growing data crisis in computational mental health: progress increasingly depends on a small number of widely reused datasets, while new clinically meaningful data remain difficult to collect, share, and validate. Moving beyond a public–private binary will be necessary to support long-term progress, through mechanisms such as controlled evaluation servers, secure data enclaves, federated learning, or high-quality synthetic and partially de-identified derivatives.

Importantly, both evaluation design and data governance highlight a broader structural gap: the limited and often late-stage involvement of domain experts. Many datasets and benchmarks are constructed without sustained participation from clinicians or mental-health professionals, which constrains the choice of labels, tasks, and metrics, and weakens the link between model performance and real-world decision-making. Clearer documentation of data provenance, consent assumptions, and access conditions developed in collaboration with domain experts would further support responsible reuse and meaningful comparison (MacAvaney et al., 2021; Milne et al., 2016; Benton et al., 2017).

Taken together, sustainable deployment in mental health NLP requires addressing a field-wide data crisis by moving beyond isolated datasets toward shared infrastructures and sustained interdisciplinary collaboration. Including domain expertise throughout dataset design, evaluation, and governance is essential for ensuring that resulting systems are both reproducible and clinically reliable.

7 Conclusion

Our survey makes three contributions: **(1)** a dataset-centric review of computational mental health research, emphasizing how mental-health states are operationalized through existing resources; **(2)** an accompanying online resource to support dataset discovery and comparison; and **(3)** findings on labeling, representation, and evaluation that expose recurring dataset design patterns shaping what current systems can learn, generalize, and improve on. We hope that this work informs the development of future datasets and benchmarks that are not only technically sound but are also clinically grounded and capable of supporting real-world impact.

Limitations

Our survey focuses on mental health datasets and resources as they are discussed in the NLP literature, with an emphasis on how mental-health states are operationalized for modeling and evaluation. As a result, we do not aim to comprehensively cover parallel work from clinical psychology, psychiatry, or the behavioral sciences that does not intersect with computational or NLP-oriented datasets. While we draw on clinical constructs to contextualize dataset design, deeper theoretical or diagnostic discussions from the clinical literature fall outside the scope of this review.

Throughout the paper, we use terms such as *clinical grounding* or *clinical alignment* to describe the extent to which dataset labels, task formulations, and signals correspond to established clinical constructs (e.g., validated instruments or severity scales). These terms are not intended to imply diagnostic readiness or suitability for deployment in clinical settings. Our analysis evaluates datasets as research resources rather than clinical tools.

The scope of this survey is also limited to papers published in English, regardless of the language of the underlying datasets. Although we include and discuss non-English datasets when they appear in English-language publications, relevant work published exclusively in other languages may be underrepresented. This constraint reflects common practice in NLP surveys but may limit coverage of region-specific or locally developed resources.

In addition, while our review follows a structured and systematic process inspired by PRISMA-style guidelines, it is not a meta-analysis and does not include controlled experimental validation. That said, we provide quantitative summaries, comparative statistics, and structured metadata through the accompanying website released with this paper, which supports dataset comparison and exploration beyond what is feasible in print. Our findings are therefore grounded in systematic curation and quantitative resource analysis, rather than empirical model evaluation.

Finally, our disorder coverage is necessarily selective. We focus on a set of representative mental health conditions discussed in Section 2, which together capture the majority of existing datasets and research activity in NLP. We anticipate that additional conditions and emerging research areas may introduce new dataset design considerations, and we hope this survey provides a foundation for future extensions that broaden coverage and deepen clinical relevance.

Acknowledgments

This work is in part supported by the funds provided by the National Science Foundation and by DoD OUSD (R&E) under Cooperative Agreement PHY-2229929 (The NSF AI Institute for Artificial and Natural Intelligence). The views, opinions and/or findings expressed are those of the authors and should not be interpreted as representing the official views or policies of the National Science Foundation or the U.S. Government.

References

Nuredin Ali Abdelkadir, Charles Zhang, Ned Mayo, and Stevie Chancellor. 2024a. Diverse perspectives, divergent models: Cross-cultural evaluation of depression detection on Twitter. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 672–680, Mexico City, Mexico. Association for Computational Linguistics.

Nuredin Ali Abdelkadir, Charles Zhang, Ned Mayo, and Stevie Chancellor. 2024b. Diverse perspectives, divergent models: Cross-cultural evaluation of depression detection on Twitter. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 672–680, Mexico City, Mexico. Association for Computational Linguistics.

Shivam Agarwal, Ramit Sawhney, Sanchit Ahuja, Ritesh Soun, and Sudheer Chava. 2022. HYPHEN: Hyperbolic Hawkes attention for text streams. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 620–627, Dublin, Ireland. Association for Computational Linguistics.

Carlos Aguirre, Keith Harrigian, and Mark Dredze. 2021. Gender and racial fairness in depression research using social media. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2932–2949, Online. Association for Computational Linguistics.

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 the Association for Computational Linguistics: EMNLP 2022*, pages 2871–2887, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ankit Aich, Avery Quynh, Pamela Osseyi, Amy Pinkham, Philip Harvey, Brenda Curtis, Colin Depp, and Natalie Parde. 2025. Using LLMs to aid annotation and collection of clinically-enriched data in bipolar disorder and schizophrenia. In *Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2025)*, pages 181–192, Albuquerque, New Mexico. Association for Computational Linguistics.

Lindsay M. Alexander, Jasmine Escalera, Lei Ai, Charissa Andreotti, Karina Febre, Alexander Mangone, Natan Vega-Potler, Nicolas Langer, Alexis Alexander, Meagan Kovacs, Shannon Litke, Bridget O'Hagan, Jennifer Andersen, Batya Bronstein, Anastasia Bui, Marijayne Bushey, Henry Butler, Victoria Castagna, Nicolas Camacho, and 48 others. 2017. An open resource for transdiagnostic research in pediatric mental health and learning disorders. *Scientific Data*, 4:170181.

American Academy of Sleep Medicine. 2023. *International Classification of Sleep Disorders (3rd ed.)*. American Academy of Sleep Medicine.

American Psychiatric Association. 2022. *Diagnostic and Statistical Manual of Mental Disorders (5th ed., text rev.)*. American Psychiatric Association.

Silvio Amir, Glen Coppersmith, Paula Carvalho, Mario J. Silva, and Bryon C. Wallace. 2017. Quantifying mental health from social media with neural user embeddings. In *Proceedings of the 2nd Machine Learning for Healthcare Conference*, volume 68 of *Proceedings of Machine Learning Research*, pages 306–321. PMLR.

Kfir Bar, Vered Zilberstein, Ido Ziv, Heli Baram, Nachum Dershowitz, Samuel Itzikowitz, and Eiran Vadim Harel. 2019. Semantic characteristics of schizophrenic speech. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 84–93, Minneapolis, Minnesota. Association for Computational Linguistics.

Behavioral Data Lab. 2020. Empathy in mental health dataset. https://github.com/behavioral-data/Empathy-Mental-Health/tree/master/dataset. Accessed 2025-11-04.

Pierre Bellec, Carlton Chu, François Chouinard-Decorte, Yassine Benhajali, Daniel S. Margulies, and R. Cameron Craddock. 2017. The neuro bureau adhd-200 preprocessed repository. *NeuroImage*, 144(Pt B):275–286.

Adrian Benton, Glen Coppersmith, and Mark Dredze. 2017. Ethical research protocols for social media health research. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 94–102, Valencia, Spain. Association for Computational Linguistics.

Giuseppe Bersani, Elisa Polli, Giuseppe Valeriani, Daiana Zullo, Claudia Melcore, Enrico Capra, Adele Quartini, Pietropaolo Marino, Amedeo Minichino, Laura Bernabei, Maddalena Robiony, Francesco Saverio Bersani, and Damien Liberati. 2013. Facial expression in patients with bipolar disorder and schizophrenia in response to emotional stimuli: A partially shared cognitive and social deficit of the two disorders. *Neuropsychiatric Disease and Treatment*, 9:1137–1144.

Willem B Bruin, Luke Taylor, Rajat M Thomas, Jonathan P Shock, and 1 others. 2020. Structural neuroimaging biomarkers for obsessive-compulsive disorder in the enigma-ocd consortium: medication matters. *Translational Psychiatry*, 10(1):342.

Ana-Maria Bucur, Andreea Moldovan, Krutika Parvatikar, Marcos Zampieri, Ashiqur Khudabukhsh, and Liviu Dinu. 2025a. Datasets for depression modeling in social media: An overview. In *Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2025)*, pages 116–126, Albuquerque, New Mexico. Association for Computational Linguistics.

Ana-Maria Bucur, Marcos Zampieri, Tharindu Ranasinghe, and Fabio Crestani. 2025b. A survey on multilingual mental disorders detection from social media data. *Preprint*, arXiv:2505.15556.

Pasquale Bufano, Marco Laurino, Sara Said, Alessandro Tognetti, and Danilo Menicucci. 2023. Digital phenotyping for monitoring mental disorders: Systematic review. *J Med Internet Res*, 25:e46778.

Zoraida Callejas, Brian Ravenet, Magalie Ochs, and Catherine Pelachaud. 2014. A model to generate adaptive multimodal job interviews with a virtual recruiter. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3615–3619, Reykjavik, Iceland. European Language Resources Association (ELRA).

Rafaeil 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.

Edward L. Campbell, Judith Dineley, Pauline Conde, Faith Matcham, Katie M. White, Carolin Oetzmann, Sara Simblett, Stuart Bruce, Amos A. Folarin, Til Wykes, Srinivasan Vairavan, Richard J. B. Dobson, Laura Docio-Fernandez, Carmen Garcia-Mateo, Vaibhav A. Narayan, Matthew Hotopf, and Nicholas Cummins. 2023. Classifying depression symptom severity: Assessment of speech representations in personalized and generalized machine learning models. In *Interspeech 2023*, pages 1738–1742.

Centers for Disease Control and Prevention. 2021a. Age-adjusted suicide rates, united states, 2021. https://www.cdc.gov/nchs/products/databriefs/db509.htm. Accessed 2025-11-04.

Centers for Disease Control and Prevention. 2021b. Web-based injury statistics query and reporting system (wisqars): Leading causes of death reports, 2021. https://www.cdc.gov/nchs/products/databriefs/db471.htm. Accessed 2025-11-04.

Stevie Chancellor, Eric P. S. Baumer, and Munmun De Choudhury. 2019. Who is the "human" in human-centered machine learning: The case of predicting mental health from social media. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW).

Stevie Chancellor, Jessica L. Feuston, and Jayhyun Chang. 2023. Contextual gaps in machine learning for mental illness prediction: The case of diagnostic disclosures. *Proc. ACM Hum.-Comput. Interact.*, 7(CSCW2).

