# NLP for Mental Health: A Survey

**Raja Kumar** and **Pushpak Bhattacharyya**

**Indian Institute of Technology Bombay, India**
190110070@.iitb.ac.in, pb@cse.iitb.ac.in

## Abstract

Mental health disorders, such as depression, anxiety, and bipolar disorder, are becoming increasingly prevalent worldwide. These conditions present significant challenges for individuals and healthcare systems alike, particularly given the exacerbation of these issues during the COVID-19 pandemic. This crisis has highlighted a critical shortage of trained mental health professionals globally, underscoring the urgent need for innovative detection and intervention methods. This survey comprehensively examines the application of Natural Language Processing (NLP) techniques in the detection of mental health disorders through text analysis, covering research over the past decade. It traces the evolution of methodologies from basic machine learning models to advanced neural networks, with a particular focus on the emergence of transformer-based models like BERT and GPT. These models have shown exceptional performance in understanding and generating human-like text, making them highly relevant for mental health applications.

## 1 Introduction

Mental illnesses, also known as mental health disorders, are highly prevalent worldwide and have emerged as one of the most serious public health concerns of our time (Rehm and Shield, 2019). These encompass a wide range of conditions, including depression, suicidal ideation, bipolar disorder, autism spectrum disorder (ASD), anxiety disorders, schizophrenia, and many others. The impact of these illnesses can be far-reaching, negatively influencing an individual's physical health and overall well-being, a situation further exacerbated by the COVID-19 pandemic (Santomauro et al., 2021).

According to the latest statistics, millions of people across the globe grapple with one or more mental health disorders (Rehm and Shield, 2019). However, early detection of these conditions can be instrumental in mitigating their progression and facilitating more effective treatment. Identifying the signs and symptoms of mental illness at an early stage can pave the way for timely intervention, potentially alleviating the long-term consequences and improving overall outcomes for those affected.

The far-reaching impact of mental health disorders on individuals, communities, and societies as a whole underscores the pressing need for heightened awareness, de-stigmatization, and the development of robust strategies for early detection and intervention. By addressing these challenges proactively and holistically, we can work towards creating a more supportive and inclusive environment for those affected by mental illnesses, ultimately promoting better mental health and well-being for all.

In the contemporary world, mental health plays a crucial role in a person's overall well-being. The World Health Organization (WHO)[1] highlights the intensity of this matter by reporting that globally, one in every eight individuals suffers from a mental disorder. A comprehensive study (McGrath et al., 2023) reveals that over 50% of people worldwide confront a mental health issue at some point in their lives. Despite the prevalence of mental health concerns, adequately trained professionals are scarce, which hinders access to timely and effective intervention. This motivates the need to leverage technology for automated mental disorder detection to bridge the gap between mental health needs and available resources.

We present a broad scope of mental illness detection using Natural Language Processing (NLP) that encompasses a decade of research, covering different types of mental illnesses and a variety of data sources. Our review aims to provide a comprehensive overview of the latest trends and recent NLP methodologies employed for text-based men-

[1] https://www.who.int/news-room/fact-sheets/detail/mental-disorders

tal illness detection, while also highlighting future challenges and directions. Our review seeks to answer the following questions:

- What are the main NLP trends and approaches for mental illness detection?
- Which features have been used for mental health detection in traditional machine learning-based models?
- Which neural architectures have been commonly used to detect mental illness?
- What are the main challenges and future directions in NLP for mental illness?

By addressing these questions, our review endeavors to provide a comprehensive and up-to-date understanding of the state-of-the-art in NLP for mental illness detection. We aim to elucidate the evolution of methodologies, from traditional machine learning techniques to cutting-edge neural architectures, while underscoring the potential of NLP to facilitate early detection, intervention, and support for individuals grappling with mental health challenges.

### 1.1 Problem Statement

**Problem Formulation:** Given a collection of social media posts or texts authored by individuals, our objective is to detect the presence or absence of a particular mental disorder. We first obtain a suitable representation of the textual data. Subsequently, we formulate this as a binary classification problem, where we aim to classify whether the subject exhibits signs of the mental disorder under consideration based on the linguistic cues and patterns present in their textual data. This formulation can be applied to both temporal and non-temporal data scenarios.

Detecting mental illness from text can be framed as a text classification or sentiment analysis problem (Nadkarni et al., 2011), where NLP techniques can be leveraged to automatically identify early indicators of mental health disorders. By harnessing the power of NLP, we can support early detection, prevention, and treatment efforts by analyzing the linguistic cues and patterns present in textual data.

The ability to process and interpret textual data on a large scale opens up new avenues for understanding and addressing mental health challenges. NLP models can be trained to recognize subtle linguistic markers that may signify mental health issues, enabling timely intervention and support. Moreover, the insights derived from these analyses can inform the development of more effective mental health surveillance systems (Ive et al., 2020; Mukherjee et al., 2020; Jackson et al., 2017), contributing to enhanced public health strategies and better outcomes for individuals affected by mental illnesses.

### 1.2 Motivation

A significant portion of mental health issues and their management unfolds through the medium of natural language (Sadock et al., 2015). This encompasses a wide array of elements, such as the evaluation of symptoms and signs associated with mental health disorders, as well as various forms of interventions, notably talk therapies. In this context, the wealth of information embedded within recorded textual expressions and interactions serves as a pivotal resource, enhancing both research endeavors and practical applications in the field of mental health.

Consequently, Natural Language Processing (NLP), a branch of computer science that enables computers to derive meaningful insights from free-form text (natural language), has demonstrated considerable promise as a tool to aid in mental health-related tasks. These tasks include identifying specific mental health conditions (Zhang et al., 2022), building emotion-support chatbots (Malgaroli et al., 2023), and supporting interventions (Li et al., 2023). The data utilized in these tasks is sourced from various domains, such as clinical data (Levis et al., 2021) and social media data (Skaik and Inkpen, 2020; Coppersmith et al., 2014).

By harnessing the power of NLP, researchers and practitioners can unlock the rich information contained within textual data, enabling more comprehensive analyses, accurate diagnoses, and tailored interventions. This interdisciplinary approach, which bridges the fields of computer science and mental health, holds the potential to revolutionize our understanding and management of mental health disorders, ultimately leading to improved outcomes for those affected by these conditions.

## 2 Background

People express their moods and mental states through various textual forms, including social media messages, interview transcripts, and clinical notes describing patients' conditions. In recent years, natural language processing (NLP), a branch of artificial intelligence (AI) technologies, has emerged as an essential tool for analyzing and managing large-scale textual data. NLP facilitates a wide range of tasks, such as information extraction, sentiment analysis, emotion detection, and mental health surveillance.

### 2.1 Psychological Background of Mental Disorders

A comprehensive psychological model of mental disorders posits that biological, social, and situational factors contribute to mental health issues by disrupting or unsettling various psychological processes.

This core concept is depicted in Figure 1. Traditionally, psychological approaches have maintained a distinction between events and the interpretation of these events. However, the model presented here effectively separates events from the psychological processes responsible for interpreting, mitigating, and producing consequences in response to these events. The psychological model examines the interplay among these different categories of causal variables. While biological, social, and circumstantial factors are all recognized as vital and believed to interact, the crucial aspect of this model lies in the combined impact of these interacting factors on psychological processes, ultimately giving rise to mental disorders.

A straightforward way to approach the concept of psychological disorders is by identifying behaviors, thoughts, and internal experiences that are unusual, distressing, impair one's functioning, and, in some cases, pose a risk to oneself or others as indicative of a disorder. For instance, when you ask a classmate out on a date and face rejection, feeling dejected is a normal response. However, suppose you experience extreme depression to the extent that it disrupts your daily life, affecting your appetite, sleep, self-esteem, and even contemplating self-harm or suicide. In that case, this response is atypical and may signify the presence of a psychological disorder. It's important to note that something being atypical doesn't automatically make it disordered.