Stevie Chancellor, Tanushree Mitra, and Munmun De Choudhury. 2016. Recovery amid pro-anorexia: Analysis of recovery in social media. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, page 2111–2123, New York, NY, USA. Association for Computing Machinery.

Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. 2024a. Humans or LLMs as the judge? a study on judgement bias. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8301–8327, Miami, Florida, USA. Association for Computational Linguistics.

Po-Chaun Chen, Mahdin Rohmatillah, You-Teng Lin, and Jen-Tzung Chien. 2024b. Convcounsel: A conversational dataset for student counseling. In *2024 27th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*, pages 1–6.

Jenny Chim, Adam Tsakalidis, Dimitris Gkoumas, Dana Atzil-Slonim, Yaakov Ophir, Ayah Zirikly, Philip Resnik, and Maria Liakata. 2024. Overview of the CLPsych 2024 shared task: Leveraging large language models to identify evidence of suicidality risk in online posts. In *Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024)*, pages 177–190, St. Julians, Malta. Association for Computational Linguistics.

CLD2Owners. 2013. Cld2: Compact language detector 2. GitHub repository.

Juan Pablo Cobá Juárez Pegueros and Jorge Rodríguez-Arce. 2025. Physiological datasets in stress and anxiety research: A systematic review. *Biomedical Signal Processing and Control*, 108:107928.

Arman Cohan, Bart Desmet, Andrew Yates, Luca Soldaini, Sean MacAvaney, and Nazli Goharian. 2018a. 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*, pages 1485–1497, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Arman Cohan, Bart Desmet, Andrew Yates, Luca Soldaini, Sean MacAvaney, and Nazli Goharian. 2018b. 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*, pages 1485–1497, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Glen Coppersmith, Mark Dredze, and Craig Harman. 2014. Quantifying mental health signals in Twitter. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 51–60, Baltimore, Maryland, USA. Association for Computational Linguistics.

Glen Coppersmith, Mark Dredze, Craig Harman, and Kristy Hollingshead. 2015a. From ADHD to SAD: Analyzing the language of mental health on Twitter through self-reported diagnoses. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 1–10, Denver, Colorado. Association for Computational Linguistics.

Glen Coppersmith, Mark Dredze, Craig Harman, Kristy Hollingshead, and Margaret Mitchell. 2015b. CLPsych 2015 shared task: Depression and PTSD on Twitter. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 31–39, Denver, Colorado. Association for Computational Linguistics.

Glen Coppersmith, Ryan Leary, Patrick Crutchley, and Alex Fine. 2018. Natural language processing of social media as screening for suicide risk. *Biomedical Informatics Insights*, 10:1178222618792860. PMID: 30158822.

Jack A. Cummins, Daniel J. Gottlieb, Tamar Sofer, and Danielle A. Wallace. 2025. Applying natural language processing techniques to map trends in insomnia treatment terms on the r/insomnia subreddit: Infodemiology study. *Journal of Medical Internet Research*, 27:e58902. © Jack A. Cummins, Daniel J. Gottlieb, Tamar Sofer, and Danielle A. Wallace. Published in J Med Internet Res, 09 Jan 2025.

Munmun De Choudhury, Scott Counts, Eric J. Horvitz, and Aaron Hoff. 2014. Characterizing and predicting postpartum depression from shared facebook data. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '14, page 626–638, New York, NY, USA. Association for Computing Machinery.

Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2021. Predicting depression via social media. *Proceedings of the International AAAI Conference on Web and Social Media*, 7(1):128–137.

David DeVault, Ron Artstein, Grace Benn, Teresa Dey, Edward Fast, Alesia Gainer, Kallirroi Georgila, Jonathan Gratch, Arno Hartholt, Margaux Lhommet, Gale Lucas, Stacy Marsella, Fabrizio Morbini, Angela Nazarian, Stefan Scherer, Giota Stratou, Apar Suri, David Traum, Rachel Wood, and 3 others. 2014. Simsensei kiosk: A virtual human interviewer for healthcare decision support. In *Proceedings of the 13th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 1061–1062, Paris, France. International Foundation for Autonomous Agents and Multiagent Systems.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019a. 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019b. 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Sri Harsha Dumpala, Katerina Dikaios, Abraham Nunes, Frank Rudzicz, Rudolf Uher, and Sageev Oore. 2024. Self-Supervised Embeddings for Detecting Individual Symptoms of Depression. In *Interspeech 2024*, pages 1450–1454.

Sri Harsha Dumpala, Chandramouli S. Sastry, Rudolf Uher, and Sageev Oore. 2025. Test-Time Training for Speech-based Depression Detection. In *Interspeech 2025*, pages 479–483.

Doğu Erdener, Şefik Evren Erdener, and Arzu Yordaml. 2019. Auditory-visual speech perception in bipolar disorder: behavioural data and physiological predictions. In *The 15th International Conference on Auditory-Visual Speech Processing*, pages 11–15.

Carol Y Espy-Wilson. 1992. Acoustic measures for linguistic features distinguishing the semivowels/wjrl/in american english. *The Journal of the Acoustical Society of America*, 92(2):736–757.

Salvatore Fara, Orlaith Hickey, Alexandra Georgescu, Stefano Goria, Emilia Molimpakis, and Nicholas Cummins. 2023a. Bayesian networks for the robust and unbiased prediction of depression and its symptoms utilizing speech and multimodal data. In *Interspeech 2023*, pages 1728–1732.

Salvatore Fara, Orlaith Hickey, Alexandra Georgescu, Stefano Goria, Emilia Molimpakis, and Nicholas Cummins. 2023b. Bayesian networks for the robust and unbiased prediction of depression and its symptoms utilizing speech and multimodal data. In *Interspeech 2023*, pages 1728–1732.

Alice Fernandez and John Smith. 2016. Exploring linguistic correlates of social anxiety in romantic stories. *Journal of Language and Social Psychology*, 35(3):345–362.

National Center for PTSD. 2025. How common is ptsd? — adults. `https://www.ptsd.va.gov/understand/common/common_adults.asp`. Accessed: 2025-10-27.

Enrique Garcia-Ceja, Michael Riegler, Tine Nordgreen, Petter Jakobsen, Ketil J. Oedegaard, and Jim Tørresen. 2018. Mental health monitoring with multimodal sensing and machine learning: A survey. *Pervasive and Mobile Computing*, 51:1–26.

Manas Gaur, Vamsi Aribandi, Amanuel Alambo, Ugur Kursuncu, Krishnaprasad Thirunarayan, Jonathan Beich, Jyotishman Pathak, and Amit Sheth. 2021. Characterization of time-variant and time-invariant assessment of suicidality on reddit using C-SSRS. *PLOS ONE*, 16(5):e0250448.

Manas Gaur and Others. 2020. Knowledge-aware assessment of severity of suicide risk for early intervention. `https://github.com/manasgaur/Knowledge-aware-Assessment-of-Severity-of-Suicide-Risk-for-Early-Intervention`. Accessed 2025-11-04.

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 the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3123–3128, Reykjavik, Iceland. European Language Resources Association (ELRA).

Keith Harrigian, Carlos Aguirre, and Mark Dredze. 2021. On the state of social media data for mental health research. In *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*, pages 15–24, Online. Association for Computational Linguistics.

Daisy Harvey, Fiona Lobban, Paul Rayson, Aaron Warner, and Steven Jones. 2022. Natural language processing methods and bipolar disorder: Scoping review. *JMIR Ment Health*, 9(4):e35928.

Kai Hong, Christian G. Kohler, Mary E. March, Amber A. Parker, and Ani Nenkova. 2012. Lexical differences in autobiographical narratives from schizophrenic patients and healthy controls. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 37–47, Jeju Island, Korea. Association for Computational Linguistics.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*.

Sheriff Tolulope Ibrahim, Madeline Li, Jamin Patel, and Tarun Reddy Katapally. 2025. Utilizing natural language processing for precision prevention of mental health disorders among youth: A systematic review. *Computers in Biology and Medicine*, 188:109859.

Mohammed Jawwadul Islam, Mohammad Fahad Al Rafi, Pranto Podder, Aysha Siddika, Moumy Kabir, Euna Mehnaz Khan, Najmul Islam, and Saddam Mukta. 2026. Irregular sleep pattern identification and analysis from social media dataset using hybrid deep learning based attention mechanism. *Data and Information Management*, 10(1):100104.

Dan Iter, Jong Yoon, and Dan Jurafsky. 2018. Automatic detection of incoherent speech for diagnosing schizophrenia. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 136–146, New Orleans, LA. Association for Computational Linguistics.

Daniel Izmaylov, Avi Segal, Kobi Gal, Meytal Grimland, and Yossi Levi-Belz. 2023. Combining psychological theory with language models for suicide risk detection. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2430–2438, Dubrovnik, Croatia. Association for Computational Linguistics.

Shaoxiong Ji, Shirui Pan, Xue Li, Erik Cambria, Guodong Long, and Zi Huang. 2021. Suicidal ideation detection: A review of machine learning methods and applications. *IEEE Transactions on Computational Social Systems*, 8(1):214–226.

Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and Erik Cambria. 2022. MentalBERT: Publicly available pretrained language models for mental healthcare. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7184–7190, Marseille, France. European Language Resources Association.

Jaehun Jung, Faeze Brahman, and Yejin Choi. 2025. Trust or escalate: Llm judges with provable guarantees for human agreement. In *International Conference on Representation Learning*, volume 2025, pages 3101–3125.

Daniel Jurafsky and James H. Martin. 2009. *Speech and Language Processing*, 2nd edition. Prentice Hall. Chapters on Lexical and Syntactic Analysis.

Gu Eon Kang, Brian J. Mickey, Melvin G. McInnis, Barry S. Krembs, and M. Melissa Gross. 2018. Motor behavior characteristics in various phases of bipolar disorder revealed through biomechanical analysis: Quantitative measures of activity and energy variables during gait and sit-to-walk. *Psychiatry Research*, 269:93–101.

Kacie Kelly, Alex Fine, and Glen Coppersmith. 2020. Social media data as a lens onto care-seeking behavior among women veterans of the US armed forces. In *Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science*, pages 184–192, Online. Association for Computational Linguistics.

Lin Sze Khoo, Mei Kuan Lim, Chun Yong Chong, and Roisin McNaney. 2024. Machine learning for multimodal mental health detection: A systematic review of passive sensing approaches. *Sensors*, 24(2).

Seoyun Kim, Junyeop Cha, Dongjae Kim, and Eunil Park. 2023. Understanding mental health issues in different subdomains of social networking services: Computational analysis of text-based reddit posts. *Journal of Medical Internet Research*, 25:e49074. © 2023 Seoyun Kim, Junyeop Cha, Dongjae Kim, and Eunil Park. Published in the Journal of Medical Internet Research.

Atharva Kulkarni, Amey Hengle, Pradnya Kulkarni, and Manisha Marathe. 2021. Cluster analysis of online mental health discourse using topic-infused deep contextualized representations. In *Proceedings of the 12th International Workshop on Health Text Mining and Information Analysis*, pages 83–93, online. Association for Computational Linguistics.

Gina R. Kuperberg. 2010. Language in schizophrenia part 1: An introduction. *Language and Linguistics Compass*, 4(8):576–589.

GT FinTech Lab. 2022. Hyphen-acl suicide ideation detection dataset. `https://github.com/gtfintechlab/HYPHEN-ACL`. Accessed 2025-11-04.

IITP NLP Lab. 2021a. Cease / self-harm identification and intent extraction datasets. `https://www.iitp.ac.in/~ai-nlp-ml/resources.html#CEASE`. Accessed 2025-11-04.

Nitay Tech Lab. 2021b. Lgbteendataset. `https://github.com/nitaytech/LGBTeenDataset`. Accessed 2025-11-04.

Alyssa M Lanzi, Anna K Saylor, Davida Fromm, Houjun Liu, Brian MacWhinney, and Matthew L Cohen. 2023. Dementiabank: Theoretical rationale, protocol, and illustrative analyses. *American Journal of Speech-Language Pathology*, 32(2):426–438. Published online 2023 Feb 15; © 2023 by the American Speech-Language-Hearing Association.

Hannah R Lawrence, Renee A Schneider, Susan B Rubin, Maja J Matarić, Daniel J McDuff, and Megan Jones Bell. 2024. The opportunities and risks of large language models in mental health. *JMIR Ment Health*, 11:e59479.

Aziliz Le Glaz, Yannis Haralambous, Deok-Hee Kim-Dufor, Philippe Lenca, Romain Billot, Taylor C Ryan, Jonathan Marsh, Jordan DeVylder, Michel Walter, Sofian Berrouiguet, and Christophe Lemey. 2021. Machine learning and natural language processing in mental health: Systematic review. *J Med Internet Res*, 23(5):e15708.

Robert Lewis, Szymon Fedor, Nelson Hidalgo Julia, Joshua Curtiss, Jiyeon Kim, Noah Jones, David Mischoulon, Thomas F Quatieri, Nicholas Cummins, Paola Pedrelli, and Rosalind Picard. 2025. Towards the Objective Characterisation of Major Depressive Disorder Using Speech Data from a 12-week Observational Study with Daily Measurements. In *Interspeech 2025*, pages 494–498.

Chuyuan Li, Maxime Amblard, Chloé Braud, Caroline Demily, Nicolas Franck, and Michel Musiol. 2021. Investigating non lexical markers of the language of schizophrenia in spontaneous conversations. In *Proceedings of the 2nd Workshop on Computational Approaches to Discourse*, pages 20–28, Punta Cana, Dominican Republic and Online. Association for Computational Linguistics.

Baihan Lin, Guillermo Cecchi, and Djallel Bouneffouf. 2024a. Working Alliance Transformer for Psychotherapy Dialogue Classification. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, pages 64–69, Mexico City, Mexico. Association for Computational Linguistics.

Baihan Lin, Guillermo Cecchi, and Djallel Bouneffouf. 2024b. Working alliance transformer for psychotherapy dialogue classification. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, pages 64–69, Mexico City, Mexico. Association for Computational Linguistics.

Shir Lissak, Nitay Calderon, Geva Shenkman, Yaakov Ophir, Eyal Fruchter, Anat Brunstein Klomek, and Roi Reichart. 2024. The colorful future of LLMs: Evaluating and improving LLMs as emotional supporters for queer youth. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human*

Language Technologies (Volume 1: Long Papers), pages 2040–2079, Mexico City, Mexico. Association for Computational Linguistics.

Zizhou Liu, Ziwei Gong, Lin Ai, Zheng Hui, Run Chen, Colin Wayne Leach, Michelle R. Greene, and Julia Hirschberg. 2025. The mind in the machine: A survey of incorporating psychological theories in llms. *Preprint*, arXiv:2505.00003.

David E. Losada and Fabio Crestani. 2016. A test collection for research on depression and language use. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 28–39, Cham. Springer International Publishing.

Daniel M Low, Laurie Rumker, Tanya Talkar, John Torous, Guillermo Cecchi, and Satrajit S Ghosh. 2020. Natural language processing reveals vulnerable mental health support groups and heightened health anxiety on reddit during covid-19: Observational study. *Journal of Medical Internet Research*, 22(10):e22635. Published online 2020 Oct 12; © 2020 by the authors.

Saturnino Luz, Fasih Haider, Sofia de la Fuente, Davida Fromm, and Brian MacWhinney. 2021a. Detecting cognitive decline using speech only: The adresso challenge. In *Interspeech 2021*, pages 3780–3784.

Saturnino Luz, Fasih Haider, Sofia de la Fuente Garcia, Davida Fromm, and Brian MacWhinney. 2021b. Editorial: Alzheimer's dementia recognition through spontaneous speech. *Frontiers in Computer Science*, 3:780169. Published online 2021 Oct 21; Conflict of interest: none declared.

Sean MacAvaney, Anjali Mittu, Glen Coppersmith, Jeff Leintz, and Philip Resnik. 2021. Community-level research on suicidality prediction in a secure environment: Overview of the CLPsych 2021 shared task. In *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*, pages 70–80, Online. Association for Computational Linguistics.

Bubai Maji, Rajlakshmi Guha, Aurobinda Routray, Shazia Nasreen, and Debabrata Majumdar. 2024. Investigation of Layer-Wise Speech Representations in Self-Supervised Learning Models: A Cross-Lingual Study in Detecting Depression. In *Interspeech 2024*, pages 3020–3024.

Bubai Maji, Anup Kumar Roy, Shazia Nasreen, Rajlakshmi Guha, Aurobinda Routray, and Debabrata Majumdar. 2023. A novel technique for detecting depressive disorder: A speech database-based approach. In *2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 1–4.

Matteo Malgaroli, Thomas D. Hull, James M. Zech, and Tim Althoff. 2023. Natural language processing for mental health interventions: a systematic review and research framework. *Translational Psychiatry*, 13(1):309.

Aishik Mandal, Prottay Kumar Adhikary, Hiba Arnaout, Iryna Gurevych, and Tanmoy Chakraborty. 2025. A comprehensive review of datasets for clinical mental health ai systems. *Preprint*, arXiv:2508.09809.

Mayo Foundation for Medical Education and Research. 2024. Schizophrenia. `https://www.mayoclinic.org/diseases-conditions/schizophrenia/symptoms-causes/syc-20354443`. Accessed: 2025-10-29.

William U Meyerson, Sarah K Fineberg, Ye Kyung Song, Adam Faber, Garrett Ash, Fernanda C Andrade, Philip Corlett, Mark B Gerstein, and Rick H Hoyle. 2023. Estimation of bedtimes of reddit users: Integrated analysis of time stamps and surveys. *JMIR Formative Research*, 7:e38112. Published online 2023 Jan 17; Conflicts of Interest: PC is a cofounder and shareholder in Tetricus Labs, unrelated to this work.

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, San Diego, CA, USA. Association for Computational Linguistics.

Hua Min, Xia Jing, Cui Tao, Joel E Williams, Sarah F Griffin, Christianne Esposito-Smythers, and Bruce Chorpita. 2025. Directory of public datasets for youth mental health to enhance research through data, accessibility, and artificial intelligence: Scoping review. *JMIR Ment Health*, 12:e73852.

Margaret Mitchell, Kristy Hollingshead, and Glen Coppersmith. 2015. Quantifying the language of schizophrenia in social media. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 11–20, Denver, Colorado. Association for Computational Linguistics.

James C. Mundt, Peter J. Snyder, Michael S. Cannizzaro, Kara Chappie, and Dayna S. Geralts. 2007. Voice acoustic measures of depression severity and treatment response collected via interactive voice response (ivr) technology. *Journal of Neurolinguistics*, 20(1):50–64.

Ankit Murarka, Balaji Radhakrishnan, and Sushma Ravichandran. 2021. Classification of mental illnesses on social media using RoBERTa. In *Proceedings of the 12th International Workshop on Health Text Mining and Information Analysis*, pages 59–68, online. Association for Computational Linguistics.

National Institute of Mental Health (NIMH). 2025. Any anxiety disorder. In *NIMH Mental Health Statistics Reports*, pages 1–1.

National Institute of Mental Health. 2025. Bipolar disorder – statistics. `https://www.nimh.nih.gov/health/statistics/bipolar-disorder`. Accessed: 2025-10-27.

Chanjun Park, Yoonna Jang, Seolhwa Lee, Sungjin Park, and Heuiseok Lim. 2022. FreeTalky: Don't be afraid! conversations made easier by a humanoid robot using persona-based dialogue. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1242–1248, Marseille, France. European Language Resources Association.

Lotem Peled-Cohen and Roi Reichart. 2025. A systematic review of nlp for dementia: Tasks, datasets, and opportunities. *Transactions of the Association for Computational Linguistics*, 13:1204–1244.

James W. Pennebaker, Ryan L. Boyd, Kayla Jordan, and Kate Blackburn. 2015. Linguistic inquiry and word count (liwc2015). Software and dictionary, Austin, TX: Pennebaker Conglomerates.

Fabrizio Piras, Federica Piras, Yoshinari Abe, Sri Mahavir Agarwal, Alan Anticevic, Stephanie Ameis, Paul Arnold, Nerisa Banaj, Núria Bargalló, Marcelo C Batistuzzo, Francesco Benedetti, Jan-Carl Beucke, Premika S W Boedhoe, Irene Bollettini, Silvia Brem, Anna Calvo, Kang Ik Kevin Cho, Valentina Ciullo, Sara Dallaspezia, and 40 others. 2021. White matter microstructure and its relation to clinical features of obsessive-compulsive disorder: findings from the enigma ocd working group. *Translational Psychiatry*, 11(1):173. The study was partially funded by the Italian Ministry of Health (Ricerca Corrente 19, 20).

Laurin Plank and Armin Zlomuzica. 2025. Linguistic trajectories of bipolar disorder on social media. *Preprint*, arXiv:2509.10035.

Nishat Raihan, Sadiya Sayara Chowdhury Puspo, Shafkat Farabi, Ana-Maria Bucur, Tharindu Ranasinghe, and Marcos Zampieri. 2024a. MentalHelp: A multi-task dataset for mental health in social media. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11196–11203, Torino, Italia. ELRA and ICCL.