### 2.2 Applications of AI and NLP to Mental Health

Recent advancements in the fields of Artificial Intelligence (AI) and machine learning offer promising solutions for addressing challenges in Mental Health Intervention (MHI). Technological and algorithmic approaches are being developed across various healthcare domains, including radiology, oncology, ophthalmology, emergency medicine, and, of particular relevance here, mental health.

One especially pertinent branch of AI is Natural Language Processing (NLP), which enables the representation, analysis, and generation of large corpora of language data. NLP facilitates the quantitative study of unstructured free-text, such as conversation transcripts and medical records, by rendering words into numerical and graphical representations. Mental Health Interventions (MHIs) heavily rely on linguistic exchanges, making them well-suited for NLP analysis (Sims, 1988), which can specify aspects of the interaction at the utterance-level detail for an extremely large number of individuals, a feat previously impossible.

NLP for MHI began with pre-packaged software tools (Tausczik and Pennebaker, 2010), followed by more computationally intense deep neural networks (Cho et al., 2014), particularly large language models (e.g., attention-based architectures like Transformers) (Vaswani et al., 2017a), and other methods for identifying meaningful trends in large amounts of data. The proliferation of digital health platforms has made these types of data more readily available, enabling the study of treatment fidelity, estimation of patient outcomes, identification of treatment components, evaluation of therapeutic alliance, and assessment of suicide risk in a transformative manner.

Moreover, NLP has been applied to mental health-relevant contexts beyond MHI, including social media (Chancellor and De Choudhury, 2020) and electronic health records (Vaci et al., 2020), demonstrating its research potential. However, questions remain about its impact on clinical practice, with a significant limiting factor being the current separation between two communities of expertise: clinical science and computer science.

Clinical researchers possess domain knowledge on MHI but face challenges in keeping up with the rapid advances in NLP. This separation is reflected in the continued reliance of clinical researchers on traditional expert-based dictionary methods (Tausczik and Pennebaker, 2010), while computer science continues to advance the state-of-the-art in large language models (Vaswani et al., 2017b). Bridging this gap between domain exper-

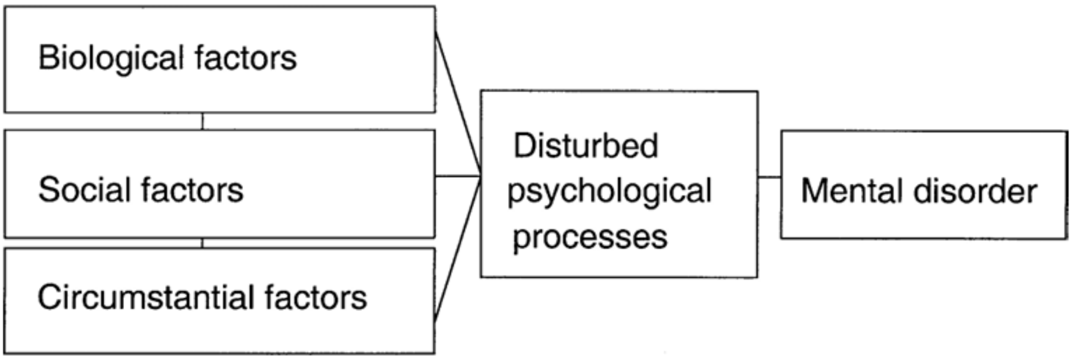

**Figure 1:** The central role of psychological processes.

tise and cutting-edge NLP techniques is crucial for fully realizing the potential of AI and machine learning in addressing mental health intervention challenges.

## 3 Datasets

### 3.1 Dataset Sources

Figure 2 illustrates the distribution of various data sources used in mental illness detection research. The majority of sources are social media posts (81%), followed by interviews (7%), electronic health records (EHRs) (6%), screening surveys (4%), and narrative writing (2%) (Zhang et al., 2022). Relational and accurate datasets are essential to train mental illness detection models. These datasets can be collected from several sources, including social media posts, screening surveys, narrative writing, interviews, and EHRs. Additionally, the datasets used for different detection tasks may vary in the types of mental illnesses they focus on and the language they use.

#### 3.1.1 Social media posts

The rise of social media platforms has provided a valuable outlet for individuals to freely express their emotions, thoughts, and personal experiences. These online spaces have become increasingly popular amongst those grappling with mental health conditions, as they offer a means to openly share their mental states, discuss related issues, and connect with others facing similar challenges. Through the publication of text messages, photographs, videos, and shared links, social media users can candidly document their journeys and seek support from virtual communities. Among the prominent platforms utilized for this purpose are Twitter, Reddit, Tumblr, Chinese microblogs, and various online forums.

**Twitter**, a widely renowned social networking service boasting over 300 million active users on a monthly basis, has emerged as a potent platform for individuals to share their perspectives through concise posts known as tweets. Users can freely compose and publish these tweets or engage with others' content by retweeting their messages. For researchers investigating mental health dynamics, Twitter offers a rich source of data that can be collected and analyzed through the platform's available application programming interfaces (APIs). For instance, Sinha et al. (2019) have harnessed Twitter data to create a manually annotated dataset aimed at identifying instances of suicidal ideation expressed through tweets. Similarly, Hu et al. (2021) employed a rule-based approach to label Twitter users' depression status based on their posted content. However, it is important to note that Twitter's privacy policies prohibit the public sharing of downloaded tweet texts, limiting the availability of datasets to only tweet identifiers, many of which may become inaccessible over time.

**Reddit**, another widely popular social media platform, facilitates the publication of posts and comments by its users. What sets Reddit apart from other data sources is its unique organization of content into distinct subreddits, each dedicated to a specific topic, such as depression or suicide. Owing to Reddit's open data policy, researchers can readily access and utilize datasets derived from this platform. For example, Yates et al. established the "Reddit Self-reported Depression Diagnosis" (RSDD) dataset (Yates et al., 2017), comprising approximately 9,000 users who self-reported depression and 100,000 control users. Similarly, the

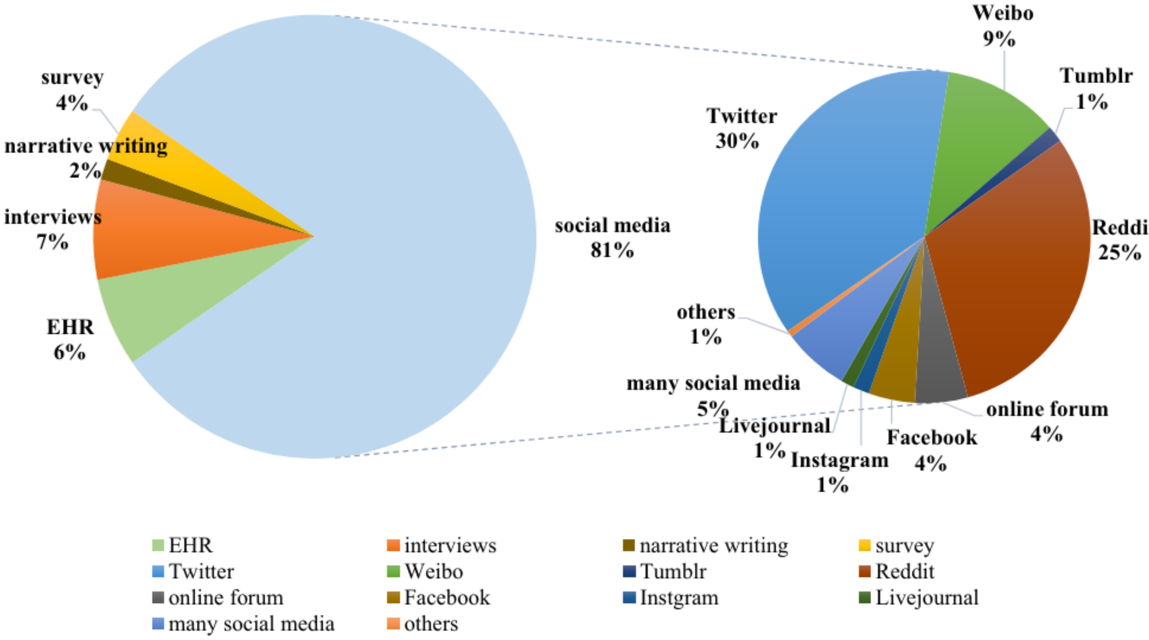

**Figure 2:** The pie chart depicts the percentages of different textual data sources based on their numbers.

CLEF risk 2019 (Naderi et al., 2019) shared task proposed an anorexia and self-harm detection task based on data sourced from the Reddit platform.