Nishat Raihan, Sadiya Sayara Chowdhury Puspo, Shafkat Farabi, Ana-Maria Bucur, Tharindu Ranasinghe, and Marcos Zampieri. 2024b. MentalHelp: A multi-task dataset for mental health in social media. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11196–11203, Torino, Italia. ELRA and ICCL.

Nairan Ramirez-Esparza, Cindy Chung, Ewa Kacewic, and James Pennebaker. 2021. The psychology of word use in depression forums in english and in spanish: Testing two text analytic approaches. *Proceedings of the International AAAI Conference on Web and Social Media*, 2(1):102–108.

Yaxuan Ren, Krithika Ramesh, Yaxing Yao, and Anjalie Field. 2025. How do we measure privacy in text? a survey of text anonymization metrics. *Preprint*, arXiv:2512.01109.

Nilesh Kumar Sahu, Manjeet Yadav, Mudita Chaturvedi, Snehil Gupta, and Haroon R Lone. 2025. Leveraging language models for summarizing mental state examinations: A comprehensive evaluation and dataset release. In *Proceedings of the 31st International Conference on Computational Linguistics (COLING 2025)*, pages 2658–2682, Abu Dhabi, UAE. Association for Computational Linguistics.

Shirin Saleem, Rohit Prasad, Shiv Vitaladevuni, Maciej Pacula, Michael Crystal, Brian Marx, Denise Sloan, Jennifer Vasterling, and Theodore Speroff. 2012. Automatic detection of psychological distress indicators and severity assessment from online forum posts. In *Proceedings of COLING 2012*, pages 2375–2388, Mumbai, India. The COLING 2012 Organizing Committee.

Kayalvizhi Sampath, Durairaj Thenmozhi, Bharathi Raja Chakravarthi, Jerin Mahibha C, Kogilavani Shanmugavadivel, and Pratik Anil Rahood. 2023. Overview of the shared task on detecting signs of depression from social media text. In *Proceedings of the Third Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 25–30, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Efsun Sarioglu Kayi, Mona Diab, Luca Pauselli, Michael Compton, and Glen Coppersmith. 2017. Predictive linguistic features of schizophrenia. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017)*, pages 241–250, Vancouver, Canada. Association for Computational Linguistics.

Moctar Abdoul Latif Sawadogo, Furkan Pala, Gurkirat Singh, Imen Selmi, Pauline Puteaux, and Alice Othmani. 2024. Ptsd in the wild: A video database for studying post-traumatic stress disorder recognition in unconstrained environments. *Multimedia Tools and Applications*, 83(14):42861–42883.

Ramit Sawhney, Harshit Joshi, Rajiv Ratn Shah, and Lucie Flek. 2021. Suicide ideation detection via social and temporal user representations using hyperbolic learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2176–2190, Online. Association for Computational Linguistics.

Philip Schmidt, Attila Reiss, Robert Duerichen, Claus Marberger, and Kristof Van Laerhoven. 2018. Introducing WESAD, a Multimodal Dataset for Wearable Stress and Affect Detection. In *Proceedings of the 20th International Conference on Multimodal Interaction (ICMI 2018)*, pages 400–408, Boulder, CO, USA. Association for Computing Machinery.

Adrian B. R. Shatte, Delyse M. Hutchinson, and Samantha J. Teague. 2019. Machine learning in mental health: a scoping review of methods and applications. *Psychological Medicine*, 49(9):1426–1448.

Ying Shen, Huiyu Yang, and Lin Lin. 2022. Automatic depression detection: an emotional audio-textual corpus and a gru/bilstm-based model. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6247–6251.

Han-Chin Shing, Suraj Nair, Ayah Zirikly, Meir Friedenberg, Hal Daumé III, and Philip Resnik. 2018. Expert, crowdsourced, and machine assessment of suicide risk via online postings. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 25–36, New Orleans, LA. Association for Computational Linguistics.

Yaara Shriki, Ido Ziv, Nachum Dershowitz, Eiran Harel, and Kfir Bar. 2022. Masking morphosyntactic categories to evaluate salience for schizophrenia diagnosis. In *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology*, pages 148–157, Seattle, USA. Association for Computational Linguistics.

Tanvir Singh and Muhammad Rajput. 2006. Misdiagnosis of bipolar disorder. *Psychiatry (Edgmont)*, 3(10):57–63.

Youngseo Son, Sean A. P. Clouston, Roman Kotov, Johannes C. Eichstaedt, Evelyn J. Bromet, Benjamin J. Luft, and H. Andrew Schwartz. 2021. World trade center responders in their own words: Predicting PTSD symptom trajectories with AI-based language analyses of interviews. *Psychological Medicine*, pages 1–9.

Vineet Srivastava, Lokesh Boggavarapu, Anthony Shin, Avisek Datta, Yingda Lu, and Runa Bhaumik. 2025. Towards understanding bipolar disorder through social media and transformer models: Challenges and insights. *medRxiv*, pages 2025–03.

Mayo Clinic Staff. 2024. Bipolar disorder – symptoms & causes. `https://www.mayoclinic.org/diseases-conditions/bipolar-disorder/symptoms-causes/syc-20355955`. Accessed: 2025-10-27.

Ian Stewart, Stevie Chancellor, Munmun De Choudhury, and Jacob Eisenstein. 2017. #anorexia, #anarexia, #anarexyia: Characterizing online community practices with orthographic variation. In *2017 IEEE International Conference on Big Data (Big Data)*, pages 4353–4361.

Suhavi, Asmit Kumar Singh, Udit Arora, Somyadeep Shrivastava, Aryaveer Singh, Rajiv Ratn Shah, and Ponnurangam Kumaraguru. 2022. Twitter-stmhd: An extensive user-level database of multiple mental health disorders. In *Proceedings of the Sixteenth International AAAI Conference on Web and Social Media (ICWSM 2022)*, volume 16, pages 1182–1184.

Shan Tang, Ryan Kriz, Seongho Cho, Sanghoon Park, Jared Harowitz, Raquel Gur, Mahesh Bhati, Daniel Wolf, Joao Sedoc, and Mark Liberman. 2021. Natural language processing methods are sensitive to sub-clinical linguistic differences in schizophrenia spectrum disorders. *npj Schizophrenia*, 7:25.

Fuxiang Tao, Anna Esposito, and Alessandro Vinciarelli. 2023. The androids corpus: A new publicly available benchmark for speech based depression detection. In *Interspeech 2023*, pages 4149–4153.

Hana Teferra and Michael Rose. 2023. Predicting generalized anxiety disorder from impromptu speech transcripts using context-aware transformer-based neural networks: Model evaluation study. *JMIR Mental Health*, 10(1):e44325.

Syauki A. Thamrin and Arbee L.P. Chen. 2024. Detection of bipolar disorder on social media data utilizing biomedical, clinical and mental health domain fine-tuned word embeddings. In *2024 IEEE 12th International Conference on Healthcare Informatics (ICHI)*, pages 612–619.

Eric Topol. 2019. High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*, 25(1):44–56.

Adam Tsakalidis, Jenny Chim, Iman Munire Bilal, Ayah Zirikly, Dana Atzil-Slonim, Federico Nanni, Philip Resnik, Manas Gaur, Kaushik Roy, Becky Inkster, Jeff Leintz, and Maria Liakata. 2022. Overview of the CLPsych 2022 shared task: Capturing moments of change in longitudinal user posts. In *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology*, pages 184–198, Seattle, USA. Association for Computational Linguistics.

Talia Tseriotou, Jenny Chim, Ayal Klein, Aya Shamir, Guy Dvir, Iqra Ali, Cian Kennedy, Guneet Singh Kohli, Anthony Hills, Ayah Zirikly, Dana Atzil-Slonim, and Maria Liakata. 2025. Overview of the CLPsych 2025 shared task: Capturing mental health dynamics from social media timelines. In *Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2025)*, pages 193–217, Albuquerque, New Mexico. Association for Computational Linguistics.

Elsbeth Turcan and Kathy McKeown. 2019a. Dreaddit: A Reddit dataset for stress analysis in social media. In *Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019)*, pages 97–107, Hong Kong. Association for Computational Linguistics.

Elsbeth Turcan and Kathy McKeown. 2019b. Dreaddit: A Reddit dataset for stress analysis in social media. In *Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019)*, pages 97–107, Hong Kong. Association for Computational Linguistics.

U.S. Department of Health and Human Services. 2025. Schizophrenia. `https://www.nimh.nih.gov/health/statistics/schizophrenia`. National Institute of Mental Health. Accessed: 2025-10-29.

Michel Valstar, Björn Schuller, Kirsty Smith, Timur Almaev, Florian Eyben, Jarek Krajewski, Roddy Cowie, and Maja Pantic. 2014. Avec 2014: 3d dimensional affect and depression recognition challenge. In *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, AVEC '14, page 3–10, New York, NY, USA. Association for Computing Machinery.

Michel Valstar, Björn Schuller, Kirsty Smith, Florian Eyben, Bihan Jiang, Sanjay Bilakhia, Sebastian Schnieder, Roddy Cowie, and Maja Pantic. 2013. Avec 2013: the continuous audio/visual emotion and depression recognition challenge. In *Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge*, AVEC '13, page 3–10, New York, NY, USA. Association for Computing Machinery.

Vasudha Varadarajan, Sverker Sikström, Oscar Kjell, and H. Andrew Schwartz. 2024. ALBA: Adaptive language-based assessments for mental health. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2466–2478, Mexico City, Mexico. Association for Computational Linguistics.

Bo Wang, Yue Wu, Niall Taylor, Terry Lyons, Maria Liakata, Alejo J. Nevado-Holgado, and Kate E.A. Saunders. 2020. Learning to detect bipolar disorder and borderline personality disorder with language and speech in non-clinical interviews. In *Interspeech 2020*, pages 437–441.

Guoxin Wang, Sheng Shi, Shan An, Fengmei Fan, Wenshu Ge, Qi Wang, Feng Yu, and Zhiren Wang. 2024. A bi-pyramid multimodal fusion method for the diagnosis of bipolar disorders. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1746–1750.

Jason Wei, Kelly Finn, Emma Templeton, Thalia Wheatley, and Soroush Vosoughi. 2021. Linguistic complexity loss in text-based therapy. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4450–4459, Online. Association for Computational Linguistics.

Lauren White, Ewan Carr, Judith Dineley, Catarina Botelho, Pauline Conde, Faith Matcham, Carolin Oetzmann, Amos Folarin, George Fairs, Agnes Norbury, Stefano Goria, Srinivasan Vairavan, Til Wykes, Richard Dobson, Vaibhav Naraya, Matthew Hotopf, Alberto Abad, Isabel Trancoso, and Nicholas Cummins. 2025. Speech Reference Intervals: An Assessment of Feasibility in Depression Symptom Severity Prediction. In *Interspeech 2025*, pages 459–463.