**Online forums**, also referred to as online communities, provide a dedicated space for individuals to discuss their mental health conditions and seek support from others facing similar challenges. These forums can take various forms, including chat rooms and discussion boards, such as Recoveryourlife and Endthislife. Researchers have leveraged data from these online forums to develop and train models for detecting psychological distress. For instance, Saleem et al. (2012) designed a psychological distress detection model utilizing discussion threads downloaded from an online forum dedicated to veterans. Additionally, Franz et al. (2020) employed text data from TeenHelp.org, an Internet support forum, to train a system for detecting self-harm.

### 3.1.2 Electronic Health Records (EHRs)

Electronic Health Records (EHRs) represent a rich and invaluable source of secondary healthcare data, meticulously documenting patients' historical medical records (Menachemi and Collum, 2011). EHRs often encompass a diverse array of data types, including patient profile information, medication records, diagnosis histories, and medical images. Notably, in the context of mental illness, EHRs frequently incorporate clinical notes written in narrative form (Kho et al., 2013), chronicling the patient's condition and treatment journey. The wealth of information contained within EHRs makes them well-suited for the application of Natural Language Processing (NLP) techniques to assist in disease diagnosis and analysis. Researchers have successfully utilized EHR datasets for tasks such as suicide risk screening (Downs et al., 2017), identification of depressive disorders (Kshatriya et al., 2021), and prediction of mental health conditions (Tran and Kavuluru, 2017).

### 3.1.3 Interviews

Another approach to detecting mental illness involves conducting interviews with participants and subsequently analyzing the linguistic information extracted from transcribed clinical interviews. Several notable datasets have been compiled using this method, including the DAIC-WoZ depression database (Ringeval et al., 2017), which comprises transcriptions of interviews with 142 participants. Additionally, the AViD-Corpus (Valstar et al., 2014) encompasses data from 48 participants, while the schizophrenic identification corpus was collected through interviews with 109 individuals (Voleti et al., 2019).

### 3.1.4 Screening surveys

To evaluate participants' mental health conditions, researchers often employ screening surveys or questionnaires designed for clinician-patient diagnosis or self-assessment purposes. These surveys are distributed to participants through crowdsourcing platforms like Crowd Flower and Amazon's Mechanical Turk, or via online platforms. Once completed, the collected data is labeled and analyzed. The content of these surveys varies to assess different psychiatric symptoms and conditions. For the evaluation of depression, widely used questionnaires include the Patient Health Questionnaire (PHQ-9) (Tlachac et al., 2019) and the Beck Depression Inventory (BDI) (Stankevich et al., 2020), both aimed at assessing the severity of depressive symptoms. The Center for Epidemiological Studies Depression Scale (CES-D) questionnaire (Wongkoblap et al., 2018), comprising 20 multiple-choice questions, is another instrument designed specifically for testing depression. In the context of suicide ideation, researchers often utilize questionnaires such as the Holmes-Rahe Social Readjustment Rating Scale (SRRS) (Delgado-Gomez et al., 2012) or the Depressive Symptom Inventory-Suicide Subscale (DSI-SS) (von Glischinski et al., 2016).

### 3.1.5 Narrative writing

In addition to data sourced from social media platforms, EHRs, interviews, and screening surveys, researchers have also utilized other types of narrative texts for mental health analysis. These include writings produced by participants as part of specific experiments or personal narratives that were not initially published on social media platforms. For instance, one study involved asking children to write stories about a time when they faced a problem or conflict with others. The researchers then analyzed these personal narratives to detect signs of Autism Spectrum Disorder (ASD) (Hilvert et al., 2020). Another study conducted a case analysis of Greek poetry from the 20th century with the aim of predicting suicidal tendencies among the poets based on their written works (Zervopoulos et al., 2019). These varied sources of data, ranging from social media posts and EHRs to interviews, surveys, and narrative writings, provide researchers with a wealth of information to develop and refine natural language processing techniques for mental health analysis and support.

## 3.2 Available Datasets

In the past, the research community has witnessed the utilization of widely available datasets such as the CLPsych shared task (Coppersmith et al., 2015), the Reddit Self-reported Depression Diagnosis dataset (Yates et al., 2017), and the Language of Mental Health dataset (Gkotsis et al., 2016), as well as the early risk prediction on the Internet (eRISK) data from the CLEF Forum (Losada et al., 2018). Only a handful of datasets are available in the public domain, with many being either reproducible or available upon request. Researchers encounter more than a dozen new datasets for predicting mental health conditions based on social media data every year.

Due to the limited availability of these datasets, this section aims to highlight the most popular and reproducible datasets or those that can be obtained through requests or signed agreements. We will delve into the details of each dataset, providing a comprehensive overview of the resources available for mental health analysis research.

The scarcity of publicly accessible datasets presents a significant challenge for researchers in this field. However, by focusing on the most widely used and accessible resources, we can facilitate knowledge sharing, reproducibility, and collaborative efforts toward advancing mental health analysis techniques.

**CLPsych Shared Task Dataset:** The CLPsych dataset (Coppersmith et al., 2015) contains three modules, namely, DepressionvControl (DvC), PTSDvControl (PvC), and DepressionvPTSD (DvP), which are available via signed agreement. Academic researchers must sign a confidentiality agreement to ensure the privacy of the data when using this dataset.

**Multimodal Dictionary Learning (MDDL):** The MDDL (Shen et al., 2017) is a depression detection dataset comprising D1, D2, and D3 modules. The Depression Dataset D1 is constructed using tweets from 2009 and 2016, where users were labeled as depressed if their anchor tweets satisfied the strict pattern "(I'm/I was/I am/I've been) diagnosed depression". The Non-Depression Dataset D2 is constructed in December 2016, where users were labeled as non-depressed if they had never posted any tweet containing the character string "depress". Although D1 and D2 are well-labeled, the depressed users in D1 are too few, thus, a larger unlabeled Depression-candidate Dataset D3 is constructed for depression behavior discovery, which contains much more noise.

**Reddit Self-reported Depression Diagnosis (RSDD)** The RSDD dataset (Yates et al., 2017) contains the Reddit posts of approximately 9,000 users who have claimed to have been diagnosed with depression ("diagnosed users") and approximately 107,000 matched control users. The introduction of the Reddit dataset has made a significant contribution, which has been utilized by many existing studies.

**Self-Reported Mental Health Diagnoses (SMHD) Dataset:** Similar to the RSDD dataset, the SMHD dataset (Cohan et al., 2018) can be obtained via signed agreement as per the privacy policy of the data. The dataset consists of Reddit posts from users diagnosed with one or several of nine mental health conditions ("diagnosed users"), and matched control users. This dataset has been used by a few studies in the literature and is related to multiple mental health conditions instead of just depression.