Annika Wiebe, Behrem Aslan, Charlotte Brockmann, Alexandra Lepartz, Dominika Dudek, Kyra Kannen, Benjamin Selaskowski, Silke Lux, Ulrich Ettinger, Alexandra Philipsen, and Niclas Braun. 2023. Multimodal assessment of adult attention-deficit hyperactivity disorder: A controlled virtual seminar room study. *Clinical Psychology & Psychotherapy*, 30(5):1111–1129. © 2023 The Authors. Clinical Psychology & Psychotherapy published by John Wiley & Sons Ltd.

Annika Wiebe, Benjamin Selaskowski, Martha Paskin, Laura Asché, Julian Pakos, Behrem Aslan, Silke Lux, Alexandra Philipsen, and Niclas Braun. 2024. Virtual reality-assisted prediction of adult adhd based on eye tracking, eeg, actigraphy and behavioral indices: a machine learning analysis of independent training and test samples. *Translational Psychiatry*, 14(1):508. Some authors received funding from BONFOR, the German Federal Ministry of Education, Medice, and other sources. EEG-based features were not included in the final model due to low predictive contribution.

Jenna Wiens, Suchi Saria, Mark Sendak, Marzyeh Ghassemi, Vincent X. Liu, Finale Doshi-Velez, Kenneth Jung, Katherine Heller, David Kale, Mohammed Saeed, Pilar N. Ossorio, Sonoo Thadaney-Israni, and Anna Goldenberg. 2019. Do no harm: a roadmap for responsible machine learning for health care. *Nature Medicine*, 25(9):1337–1340.

Akkapon Wongkoblap, Miguel A Vadillo, and Vasa Curcin. 2017. Researching mental health disorders in the era of social media: Systematic review. *J Med Internet Res*, 19(6):e228.

World Health Organization. 2022. *International Classification of Diseases (11th rev.)*. World Health Organization.

World Health Organization. 2025. Depressive disorder (depression).

Xinyi Wu, Changqing Xu, Nan Li, Rongfeng Su, Lan Wang, and Nan Yan. 2024. Depression Enhances Internal Inconsistency between Spoken and Semantic Emotion: Evidence from the Analysis of Emotion Expression in Conversation. In *Interspeech 2024*, pages 4219–4223.

Jia Xu, Tianyi Wei, Bojian Hou, Patryk Orzechowski, Shu Yang, Ruochen Jin, Rachael Paulbeck, Joost Wagenaar, George Demiris, and Li Shen. 2025. Mentalchat16k: A benchmark dataset for conversational mental health assistance. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, KDD '25, page 5367–5378, New York, NY, USA. Association for Computing Machinery.

Binwei Yao, Chao Shi, Likai Zou, Lingfeng Dai, Mengyue Wu, Lu Chen, Zhen Wang, and Kai Yu. 2022. D4: a Chinese dialogue dataset for depression-diagnosis-oriented chat. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2438–2459, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Andrew Yates, Arman Cohan, and Nazli Goharian. 2017. Depression and self-harm risk assessment in online forums. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2968–2978, Copenhagen, Denmark. Association for Computational Linguistics.

Congchi Yin, Feng Li, Shu Zhang, Zike Wang, Jun Shao, Piji Li, Jianhua Chen, and Xun Jiang. 2025. Mdd-5k: A new diagnostic conversation dataset for mental disorders synthesized via neuro-symbolic llm agents. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(24):25715–25723.

Sophie Young, Fuxiang Tao, Bahman Mirheidari, Madhurananda Pahar, Markus Reuber, and Heidi Christensen. 2025. Can Speech Accurately Detect Depression in Patients With Comorbid Dementia? An Approach for Mitigating Confounding Effects of Depression and Dementia. In *Interspeech 2025*, pages 499–503.

Tong Zhang, Annika M. Schoene, Shaoxiong Ji, and Sophia Ananiadou. 2022. Natural language processing applied to mental illness detection: A narrative review. *npj Digital Medicine*, 5(1):46.

Yuqiu Zhou, Yongjie Zhou, Yudong Yang, Yang Liu, Jun Huang, Shuzhi Zhao, Rongfeng Su, Lan Wang, and Nan Yan. 2025. Emotion-Guided Graph Attention Networks for Speech-Based Depression Detection under Emotion-Inducting Tasks. In *Interspeech 2025*, pages 469–473.

Jonathan Zomick, Sarah Ita Levitan, and Mark Serper. 2019. Linguistic analysis of schizophrenia in Reddit posts. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 74–83, Minneapolis, Minnesota. Association for Computational Linguistics.

A Paper Collection Method and Example

To systematically collect mental health datasets, we organized our literature review by mental health aspect, including *depression* (and related affective disorders), *bipolar disorder*, *anxiety* (including stress), *PTSD*, *suicidal ideation*, *schizophrenia*, and a set of additional conditions grouped as "Other" disorders, including *Attention-Deficit/Hyperactivity Disorder (ADHD), obsessive-compulsive disorder (OCD), eating disorders, borderline personality disorder, seasonal affective disorder, insomnia,* and *dementia*. This disorder-centric organization guided both paper discovery and dataset categorization throughout the survey.

We conducted keyword-based searches across a broad set of peer-reviewed venues commonly used in NLP, speech, and adjacent AI communities. These included major NLP and computational linguistics venues (ACL Anthology, COLM, COLING), speech and audio processing conferences (Interspeech, ICASSP), machine learning and artificial intelligence venues (AAAI, NeurIPS, ICML, ICLR), and selected computer vision conferences (eCVPR, ICCV, ECCV). Searches covered publications from 2001 to 2025 and were supplemented by Google Scholar queries to capture relevant interdisciplinary work. For each venue, we queried disorder-specific keywords in combination with dataset- and resource-related terms, iteratively expanding the keyword library as new terminology emerged.

Identified papers were manually reviewed to determine whether they introduced, curated, or substantially analyzed a mental health dataset. In addition to primary dataset papers, we followed citation chains and related-work sections to identify earlier resources and derivative datasets. For each dataset, we recorded key metadata, including disorder coverage, modality, task formulation, label source, and access conditions, in a shared tracking spreadsheet, which we additionally release an accompanying online resource that consolidates all 229 reviewed datasets into structured tables and interactive visualizations. (Available at `https://anonymous.4open.science/r/mental-health-datasets-resources-review-anonymize-0772` and Appendix G for details). To avoid duplication and ensure consistency, all entries were cross-checked against a centralized meta-table before inclusion.

Overall, our collection process follows a structured and systematic workflow inspired by PRISMA-style guidelines, emphasizing transparent inclusion criteria and comprehensive coverage within the defined scope. However, as discussed in Section 7, this survey is not a meta-analysis and does not include controlled experimental validation. Instead, our findings are grounded in systematic curation, quantitative summaries, and structured metadata analysis. To support transparency and reuse beyond what is feasible in print, we release an accompanying website that consolidates all reviewed datasets and provides comparative statistics and interactive visualizations.

Paper Collection Example: Bipolar Disorder As a concrete example, for bipolar disorder we searched across a broad set of sources, including the ACL Anthology,COLM, Interspeech, ACM venues, IEEE conferences, Google Scholar, and selected computer vision venues (e.g., CVPR). Keywords included "bipolar disorder," "bipolar disorder NLP," and "bipolar," with additional filtering required to remove unrelated results. In total, we identified 44 bipolar-related datasets, of which 13 originated from ACL venues and 8 from Interspeech, with the remainder distributed across mixed NLP and interdisciplinary outlets. Notably, despite surveying major computer vision and machine learning venues, we did not identify bipolar-specific datasets originating from those communities, reflecting both disorder-specific research focus and broader venue biases discussed in Section 7.

B Clinical Background and Terminology

This appendix provides brief clinical background and terminology to contextualize how mental health conditions are referenced throughout the survey. Our goal is not to offer diagnostic definitions or exhaustive clinical taxonomies, but to clarify how disorder names used in NLP datasets (e.g., *depression*, *anxiety*, *PTSD*) map—often imperfectly—onto clinical constructs.

Many conditions discussed in this paper encompass heterogeneous symptom profiles, subtypes, and severity levels (e.g., major depressive disorder vs. dysthymia, bipolar I vs. bipolar II), which are frequently collapsed or underspecified in dataset design. In the main body of the paper, we also summarize commonly used clinical distinctions and assessment practices to support interpretation of dataset labels, task formulations, and evaluation choices, and to highlight where abstraction or simplification may affect downstream modeling and claims.

B.1 Depression Disorders background info

Depression disorders are characterized by persistent low mood and loss of interest or pleasure, typically accompanied by changes in sleep or appetite, fatigue, and difficulty concentrating that interfere with daily functioning. Clinically, these conditions are diagnosed using standardized criteria in the DSM-5-TR (Diagnostic and Statistical Manual of Mental Disorders, Text Revision), a reference manual published by the American Psychiatric Association (American Psychiatric Association, 2022). A commonly studied clinical construct is the major depressive episode (MDE), which involves a period of at least two weeks of depressed mood or loss of interest/pleasure along with additional symptoms and impairment (World Health Organization, 2022). Globally, depression affects hundreds of millions of people and is a leading contributor to disability. The World Health Organization estimates that hundreds of millions of people worldwide experience depression, and notes that in high-income countries only about one-third of people with depression receive mental health treatment (World Health Organization, 2025). Motivated by the scale of unmet need, recent work in computational mental health has explored how language signals from text (e.g., social media, patient-generated text, or clinical documentation) can support depression screening or monitoring.

B.2 Bipolar background info

Bipolar disorder is characterized by extreme swings in mood, energy, and activity levels (Staff, 2024). These shifts are known as manic or hypomanic and depressive episodes, and can cause significant disruptions in daily life. The shifts from depression to mania may occur rarely or multiple times a year, and each episode usually lasts several days (Staff, 2024). This mental health disorder affects around 7 million American adults in a given year (of Mental Health, 2025). It is a lifelong disease, but can be controlled with medication. It's estimated that around 8 million Americans are affected with bipolar disorder but are undiagnosed. With this review, our goal is to organize these datasets by modality, type, and public availability into a website that is easily accessible for future research.

Also, around 69% of patients with bipolar disorder are initially misdiagnosed, and more than 30% remain misdiagnosed for 10 or more years (Singh and Rajput, 2006). In recent years, many researchers have looked into using natural language processing to improve mental health diagnosis. In computational and linguistic studies, researchers have used observable language patterns and features associated with these disorders to attempt to automate or improve clinical diagnosis.

B.3 Anxiety background info

Anxiety is an extremely common mental health disorder that is characterized by excessive nervousness and fear that can interfere with daily life. It is a normal response to stress but when it becomes severe and/or persistent, it is referred to as anxiety disorder.

It is estimated that as many as 1/3 of US adolescents and adults, as a lifetime prevalence, have experienced anxiety at one point in their lifetime. This includes a multitude of different types of anxieties including generalized anxiety disorder, panic disorder, social anxiety disorder, phobia-related disorder, among many others. Anxiety is often caused by previous trauma experienced by the individual or mental disorders in biological relatives that is passed down genetically. (National Institute of Mental Health (NIMH), 2025)

B.4 PTSD background info

It is common for people of all ages and backgrounds to be affected by PTSD, however most commonly is associated with veterans dealing with combat trauma. The challenges with diagnosing PTSD lie in the fact that the patient must describe their experiences accurately and in detail, which can be challenging due to memory and time passing in between the event and the presenting of symptoms. Additionally, PTSD can manifest differently in different patients, depending on each patient's experience. This research can greatly affect the rate of correct diagnosis for PTSD, as it is often misdiagnosed as other conditions, such as depression and anxiety.

PTSD, also known as Post Traumatic Stress Disorder, is a mental health disorder that develops after witnessing or experiencing a traumatic event. PTSD arises from experiencing or witnessing events that involve actual or threatened death, serious injury, or sexual violence (American Psychiatric Association, 2022). These events can include serious accidents, physical or sexual assault, abuse, combat, or exposure to traumatic events at work. PTSD is characterized by intrusive thoughts, avoidance behaviors, negative alterations in cognition and mood, and alterations in arousal and reactivity (World Health Organization, 2022). It often co-occurs with other conditions, such as substance abuse, depression, and anxiety disorders. PTSD affects 6% of the American adult population, about 13 million people annually (for PTSD, 2025).

B.5 Suicidal Ideation backgroud info

In the United States, suicide was the second leading cause of death in 2021 for people ages 10–24 (Centers for Disease Control and Prevention, 2021b). That year, the age-adjusted suicide rate for the total population was approximately 14.1 deaths per 100,000 (Centers for Disease Control and Prevention, 2021a). Suicide rates also vary markedly by sex: in 2022, the age-adjusted rate for males was 22.8 per 100,000, compared with 5.7 per 100,000 for females, roughly four times lower. Older men exhibit the highest suicide rates of any age-sex group, with men aged 75 years and older reaching approximately 43.9 per 100,000 in 2022. In terms of race and ethnicity, non-Hispanic American Indian or Alaska Native persons had the highest age-adjusted suicide rate in 2021 at 28.1 per 100,000, while non-Hispanic Asian persons had among the lowest rates, around 6.5 per 100,000 in 2023.

Suicidal ideation is often the starting point for what eventually leads to an individual deciding to take their life. Suicidal ideation is defined as thinking about or formulating plans for suicide consistently, to the point where it significantly and adversely affects the individual's life. The National Institute of Health (NIH) states "A helpful analogy is to view suicidal ideation as the more significant, unseen portion of an iceberg, with the act of suicide as the visible tip. This perspective emphasizes the need for early identification and targeted intervention of those with suicidal ideation to prevent progression to suicide.". Additionally, suicidal ideation has increased in recent years and has become more prevalent due to an increasingly busy and stressful world, along with continued substance abuse (including various drugs and alchohol) among the youth).

Trends. The graph below depicts the number of databases on ACL anthology that papers utilized relating to the key words "suicidal ideation" per year from 2014 to 2025. There does not seem to be a consistent increase in the number of databases that were used but there does seem to be a marked increase in the year 2025. One important thing to note is that this graphic depicts the number of unique datasets each year relating to suicidal ideation. In previous years, including 2021 and 2022, CLPsych released their own suicidal ideation datasets and many papers utilized this dataset, leading to less unique datasets for suicidal ideation being used.

B.6 Schizophrenia background info

Schizophrenia is a mental disorder that manifests itself in a number of ways, typically split into two categories: positive and negative symptoms. Positive symptoms include hallucinations, delusions and disorganized thinking and speech, while negative symptoms include reduced emotional expression, lack of motivation and social withdrawal

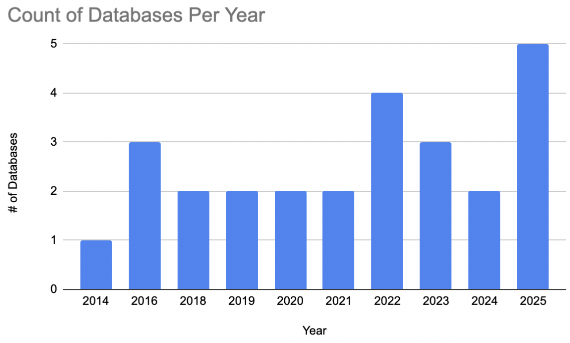


Figure 2: Number of suicidal ideation datasets per year

(Mayo Foundation for Medical Education and Research, 2024). Schizophrenia is a lifetime illness, affecting between 0.33% and 0.75% of the non-institutionalized population (U.S. Department of Health and Human Services, 2025), but can be managed with medication and therapy. Those with schizophrenia are far more prone to premature death and suicide relative to the general population (U.S. Department of Health and Human Services, 2025), creating strong incentives to improve diagnostic techniques. Some studies have attempted to address the patterns of schizophrenic speech in the hope of using natural language processing (NLP) to identify corresponding abnormalities in patients' speech.

C Detailed Trends and Observations Based on Disorders

C.1 Bipolar

As seen in Fig 3, 43/44 of the datasets collected have been created post 2014, and one is from 2002. This data is an outlier, and was a study done to analyze speech differences between different mental health patients. This study was not originally relevant to artificial intelligence, but this transcription is helpful to us and its data can be used today. Most of the studies are from 2021 or 2022, which is because this is after the introduction of LLMs, and studies related to natural language processing methods in bipolar disorder became more common.

The newer studies use methods such as transformer architecture, and models such as Google's BERT (Devlin et al., 2019b), and pretrained BERT models such as MentalBERT and ClinicalBERT (Ji et al., 2022), (Huang et al., 2019). The earlier articles (pre 2020), use analysis based on features such as LIWC analysis (Pennebaker et al., 2015), and

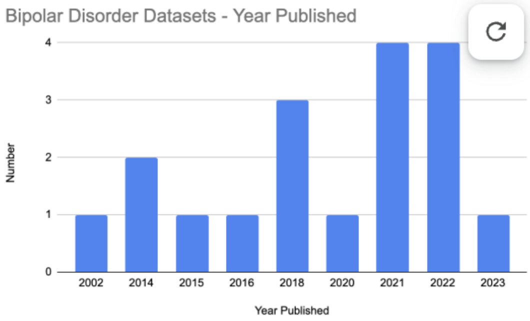


Figure 3: Datasets published by year, N=44, 43/44 (97.7%) of datasets are created post 2014

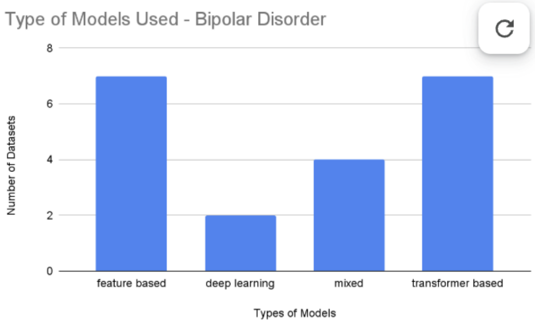


Figure 4: Types of models used to analyze bipolar disorder datasets

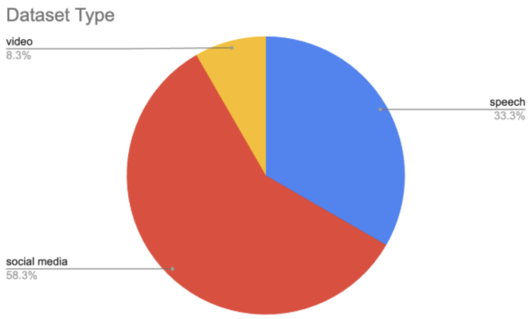


Figure 5: Type of dataset collected (social media, text, or other type such as video)

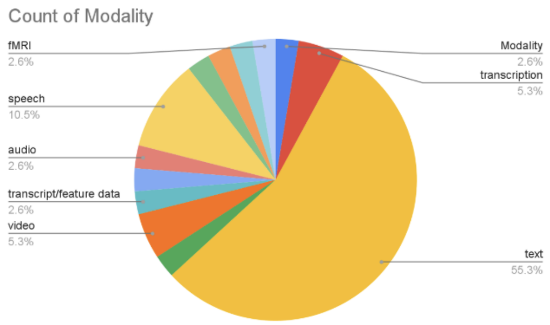


Figure 6: Count of each type of modality of the datasets collected for Bipolar Disorder

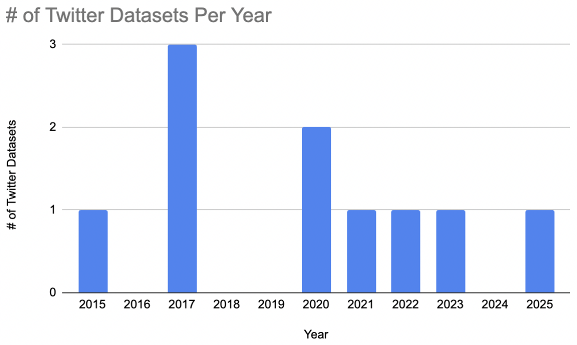


Figure 7: Number of anxiety datasets from Twitter per year

lexical and syntax analysis (Jurafsky and Martin, 2009). Fig 4 shows how the feature based models and transformer based models used are essentially the same amount of papers, but trajectories of the datasets we have collected show that transformer based models are more commonly now used for analysis.

As shown in the Fig 5, 58% of the past literature that we looked at in bipolar disorder detection has utilized social media datasets, and only 8 percent included data that was a modality different from text (video). Due to its size and availability, social media data is very prominent in computational linguistics research. Reddit and Twitter are both popular choices of data collection sites, due to their convenience, post length, informal language use, and possible anonymity. Most of the datasets available for analysis on bipolar disorder are Reddit or Twitter based datasets, with only 9 out of 30+ reviewed datasets being a type of data other than social media posts or profiles. However, only four of the 17 social media datasets collected are publicly available in their annotated and filtered form. Some are available upon request or available to redownload and filter through from the public API, but only a small few are linked in the papers. Many of the papers follow the (Coppersmith et al., 2014) method of approaching data collection, which includes filtering social media data to pick out the ones with self diagnoses, and then remove the non English posts using the Compact Language Detector (CLD2Owners, 2013).