**eRISK:** The eRISK dataset (Losada et al., 2019) is available online for experiments and analysis to meet the targets of a shared task for a few years. The dataset for early risk detection by the CLEF Lab is provided to solve the problems of detecting depression, anorexia, and self-harm over the past few years.

**Pirina:** A new dataset, named Pirina (Triantafyllopoulos et al., 2023), is proposed and available online for research purposes. A filtered data is extracted from the Reddit social media platform for the depression detection task. Although this dataset is not actively maintained, it can be extracted and used for pilot studies.

**Ji:** A new Reddit dataset (Ji et al., 2018, 2020) of 5,326 suicidal posts out of 20,000 posts and 594 Suicidal Tweets out of 10,000 Tweets were extracted for experiments and evaluation of the proposed classification approach for suicidal risk detection. This dataset is referred to as the Ji dataset in this study and is available upon request.

**Sina Weibo:** Another dataset (Liao et al., 2013), proposed for the public domain and remains unnamed, is given the name of the social media platform, Sina Weibo, to refer to it in this study. The dataset comprises 3,652 users with suicidal tendencies and 3,677 users without suicidal risk, extracted from Sina Weibo, a Chinese social media platform.

**Dreaddit:** Dreaddit (Turcan and McKeown, 2019) is a new text corpus of lengthy multi-domain social media data for the identification of stress. This dataset consists of 190K posts from five different categories of Reddit communities; the authors additionally label 3.5K total segments taken from 3K posts using Amazon Mechanical Turk. The lexical features used in this dataset are the Dictionary of Affect in Language, LIWC features, and the patterns sentiment library; syntactic features like unigrams and bigrams, the Flesch-Kincaid Grade level, and the automated reliability index; social media features like timestamp, upvote ratio, karma (upvote–downvote), and the total number of comments.

**Suicide Risk Assessment using Reddit (SRAR):** The SRAR dataset (Gaur et al., 2019) is available in the public domain. The dataset is composed of 500 Redditors (anonymized), their posts, and domain expert annotated labels. The SRAR is used along with different lexicons built from the knowledge base associated with mental health like SNOMED-CT, ICD-10, UMLS, and Clinical Trials. This dataset has been recently used, and the research community looks forward to utilizing it soon to enhance the proposed techniques.

**Aladaug:** This dataset (Aladağ et al., 2018) was built by Aladaug during his study on the identification of suicidal tendencies from social media posts. Since no name was given to this dataset, it is referred to as Aladaug in this study. Among 10,785 posts, 785 were manually labeled for this study. This dataset is available upon request from the authors.

**The University of Maryland Reddit Suicidality Dataset (UMD-RD):** The UMD-Reddit Dataset (Shing et al., 2018; Zirikly et al., 2019) contains one sub-directory with data pertaining to 11,129 users who posted on SuicideWatch, and another for 11,129 users who did not. For each user, there is full longitudinal data from the 2015 Full Reddit Submission Corpus. The UMD-Reddit dataset has been actively used by academic researchers since 2019, as it is available via signed agreement.

**GoEmotion:** The GoEmotion dataset (Demszky et al., 2020) contains 58K carefully curated comments extracted from Reddit, with human annotations to 27 emotion categories or Neutral. It also contains a filtered version based on rater agreement, which includes a train/test/validation split. This dataset was proposed in 2020 for emotion detection and has been used to validate the scalability of the proposed models for stress detection.

**SDCNL Dataset:** The SDCNL (Haque et al., 2021) dataset was collected using the Reddit API and scraped from two subreddits, r/SuicideWatch and r/Depression, containing a total of 1,895 posts. Two fields were utilized from the scraped data: the original text of the post as inputs, and the subreddit it belongs to as labels. Posts from r/SuicideWatch are labeled as suicidal, and posts from r/Depression are labeled as depressed.

**CAMS:** CAMS stands for Causal Analysis for Mental illness in Social media posts. The introduction of the CAMS dataset enables academic researchers to perform causal inference, causal explanation extraction, and causal categorization. The dataset contains 5,051 samples and categorizes each sample into one of the five different causal categories, namely, bias/abuse, jobs and careers, medication, relationships, and alienation. This dataset is publicly available (Garg et al., 2022).

**RHMD:** The RHMD stands for a Real-world Dataset for Health Mention classification on Reddit data (Naseem et al., 2022). The health mention is defined as a problem to find symptoms and understand its semantics. These semantics specify the contextual perspective of a given symptom in texts. Every sample of this dataset categorizes a given post into five categories: health mention, non-health mention, hyperbolic mention, figurative mention, and uninformative.

**Kayalvizhi:** A unique dataset (S et al., 2022) that not only detects depression from social media but also analyzes the level of depression. Initially, 20,088 instances of postings data were annotated, out of which 16,613 instances were found to be mutually annotated instances by the two judges, and thus they were considered as instances of the dataset with their corresponding labels.

## 4 Coverage of Papers

In this section, we discuss traditional machine learning and deep learning approaches for mental illness detection, including temporal representation models. We highlight the shift from manual feature extraction in ML models to automated, advanced feature extraction in DL models like CNNs, RNNs, and transformers. We also explore the emerging role of large language models, such as GPT-3.5, GPT-4, and LLaMA, in enhancing mental health applications through techniques like few-shot prompting and fine-tuning.

### 4.1 Machine Learning based Approach

Traditional machine learning methods, such as Support Vector Machines (SVM), Adaptive Boosting (AdaBoost), and Decision Trees, have long been employed for various NLP downstream tasks, including mental illness detection. These methods typically follow a pipeline approach involving data pre-processing, feature extraction, modeling, optimization, and evaluation. Key to training effective ML models is the selection of main contributing features, which also help identify key predictors of illness.

Linguistic features encompass syntactic features like Part-of-Speech (POS) tagging, which categorizes words based on grammatical use and function (Birjali et al., 2017; Trifan et al., 2020; Briand et al., 2018), and dependency parsing, which examines the grammatical structure of sentences (Tadisetty and Ghazinour, 2021; Doan et al., 2019). Lexicon-based features include the Bag-of-Words (BoW) model, a basic text representation method (Trifan and Oliveira, 2019; Lin et al., 2017b; Chomutare, 2014), and measures of lexical diversity and density, which assess unique vocabulary usage and the proportion of content words (Voleti et al., 2019). Emotion features involve sentiment scores, quantifying the sentiment polarity of texts using tools like VADER, SenticNet, and AFINN lexicons (Leiva and Freire, 2017; Stephen and Prabu, 2019), and emotion scores, which gauge the emotional intensity in texts using resources such as the NRC Affect Intensity Lexicon (Guntuku et al., 2018; Delahunty et al., 2018). Topic features are derived from topic modeling algorithms like Latent Dirichlet Allocation (LDA), Latent Semantic Analysis (LSA), and Non-negative Matrix Factorization (NMF) to extract topics from texts (Shickel et al., 2020; Hwang et al., 2020; Fodeh et al., 2019). The Linguistic Inquiry and Word Count (LIWC) tool automatically extracts linguistic styles from texts by calculating the percentage of words in various categories, including linguistic, social, and affective (Islam et al., 2018; Su et al., 2018).