There is a significant lack in any data that is a modality other than text or transcription (only 9% of the papers we reviewed were of a modality other than text or transcription). There are multi modal (datasets of more than one modality, combining two or more of the following modalities: audio, video, physiological, or text) datasets available for depression, but none are specific to bipolar disorder. Fig 6 shows how overwhelmingly, the majority of data is text, transcription, or speech (all based on the text modality). Very few of the datasets include fMRI, audio, or video data, and only a handful use more than one modality when collecting data. This is a big gap in research because bipolar disorder is inherently multi modal in its manifestation in patients, it affects more than just what people say. Their behavior, emotions, facial expression, and speech all are changed and affected by the disorder. Facial expression is slightly different between bipolar patients and the control group (Bersani et al., 2013), and motor behavior also differs between various phases of bipolar disorder and the control group (Kang et al., 2018). Thus, multi modal data could be beneficial to our ultimate goal of creating an agent which can assist with diagnosis. Unlike depression, which is a relatively consistent behavioral pattern, bipolar disorder involves extreme shifts in behavior which could be studied through modalities other than text. This limits the diagnostic capabilities of ML models because they may not be able to detect the difference between manic and depressive states, non verbal indicators, or facial cues. Therefore, the development of bipolar disorder specific multi modal datasets is an important next step in advancing this research.

C.2 Anxiety

From observing trends of dataset source (Fig 8 and 7), we see a decrease in the number of Twitter datasets and an increase in the number of Reddit datasets as time has progressed.

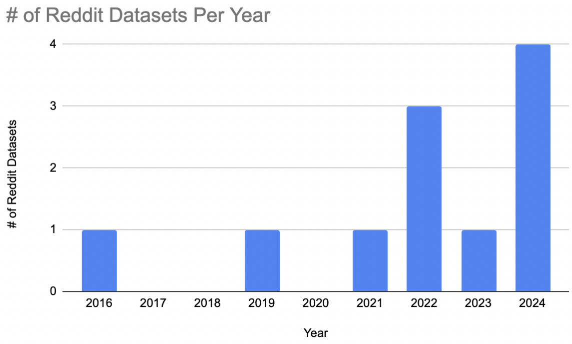


Figure 8: Number of anxiety datasets from Reddit per year

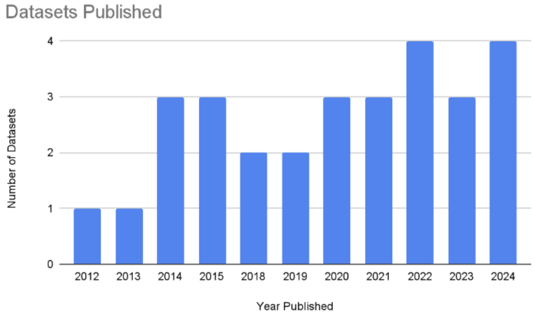


Figure 9: Number of datasets published each year

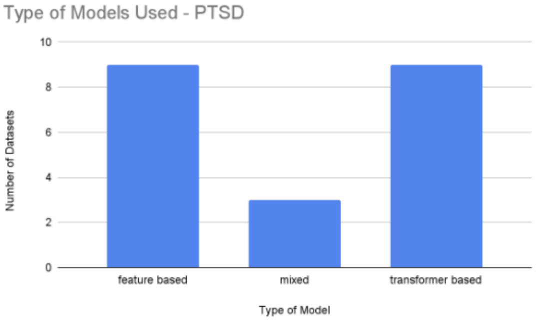


Figure 10: Type of analysis done on each dataset

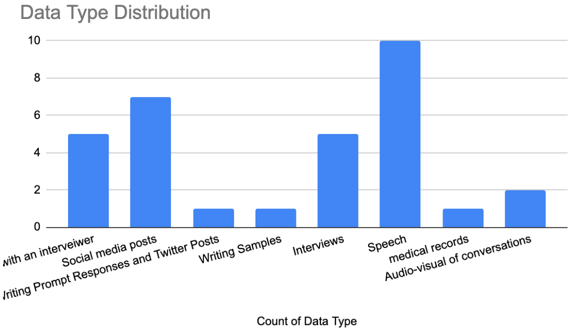


Figure 11: Distribution of data sources in schizophrenia-related studies.

C.3 PTSD

C.4 schizophrenia

Distribution of data sources in schizophrenia- related studies.

D Evaluation Case study: Depression Task Eval

Across the depression literature, task formulations concentrate on (i) binary detection over social-media timelines and user histories, (ii) severity or symptom prediction using standardized instruments such as PHQ-8/9, BDI and MADRS, and (iii) early-risk or temporal modeling where observations arrive sequentially. Evaluations follow modality and objective: social-media classification typically reports precision/recall/F1 (with occasional AUROC), while severity prediction emphasizes MAE/RMSE; early-risk and temporal setups employ time-aware metrics, such as the Early Risk Detection Error (ERDE) (Losada and Crestani, 2016). Across establiashed benchmarks, modeling has diversified from feature-engineered SVM/SVR to deep learning approaches—on AVEC 2013/2014 (Valstar et al., 2013, 2014) and DAIC-WOZ (Gratch et al., 2014) for speech/multimodal severity, within eRisk (Losada and Crestani, 2016) for early-risk detections, and in the CLPsych shared tasks (Coppersmith et al., 2015b) and RSDD (Yates et al., 2017) for social-media detection—while recent datasets extend these trends with self-supervised audio and new fusion strategies (e.g., Androids (Tao et al., 2023), CLIND (Dumpala et al., 2024), VMD (Dumpala et al., 2025), EATD (Shen et al., 2022)).

E Benchmarking and Interpretability – Feature Analysis

Benchmarking efforts typically compare traditional machine learning models such as Support Vector Machines (SVMs), Logistic Regression (LR), and Multi-Layer Perceptrons (MLPs) with advanced deep learning architectures including Convolutional Neural Networks (CNNs) (Campbell et al., 2023), Long Short-Term Memory networks (LSTMs) (Lewis et al., 2025; Fara et al., 2023b; Campbell et al., 2023; Thamrin and Chen, 2024; Dumpala et al., 2025; Zhou et al., 2025; Sampath et al., 2023), and transformer-based architectures (Lin et al., 2024b; Varadarajan et al., 2024). More recent studies also benchmark large language models such as MentalLLaMA, LLaMA (Thamrin and Chen, 2024), and clinically augmented variants of LLaMA, demonstrating state-of-the-art performance across multiple benchmark datasets.

A significant trend in current research is the comparison of features extracted from Self-Supervised Learning (SSL) models such as WavLM, HuBERT, AudioMAE, and TRILLsson with traditional acoustic feature sets including eGeMAPS, COVAREP, and Mel-spectrograms (Dumpala et al., 2024). These comparisons highlight the superiority of SSL-derived representations in capturing clinically relevant acoustic patterns.

Beyond benchmarking accuracy, interpretability is crucial in mental health applications where transparency is necessary for clinical adoption. Feature importance is examined using model-agnostic interpretability tools such as LIME and Gini impurity for Random Forest models (Srivastava et al., 2025), as well as statistical frameworks such as linear mixed-effects models (LMMs), which incorporate fixed and random effects to analyze group differences while controlling for demographic variables (Plank and Zlomuzica, 2025; Wei et al., 2021).

Additional analytical techniques include UMAP-based topic visualization, regression coefficients (Plank and Zlomuzica, 2025), Spearman correlation analysis, two-way ANOVA, and Mann–Whitney U tests (Abdelkadir et al., 2024a; Wu et al., 2024; Thamrin and Chen, 2024). Recent research has also introduced the concept of Reference Intervals (RIs) as a clinically interpretable method to define normative "healthy" ranges of acoustic features, enabling the objective detection of deviations associated with mental health disturbances (White et al., 2025).

E.1 Feature Analysis

Bipolar Methodologically, early studies relied primarily on hand-crafted linguistic features such as LIWC and lexical–syntactic analysis, whereas more recent work increasingly adopts transformer-based models including BERT, MentalBERT, and ClinicalBERT (Devlin et al., 2019b; Ji et al., 2022; Huang et al., 2019; Pennebaker et al., 2015; Jurafsky and Martin, 2009). As illustrated in Fig 4, while feature-based and transformer-based approaches appear in comparable numbers overall, there is a clear temporal shift toward transformers.

Anxiety In the literature on anxiety detection, older papers tend to rely heavily on hand-crafted linguistic feature extraction methods such as LIWC. For example, Fernandez and Smith (2016) used LIWC categories (e.g., "anxiety" and negative emotion words) to predict anxiety without employing neural networks. In contrast, newer work increasingly adopts transformer-based models; for instance, Teferra and Rose (2023) fine-tuned a transformer on speech transcripts and outperformed a LIWC-based logistic regression baseline. A smaller subset of studies employ a mixed approach that combines traditional linguistic features with deep learning methods. The distribution of these approaches can be seen in the pie chart below, which highlights the shift over time from handcrafted features to more advanced neural architectures.

PTSD When looking at all the datasets, we also noted how many datasets used feature based analysis, commonly LIWC, and how many used deep learning and transformer based architecture to analyze the datasets. What we found was that 9/37 papers use feature based, and 9/37 use transformer based and the rest are mixed, as seen in Fig 10.

F Extended discussion: Consent and Anonymization

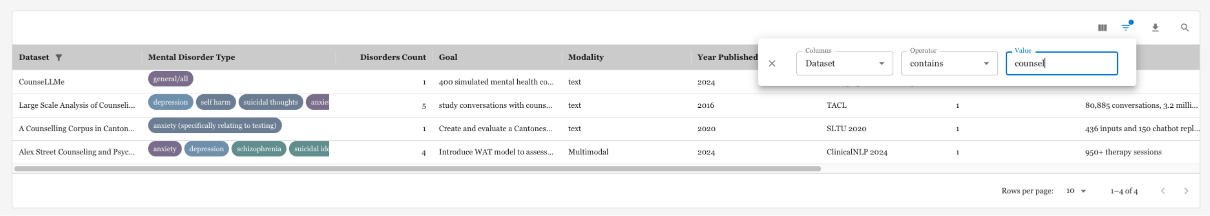
In the mental health NLP domain, data collection as well as sharing practices must carefully balance research needs with privacy, consent, and sustainability. A review of datasets used in our paper reveals a variety of approaches to user consent, anonymization, and sustainability.

User Consent Researchers have adopted different consent practices depending on data source and context. Some datasets are derived from volunteer contributions or surveys, where participants explicitly consented to share their data for research(De Choudhury et al., 2014; Gratch et al., 2014; MacAvaney et al., 2021). In contrast, many mental health datasets leverage public social media posts without individual consent for each user, under the assumption that publicly posted data can be ethically mined if handled properly and in line with platform policies(Coppersmith et al., 2015a; Milne et al., 2016; Yates et al., 2017; Cohan et al., 2018a; Turcan and McKeown, 2019a). Even when per-user consent is not feasible for large-scale corpora, researchers are urged to follow institutional ethics guidelines and obtain at least an IRB exemption or approval. It is important to remember that "public" does not equate to permission for unrestricted use. Indeed, users often share sensitive information like a depression diagnosis or suicidal thoughts in online communities with an expectation of privacy or anonymity, even if the forum is technically open-access. Because of this, some researchers advocate creative alternatives to classic informed consent, such as opt-in data donation and access-controlled evaluation environment(MacAvaney et al., 2021; Milne et al., 2016). Regardless of how data is sourced, an emerging best practice is to be transparent with users and communities. Some researchers engage directly with online health communities or moderators before using their data, or at minimum publish a public notice about data use(MacAvaney et al., 2021; Milne et al., 2016). That being said, robust anonymization remains essential.

Anonymization Mental health data can be rich in personal details such as names, locations, or health revelations, which require anonymization or de-identification before sharing or analysis. Most text-based datasets undergo some form of de-identification preprocessing. This can include removing or replacing usernames, user IDs, and mentions, stripping sensitive URLs and references to specific groups, and deleting metadata fields like timestamps if they could be identifying (Benton et al., 2017). In the 2015 CLPsych Twitter dataset, these steps were taken systematically, and an open-source tool was even released to help de-identify Twitter data in future studies(Coppersmith et al., 2015a). For forum or interview transcripts, researchers often run named entity recognition to find any person names or locations in the text and then either blank them out or substitute with generic tokens (e.g., "[NAME]", "[CITY]"). For example, one study on online breast cancer forums applied an automated de-identification to remove names/places from user messages(Benton et al., 2017). Similarly, clinical interview recordings (e.g. speech datasets for schizophrenia or depression) are usually transcribed and edited to omit names or any protected health information before researchers analyze them.

G Website

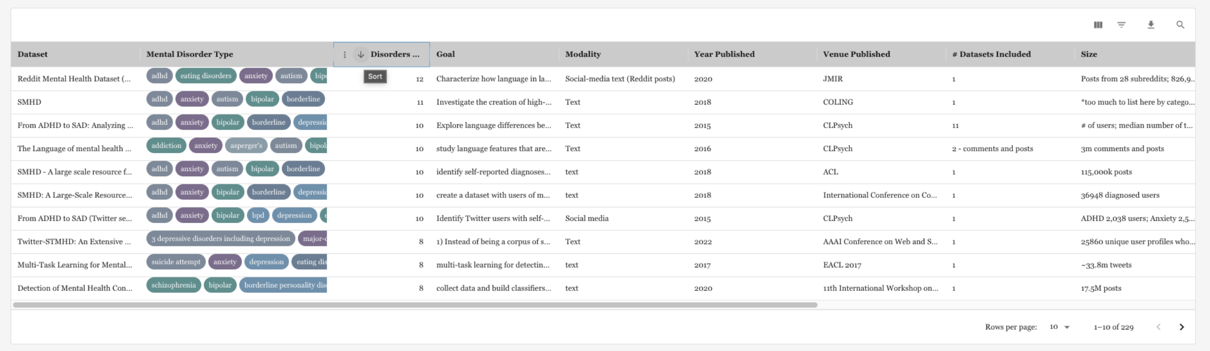
The website (`https://ziweig.github.io/mental-health-datasets-resources-review/`) serves as the interactive companion to our paper hosting the summary tables and visualizations with interactive components. Figure 12 and Figure 13 illustrate the filtering and sorting functionalities of the dataset summary table. Users can apply string and number matching, and sort the table by individual columns. In addition to the tabular interface, aggregate insights are presented through visualizations below the table. Figure 14 presents a paginated and sortable bar chart (By Paper Count or Mental Disorder Name) showing the number of papers per mental disorder group. Figure 15 shows a bar chart illustrating the number of papers by the number of disorders included. Figure 16 illustrates a line chart visualizes the number of papers collected per year.



Dataset	Mental Disorder Type	Disorders Count	Goal	Modality	Year Published			
CounseLLMe	general/all	1	400 simulated mental health co...	text	2024			
Large Scale Analysis of Counseli...	depression, self harm, suicidal thoughts, anxie	5	study conversations with couns...	text	2016	TACL	1	80,885 conversations, 3.2 milli...
A Counselling Corpus in Canton...	anxiety (specifically relating to testing)	1	Create and evaluate a Cantones...	text	2020	SLTU 2020	1	436 inputs and 150 chatbot repl...
Alex Street Counseling and Psyc...	anxiety, depression, schizophrenia, suicidal ide	4	Introduce WAT model to assess...	Multimodal	2024	ClinicalNLP 2024	1	950+ therapy sessions

Rows per page: 10 1–4 of 4

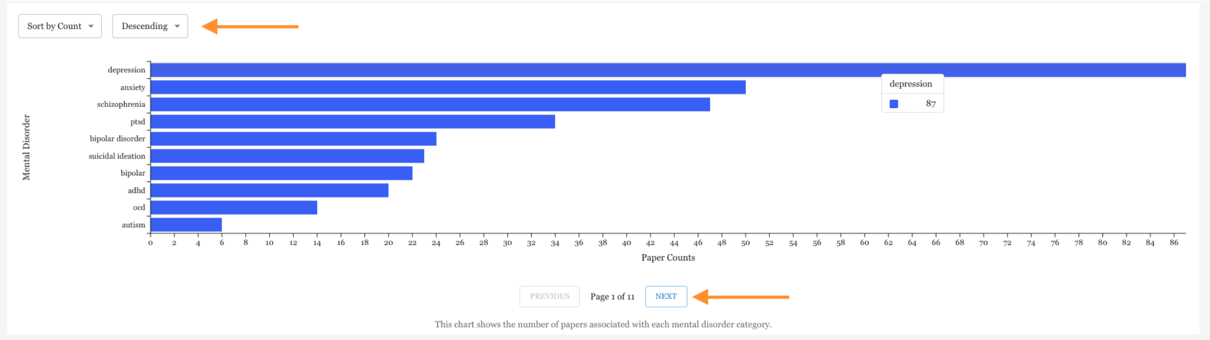
Figure 12: Filtering feature of the summary table.



Dataset	Mental Disorder Type	Disorders ...	Goal	Modality	Year Published	Venue Published	# Datasets Included	Size
Reddit Mental Health Dataset (...	adhd, eating disorders, anxiety, autism, bip	12	Characterize how language in la...	Social-media text (Reddit posts)	2020	JMIR	1	Posts from 28 subreddits; 826,9...
SMHD	adhd, anxiety, autism, bipolar, borderline	11	Investigate the creation of high-...	Text	2018	COLING	1	*too much to list here by catego...
From ADHD to SAD: Analyzing ...	adhd, anxiety, bipolar, borderline, depressi	10	Explore language differences be...	Text	2015	CLPsych	11	# of users; median number of t...
The Language of mental health ...	addiction, anxiety, asperger's, autism, bipol	10	study language features that are...	Text	2016	CLPsych	2 - comments and posts	3m comments and posts
SMHD - A large scale resource f...	adhd, anxiety, autism, bipolar, borderline	10	identify self-reported diagnoses...	text	2018	ACL	1	115,000k posts
SMHD: A Large-Scale Resource...	adhd, anxiety, bipolar, borderline, depressi	10	create a dataset with users of m...	text	2018	International Conference on Co...	1	36948 diagnosed users
From ADHD to SAD (Twitter se...	adhd, anxiety, bipolar, bpd, depression	10	Identify Twitter users with self-...	Social media	2015	CLPsych	1	ADHD 2,038 users; Anxiety 2,5...
Twitter-STMHD: An Extensive ...	3 depressive disorders including depression, major-c	8	1) Instead of being a corpus of s...	Text	2022	AAAI Conference on Web and S...	1	25860 unique user profiles who...
Multi-Task Learning for Mental...	suicide attempt, anxiety, depression, eating dis	8	multi-task learning for detectin...	text	2017	EACL 2017	1	~33.8m tweets
Detection of Mental Health Con...	schizophrenia, bipolar, borderline personality dis	8	collect data and build classifiers...	text	2020	11th International Workshop on...	1	17.5M posts

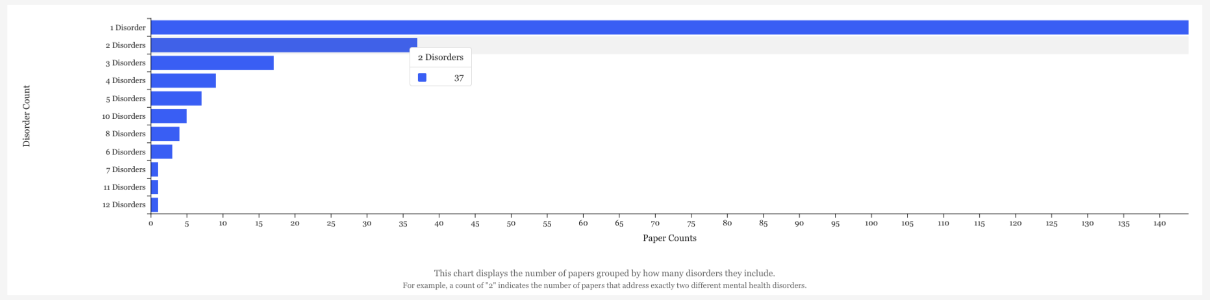
Rows per page: 10 1–10 of 229

Figure 13: Sorting feature of the summary table.



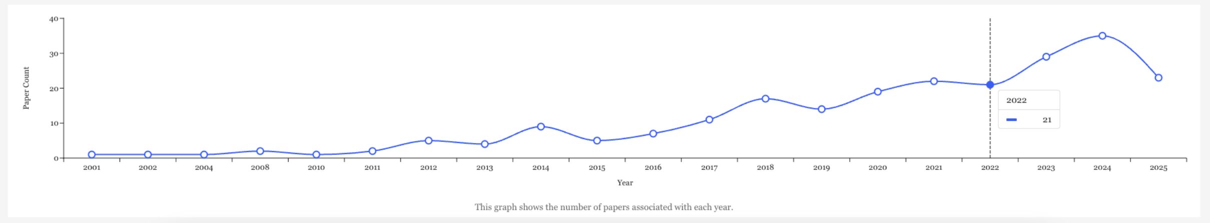
This chart shows the number of papers associated with each mental disorder category.

Figure 14: Pagination and Sorting for paper counts per mental disorder bar chart.



This chart displays the number of papers grouped by how many disorders they include.
For example, a count of "2" indicates the number of papers that address exactly two different mental health disorders.

Figure 15: Bar chart for paper count per disorder count



This graph shows the number of papers associated with each year.

Figure 16: Bar chart for paper count per disorder count