Statistical features are divided into statistical corpus features and vector-based features. Statistical corpus features include n-grams, contiguous sequences of n words (Shickel et al., 2020; He et al., 2017), term frequency-inverse document frequency (TF-IDF), which reflects a word's importance in a document (Boag et al., 2021; Adamou et al., 2018), and length statistics, which measure the length of posts, documents, or average sentences (Saleem et al., 2012; Trifan and Oliveira, 2021). Vector-based features include word embeddings, which are vector representations of words like word2vec and GloVe (Lin et al., 2017b; Guntuku et al., 2018), and document embeddings, which are vector representations of entire documents (Bandyopadhyay et al., 2019).

Domain knowledge features consist of conceptual features such as the Unified Medical Language System (UMLS), which provides key terminology, coding standards, and resources related to biomedical information (Zhong et al., 2018), and linguistic dictionaries that contain words related to mental health illnesses (Huang et al., 2019; Lv et al., 2015).

Auxiliary features encompass social, behavioral, time, and user profile features. Social behavioral features include measures of social connectivity, such as the number of followers, friends, and communities joined on social media (Dao et al., 2014), and user behaviors, like the frequency of comments and forwards (Hwang et al., 2020; Katchapakirin et al., 2018). Time features focus on time-related aspects like sending time and time intervals (Guntuku et al., 2019; Zhao et al., 2015). User's profile features include individual information on social networks (Chang and Tseng, 2020; Tong et al., 2022).

Machine learning models for mental illness detection often combine various extracted features, predominantly using supervised learning methods. These methods include SVM, AdaBoost, k-Nearest Neighbors (KNN), Decision Trees, Random Forests, Logistic Model Trees (LMT), Naive Bayes (NB), Logistic Regression, XGBoost, and ensemble models. Supervised learning is advantageous due to its ability to learn patterns from labeled data, ensuring better performance. However, accurately labeling large datasets is time-consuming and challenging, although some methods help reduce the burden of human annotation.

To address the limitations of relying on labeled data, unsupervised learning methods are used to discover patterns from unlabeled data, such as through clustering or using LDA topic models. These unsupervised models often extract additional features for developing supervised learning classifiers. Some research has also explored semi-supervised learning, which combines large amounts of unlabeled data as additional information, including methods like semi-supervised topic modeling over time and classic semi-supervised algorithms like YATSI and LLGC.

## 4.2 Deep Learning based Approach

Deep learning methods offer significant advancements over traditional machine learning by automating feature extraction, thereby eliminating the need for extensive feature engineering. This automation enables models to capture valuable features more effectively, leading to notable improvements across various fields, including computer vision, natural language processing (NLP), and signal processing. Recently, deep learning techniques have shown superior performance in detecting mental illness from text compared to traditional machine learning methods.

Deep learning frameworks typically consist of two layers: an embedding layer and a classification layer. The embedding layer converts sparse one-hot encoded vectors into dense vectors that preserve semantic and syntactic information, enhancing the training of deep learning models. Various embedding techniques are used, such as ELMo, GloVe, word2vec, and contextual language representations like BERT and ALBERT.

Based on the structure of their classification layers, deep learning methods for mental illness detection can be categorized into four main types: convolutional neural networks (CNN)-based methods, recurrent neural networks (RNN)-based methods, transformer-based methods, and hybrid-based methods that combine different neural network structures.

### 4.2.1 CNN-Based Models

Standard CNN structures include convolutional layers and pooling layers, followed by fully-connected layers. Some studies have used standard CNNs combined with features like LIWC, TF-IDF, BOW, and POS to build classification models (Gaur et al., 2019; Boukil et al., 2019; Phan et al., 2020; Wang et al., 2018; Trotzek et al., 2018; Obeid et al., 2019). For example, a hierarchical MGL-CNN model was proposed to capture sentiment information (Rao et al., 2020), and another framework combined CNN with a graph model to leverage tweet content and social interaction information (Lin et al., 2017a).

### 4.2.2 RNN-Based Models

RNNs are beneficial for sequential data such as text because their architecture allows previous outputs to be used as inputs. LSTM and GRU models, which address the vanishing gradient problem of traditional RNNs, are commonly used in

NLP. Numerous studies have utilized LSTM or GRU (Tran and Kavuluru, 2017; Ghosh and Anwar, 2021; Ahmed et al., 2021; Wu et al., 2020; Zogan et al., 2022; Yao et al., 2021), some incorporating attention mechanisms to highlight significant word information. Hierarchical attention networks based on LSTM or GRU structures have also been developed to better exploit different levels of semantic information (Sekulić and Strube, 2020). Other approaches include transfer learning (Rutowski et al., 2021, 2020), multi-task learning (Ghosh et al., 2022; Dinkel et al., 2019; Zhou et al., 2017; Wang et al., 2020b), reinforcement learning (Gui et al., 2019b,a), and multiple instance learning (Wongkoblap et al., 2019b,a). Additionally, Sawhney et al. (2021b) introduced an ordinal hierarchical LSTM attention model called SISMO.

#### 4.2.3 Transformer-Based Models

Transformer architectures effectively solve long-range dependencies using attention mechanisms. For instance, the C-Attention network uses a transformer encoder block with multi-head self-attention and convolution processing (Wang et al., 2021). Other researchers have developed TransformerRNN with multi-head self-attention (Zhang et al., 2021). Transformer-based pre-trained language models, including BERT, DistilBERT, Roberta, ALBERT, BioClinical BERT, XLNET, and GPT, have proven the potential of large-scale pre-training models for mental illness detection (Haque et al., 2020; Chaurasia et al., 2021; Malviya et al., 2021; Murarka et al., 2020; Wang et al., 2020a).

#### 4.2.4 Hybrid-Based Models

Hybrid methods combine multiple neural networks to enhance mental illness detection. For example, hybrid frameworks combining CNN and LSTM models capture both local and long-dependency features, outperforming individual CNN or LSTM classifiers (Gaur et al., 2021; Tadesse et al., 2019; Zhou et al., 2020; Deshpande and Warren, 2021; Solieman and Pustozerov, 2021). The STATENet model (Sawhney et al., 2020), which includes an individual tweet transformer and a Plutchik-based emotion transformer, jointly learns linguistic and emotional patterns. Inspired by improved performance using sub-emotion representations, a deep emotion attention model was presented (Aragón et al., 2020; Lara et al., 2021; Aragon et al., 2021), incorporating sub-emotion embedding, CNN, GRU, and an attention mechanism. The PHASE model (Sawhney et al., 2021a) learns the chronological emotional progression of users through a time-sensitive emotion LSTM and Hyperbolic Graph Convolution Networks, using BERT fine-tuned for emotions and a heterogeneous social network graph. The Events and Personality traits for Stress Prediction (EPSP) model (Li et al., 2021) is a joint memory network for learning the dynamics of user's emotions and personality. Hyperbolic graph convolution networks (Sawhney et al., 2021c) combine hyperbolic graph convolutions with the Hawkes process to learn the historical emotional spectrum of a user.

### 4.3 Temporal Representation based Approach

Recent research leverages longitudinal data to capture unique patterns of emotional transitions in mental health patients. These approaches typically process $m$ words (Trotzek et al. (2018); Uban et al. (2021); Orabi et al. (2018)) or $n$ posts (Ragheb et al. (2019); Mitchell et al. (2015)) sequentially, using chunking and majority voting for classification. An alternative method involves concatenating all posts related to a specific subject for feature extraction (Aguilera et al. (2021); Jamil (2017)). However, both methods fail to incorporate the temporal variations between posts, as chunking and majority voting do not account for these variations.

Some studies align more closely with approach to temporal representation of social media posts. For instance, De Choudhury et al. (2013) analyzed a user's daily tweets to derive behavioral measures such as engagement, ego network, emotion, linguistic style, depressive language, and demographics. These measures were compiled daily, creating time series data over an entire year of Twitter activity. Nevertheless, irregular or sporadic tweeting patterns can hinder accurate behavioral analysis over time. Similarly, Reece et al. (2017) employed state-space temporal analysis with daily and weekly time windows to detect depression, but their reliance on low-level features like tweet counts, average word count, and part-of-speech counts lacked the semantic depth needed for properly representing the emotional nuances of human language. Chen et al. (2020) experimented with various time window sizes to create time series representations of subjects' mood profiles, yet their sentiment retrieval was limited by a word-counting approach to identify positive and negative affects.

Guo et al. (2021) used a pre-trained BERT model fine-tuned on the "SemEval shared task on Affect in Tweets" data to classify emotions such as joy, sadness, anger, and fear in sentences. These sentences were represented by binary vectors indicating the presence of each emotion, and the vectors were aggregated to form emotional representations for posts. Emotional transition matrices, derived from user data through sliding windows and Markov's assumption, were used as features for training classifiers.

Recent advances in language-based models have furthered the identification of mental disorders. For example, Ji et al. (2022) pre-trained BERT specifically for the mental health domain by collecting data from Reddit, comprising approximately 13.67 million sentences. This computationally expensive process took over eight days and utilized four Tesla V100 GPUs. Aragon et al. (2023) implemented a two-stage adaptation of BERT, which reduced resource utilization. The first stage fine-tuned the model to match social media language using the Reddit TIFU dataset (about 120k text-summary pairs), as introduced by Kim et al. (2019). The second stage adapted the model to the mental health domain using a specialized dataset from Reddit, comprising over 105k posts related to mental disorders from mental health subreddits. This fine-tuning approach proved proficient in detecting mental disorders by analyzing the distinct writing styles associated with them.

### 4.4 Large Language Models in Mental Health Care

A significant body of existing work has focused on directly prompting large language models such as GPT-3.5 or GPT-4 for mental health applications without additional training. Some examples include depression detection (Bao et al., 2023; Hayati et al., 2022), suicide detection (Zhou et al., 2023), cognitive distortion detection (Chen et al., 2023), and relationship counseling (Vowels et al., 2023). In these studies, large language models (LLMs) function as intelligent chatbots, engaging with users to provide a range of mental health services, including analysis (Ma et al., 2023; Yang et al., 2023a), prediction (Xu et al., 2024), and support (Lai et al., 2023; Fu et al., 2023).

To enhance their effectiveness, methods like few-shot prompting, which presents the LLMs with a small number of task demonstrations before requiring them to perform a task, and chain-of-thought (CoT) prompting (Wei et al., 2022), which prompts the model to generate intermediate steps or paths of reasoning when dealing with problems, are employed. Drawing inspiration from the success of CoT in natural language processing (NLP), novel approaches such as chain-of-empathy prompting (Lee et al., 2023), which incorporates insights from psychotherapy (i.e., the therapists' reasoning process) to prompt the LLMs to generate the cognitive reasoning of human emotion, and diagnosis-of-thought prompting (Chen et al., 2023), which prompts the LLMs to make a decision using three diagnosis stages (i.e., subjectivity assessment, contrastive reasoning, and schema analysis), have been proposed to further improve LLM performance in the mental health domain.

Another stream of research focuses on the further training or fine-tuning of general LLMs using mental health-specific texts. This approach aims to inject mental health knowledge into existing base LLMs, leading to more relevant and accurate analyses and support. Notable examples include MentaLLaMA (Yang et al., 2023b) and Mental-LLM (Xu et al., 2024), which fine-tuned the LLaMA-2 model and the Alpaca (Taori et al., 2023)/FLAN-T5 (Chung et al., 2024) model, respectively, using social media data for enhanced mental health predictions. Similarly, ChatCounselor (Liu et al., 2023) leveraged the Psych8k dataset, comprising real interactions between clients and psychologists, to fine-tune the LLaMA model. Additionally, ExTES-LLaMA (Zheng et al., 2023) employed the LLaMA model, fine-tuned on emotional support dialogues.

By leveraging the powerful language understanding and generation capabilities of LLMs, combined with techniques like few-shot prompting, chain-of-thought prompting, and fine-tuning on mental health-specific data, researchers are exploring ways to enhance the performance of these models in mental health applications, ultimately aiming to provide more accurate and effective analysis, prediction, and support for individuals grappling with mental health challenges.

#### 4.4.1 Fine-Tuning Techniques

All the studies that involved fine-tuning large language models (LLMs) for mental health applications adopted the instruction fine-tuning (IFT) technique. IFT is a type of fine-tuning where a model is trained for an instruction-following task that involves instructing the model to perform another task. In contrast, classical types of supervised fine-tuning do not involve providing instructions to the model; instead, the model is directly fine-tuned to perform a single downstream task.

IFT allows domain knowledge to be injected into LLMs while improving the model's capability to follow human instructions accurately. For example, in the case of the ChatCounselor (Liu et al., 2023) study, which utilized conversations between clients and psychologists, the researchers used prompts such as: "*Your task is to identify the patient and counselor in the conversation. Summarize the conversation into only one round of conversation, one query or description by the patient, and one feedback by the counselor*" to prompt GPT-4 (Achiam et al., 2023) to generate instruction-input-output triples. In these triples, the instruction could be "*If you are a counselor, please answer the questions based on the description of the patient.*" The input would be the query or question from the patient, and the output would be the feedback or answer generated by the model, acting as a counselor.

By leveraging IFT, researchers can fine-tune LLMs on mental health-specific data while providing explicit instructions on how to process and respond to the input data. This approach not only infuses the LLMs with domain-specific knowledge but also enhances their ability to follow human instructions accurately, a critical requirement for mental health applications where precise understanding and appropriate responses are crucial.

## 5 Summary

The increasing prevalence of mental health disorders, including depression, anxiety, and bipolar disorder, highlights the urgent need for effective detection and intervention methods. The COVID-19 pandemic has further exacerbated these challenges, emphasizing the global shortage of trained mental health professionals. This survey provides a comprehensive examination of NLP techniques used in detecting mental health disorders through text analysis. Spanning a decade of research, it focuses on the evolution of methodologies from basic machine learning models to sophisticated neural networks. Notable trends include the rise of transformer-based models like BERT and GPT, which have demonstrated superior performance in understanding and generating human-like text.

The survey addresses several critical questions:

- **Approaches to Mental Illness Detection:** Various methods are explored, ranging from simple keyword spotting to complex deep learning models.

- **Features Used in Models:** Traditional features such as word frequencies and syntactic structures, along with modern embeddings and contextual features, are discussed.

- **Common Neural Architectures:** The survey highlights the transition from recurrent neural networks (RNNs) to transformers, which provide better context understanding and scalability.

- **Future Challenges:** Issues such as data privacy, the need for diverse datasets, and the development of interpretable models are identified as key areas for future research.

By answering these questions, the survey offers a detailed overview of current NLP applications in mental health and suggests potential future research directions.

## 6 Conclusions and Future Work

NLP technologies hold significant promise for enhancing the early detection and intervention of mental health disorders. Despite considerable advancements, several challenges persist. These include ensuring data privacy, creating more diverse and representative datasets, and developing models that can generalize across different populations and languages. Future research should focus on addressing these challenges, improving the accuracy and interpretability of NLP systems, and integrating them effectively into clinical workflows. By advancing these technologies, we can better support mental health professionals and improve outcomes for individuals affected by mental health disorders. The ongoing refinement of NLP methodologies has the potential to profoundly impact the field of mental health positively, offering new tools for diagnosis and treatment.

## References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Marios Adamou, Grigoris Antoniou, Elissavet Greasidou, Vincenzo Lagani, Paulos Charonyktakis, and Ioannis Tsamardinos. 2018. Mining free-text medical notes for suicide risk assessment. In *Proceedings of the 10th hellenic conference on artificial intelligence*, pages 1–8.

Juan Aguilera, Delia Irazú Hernández Farías, Rosa María Ortega-Mendoza, and Manuel Montes-y Gómez. 2021. Depression and anorexia detection in social media as a one-class classification problem. *Applied Intelligence*, 51:6088–6103.

Usman Ahmed, Suresh Kumar Mukhiya, Gautam Srivastava, Yngve Lamo, and Jerry Chun-Wei Lin. 2021. Attention-based deep entropy active learning using lexical algorithm for mental health treatment. *Frontiers in Psychology*, 12:642347.

Ahmet Emre Aladağ, Serra Muderrisoglu, Naz Berfu Akbas, Oguzhan Zahmacioglu, and Haluk O Bingol. 2018. Detecting suicidal ideation on forums: proof-of-concept study. *Journal of medical Internet research*, 20(6):e9840.

Mario Aragon, Adrian Pastor Lopez Monroy, Luis Gonzalez, David E. Losada, and Manuel Montes. 2023. DisorBERT: A double domain adaptation model for detecting signs of mental disorders in social media. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15305–15318, Toronto, Canada. Association for Computational Linguistics.

Mario Ezra Aragón, A Pastor López-Monroy, Luis C González, and Manuel Montes-y Gómez. 2020. Attention to emotions: detecting mental disorders in social media. In *Text, Speech, and Dialogue: 23rd International Conference, TSD 2020, Brno, Czech Republic, September 8–11, 2020, Proceedings 23*, pages 231–239. Springer.

Mario Ezra Aragon, Adrian Pastor Lopez-Monroy, Luis Carlos González-Gurrola, and Manuel Montes-y Gómez. 2021. Detecting mental disorders in social media through emotional patterns-the case of anorexia and depression. *IEEE transactions on affective computing*, 14(1):211–222.

Ayan Bandyopadhyay, Linda Achilles, Thomas Mandl, Mandar Mitra, and Sanjoy Kr Saha. 2019. Identification of depression strength for users of online platforms: a comparison of text retrieval approaches. In *Proc. CEUR Workshop Proceedings*, volume 2454, pages 331–342.

Eliseo Bao, Anxo Pérez, and Javier Parapar. 2023. Explainable depression symptom detection in social media. *arXiv e-prints*, pages arXiv–2310.

Marouane Birjali, Abderrahim Beni-Hssane, and Mohammed Erritali. 2017. Machine learning and semantic sentiment analysis based algorithms for suicide sentiment prediction in social networks. *Procedia Computer Science*, 113:65–72.

William Boag, Olga Kovaleva, Thomas H McCoy Jr, Anna Rumshisky, Peter Szolovits, and Roy H Perlis. 2021. Hard for humans, hard for machines: predicting readmission after psychiatric hospitalization using narrative notes. *Translational psychiatry*, 11(1):32.

Samir Boukil, Fatiha El Adnani, Loubna Cherrat, Abd Elmajid El Moutaouakkil, and Mostafa Ezziyyani. 2019. Deep learning algorithm for suicide sentiment prediction. In *Advanced Intelligent Systems for Sustainable Development (AI2SD'2018) Vol 4: Advanced Intelligent Systems Applied to Health*, pages 261–272. Springer.

Antoine Briand, Hayda Almeida, and Marie-Jean Meurs. 2018. Analysis of social media posts for early detection of mental health conditions. In *Advances in Artificial Intelligence: 31st Canadian Conference on Artificial Intelligence, Canadian AI 2018, Toronto, ON, Canada, May 8–11, 2018, Proceedings 31*, pages 133–143. Springer.

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

Ming-Yi Chang and Chih-Ying Tseng. 2020. Detecting social anxiety with online social network data. In *2020 21st IEEE International Conference on Mobile Data Management (MDM)*, pages 333–336. IEEE.

Aditi Chaurasia, Suhani Vinod Prajapati, Priya A Tiru, Shobhan Kumar, Riya Gupta, and Arun Chauhan. 2021. Predicting mental health of scholars using contextual word embedding. In *2021 8th International Conference on Computing for Sustainable Global Development (INDIACom)*, pages 923–930. IEEE.

Lushi Chen, Walid Magdy, Heather Whalley, and Maria Klara Wolters. 2020. Examining the role of mood patterns in predicting self-reported depressive symptoms. In *Proceedings of the 12th ACM Conference on Web Science*, pages 164–173.

Zhiyu Chen, Yujie Lu, and William Yang Wang. 2023. Empowering psychotherapy with large language models: Cognitive distortion detection through diagnosis of thought prompting. *arXiv preprint arXiv:2310.07146*.

Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*.

Taridzo Chomutare. 2014. Text classification to automatically identify online patients vulnerable to depression. In *Pervasive Computing Paradigms for Mental Health: 4th International Symposium, MindCare 2014, Tokyo, Japan, May 8-9, 2014, Revised Selected Papers 4*, pages 125–130. Springer.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Arman Cohan, Bart Desmet, Andrew Yates, Luca Soldaini, Sean MacAvaney, and Nazli Goharian. 2018. 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 (COLING)*, pages 1485—-1497. 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.

Glen Coppersmith, Mark Dredze, Craig Harman, Kristy Hollingshead, and Margaret Mitchell. 2015. 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.

Bo Dao, Thin Nguyen, Dinh Phung, and Svetha Venkatesh. 2014. Effect of mood, social connectivity and age in online depression community via topic and linguistic analysis. In *Web Information Systems Engineering–WISE 2014: 15th International Conference, Thessaloniki, Greece, October 12-14, 2014, Proceedings, Part I 15*, pages 398–407. Springer.

Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting depression via social media. In *Proceedings of the international AAAI conference on web and social media*, volume 7, pages 128–137.

Fionn Delahunty, Ian D Wood, and Mihael Arcan. 2018. First insights on a passive major depressive disorder prediction system with incorporated conversational chatbot. In *Irish Conference on Artificial Intelligence and Cognitive Science*. AICS 2018 and CEUR-WS. org.

David Delgado-Gomez, Hilario Blasco-Fontecilla, Federico Sukno, Maria Socorro Ramos-Plasencia, and Enrique Baca-Garcia. 2012. Suicide attempters classification: Toward predictive models of suicidal behavior. *Neurocomputing*, 92:3–8.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. Goemotions: A dataset of fine-grained emotions. *arXiv preprint arXiv:2005.00547*.

Saahil Deshpande and Jim Warren. 2021. Self-harm detection for mental health chatbots. In *MIE*, pages 48–52.

Heinrich Dinkel, Mengyue Wu, and Kai Yu. 2019. Text-based depression detection on sparse data. *arXiv preprint arXiv:1904.05154*.

Son Doan, Elly W Yang, Sameer S Tilak, Peter W Li, Daniel S Zisook, and Manabu Torii. 2019. Extracting health-related causality from twitter messages using natural language processing. *BMC medical informatics and decision making*, 19:71–77.

Johnny Downs, Sumithra Velupillai, Gkotsis George, Rachel Holden, Maxim Kikoler, Harry Dean, Andrea Fernandes, and Rina Dutta. 2017. Detection of suicidality in adolescents with autism spectrum disorders: developing a natural language processing approach for use in electronic health records. In *AMIA annual symposium proceedings*, volume 2017, page 641. American Medical Informatics Association.

Samah Fodeh, Taihua Li, Kevin Menczynski, Tedd Burgette, Andrew Harris, Georgeta Ilita, Satyan Rao, Jonathan Gemmell, and Daniela Raicu. 2019. Using machine learning algorithms to detect suicide risk factors on twitter. In *2019 international conference on data mining workshops (ICDMW)*, pages 941–948. IEEE.

Peter J Franz, Erik C Nook, Patrick Mair, and Matthew K Nock. 2020. Using topic modeling to detect and describe self-injurious and related content on a large-scale digital platform. *Suicide and Life-Threatening Behavior*, 50(1):5–18.

Guanghui Fu, Qing Zhao, Jianqiang Li, Dan Luo, Changwei Song, Wei Zhai, Shuo Liu, Fan Wang, Yan Wang, Lijuan Cheng, et al. 2023. Enhancing psychological counseling with large language model: A multifaceted decision-support system for non-professionals. *arXiv preprint arXiv:2308.15192*.

Muskan Garg, Chandni Saxena, Veena Krishnan, Ruchi Joshi, Sriparna Saha, Vijay Mago, and Bonnie J Dorr. 2022. Cams: An annotated corpus for causal analysis of mental health issues in social media posts. *arXiv preprint arXiv:2207.04674*.

Manas Gaur, Amanuel Alambo, Joy Prakash Sain, Ugur Kursuncu, Krishnaprasad Thirunarayan, Ramakanth Kavuluru, Amit Sheth, Randy Welton, and Jyotishman Pathak. 2019. Knowledge-aware assessment of severity of suicide risk for early intervention. In *The world wide web conference*, pages 514–525.

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.

Shreya Ghosh and Tarique Anwar. 2021. Depression intensity estimation via social media: a deep learning approach. *IEEE Transactions on Computational Social Systems*, 8(6):1465–1474.

Soumitra Ghosh, Asif Ekbal, and Pushpak Bhattacharyya. 2022. A multitask framework to detect depression, sentiment and multi-label emotion from suicide notes. *Cognitive Computation*, 14(1):110–129.

George Gkotsis, Anika Oellrich, Tim Hubbard, Richard Dobson, Maria Liakata, Sumithra Velupillai, and Rina Dutta. 2016. The language of mental health problems in social media. In *Proceedings of the third workshop on computational linguistics and clinical psychology*, pages 63–73.

Tao Gui, Qi Zhang, Liang Zhu, Xu Zhou, Minlong Peng, and Xuanjing Huang. 2019a. Depression detection on social media with reinforcement learning. In *Chinese Computational Linguistics: 18th China National Conference, CCL 2019, Kunming, China, October 18–20, 2019, Proceedings 18*, pages 613–624. Springer.

Tao Gui, Liang Zhu, Qi Zhang, Minlong Peng, Xu Zhou, Keyu Ding, and Zhigang Chen. 2019b. Cooperative multimodal approach to depression detection in twitter. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 110–117.

Sharath Chandra Guntuku, Anneke Buffone, Kokil Jaidka, Johannes C Eichstaedt, and Lyle H Ungar. 2019. Understanding and measuring psychological stress using social media. In *Proceedings of the international AAAI conference on web and social media*, volume 13, pages 214–225.

Sharath Chandra Guntuku, Salvatore Giorgi, and Lyle Ungar. 2018. Current and future psychological health prediction using language and socio-demographics of children for the clpysch 2018 shared task. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 98–106.

Xiaobo Guo, Yaojia Sun, and Soroush Vosoughi. 2021. Emotion-based modeling of mental disorders on social media. In *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, pages 8–16.

Ayaan Haque, Viraaj Reddi, and Tyler Giallanza. 2021. Deep learning for suicide and depression identification with unsupervised label correction.

Farsheed Haque, Ragib Un Nur, Shaeekh Al Jahan, Zarar Mahmud, and Faisal Muhammad Shah. 2020. A transformer based approach to detect suicidal ideation using pre-trained language models. In *2020 23rd international conference on computer and information technology (ICCIT)*, pages 1–5. IEEE.

Mohamad Farid Mohd Hayati, Mohd Adli Md Ali, and Ahmad Nabil Md Rosli. 2022. Depression detection on malay dialects using gpt-3. In *2022 IEEE-EMBS Conference on Biomedical Engineering and Sciences (IECBES)*, pages 360–364. IEEE.

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

Elizabeth Hilvert, Denise Davidson, and Perla B Gámez. 2020. Assessment of personal narrative writing in children with and without autism spectrum disorder. *Research in Autism Spectrum Disorders*, 69:101453.

Pengwei Hu, Chenhao Lin, Hui Su, Shaochun Li, Xue Han, Yuan Zhang, and Jing Mei. 2021. Bluememo: depression analysis through twitter posts. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 5252–5254.

Yan Huang, Xiaoqian Liu, and Tingshao Zhu. 2019. Suicidal ideation detection via social media analytics. In *Human Centered Computing: 5th International Conference, HCC 2019, Čačak, Serbia, August 5–7, 2019, Revised Selected Papers 5*, pages 166–174. Springer.

Youjin Hwang, Hyung Jun Kim, Hyung Jin Choi, and Joonhwan Lee. 2020. Exploring abnormal behavior patterns of online users with emotional eating behavior: topic modeling study. *Journal of Medical Internet Research*, 22(3):e15700.

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

Julia Ive, Natalia Viani, Joyce Kam, Lucia Yin, Somain Verma, Stephen Puntis, Rudolf N Cardinal, Angus Roberts, Robert Stewart, and Sumithra Velupillai. 2020. Generation and evaluation of artificial mental health records for natural language processing. *NPJ digital medicine*, 3(1):69.

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

Zunaira Jamil. 2017. *Monitoring tweets for depression to detect at-risk users*. Ph.D. thesis, Université d'Ottawa/University of Ottawa.

Shaoxiong Ji, Shirui Pan, Xue Li, Erik Cambria, Guodong Long, and Zi Huang. 2020. Suicidal ideation detection: A review of machine learning methods and applications. *arXiv preprint arXiv:1910.12611*.

Shaoxiong Ji, Celina Ping Yu, Sai-fu Fung, Shirui Pan, and Guodong Long. 2018. Supervised learning for suicidal ideation detection in online user content. *Complexity*, 2018.

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.

Kantinee Katchapakirin, Konlakorn Wongpatikaseree, Panida Yomaboot, and Yongyos Kaewpitakkun. 2018. Facebook social media for depression detection in the thai community. In *2018 15th international joint conference on computer science and software engineering (jcsse)*, pages 1–6. IEEE.

Abel N Kho, Luke V Rasmussen, John J Connolly, Peggy L Peissig, Justin Starren, Hakon Hakonarson, and M Geoffrey Hayes. 2013. Practical challenges in integrating genomic data into the electronic health record. *Genetics in Medicine*, 15(10):772–778.

Byeongchang Kim, Hyunwoo Kim, and Gunhee Kim. 2019. Abstractive summarization of Reddit posts with multi-level memory networks. 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 2519–2531, Minneapolis, Minnesota. Association for Computational Linguistics.

Bhavani Singh Agnikula Kshatriya, Nicolas A Nunez, Manuel Gardea Resendez, Euijung Ryu, Brandon J Coombes, Sunyang Fu, Mark A Frye, Joanna M Biernacka, and Yanshan Wang. 2021. Neural language models with distant supervision to identify major depressive disorder from clinical notes. *arXiv preprint arXiv:2104.09644*.

Tin Lai, Yukun Shi, Zicong Du, Jiajie Wu, Ken Fu, Yichao Dou, and Ziqi Wang. 2023. Psy-llm: Scaling up global mental health psychological services with ai-based large language models. *arXiv preprint arXiv:2307.11991*.

Juan S Lara, Mario Ezra Aragón, Fabio A González, and Manuel Montes-y Gómez. 2021. Deep bag-of-sub-emotions for depression detection in social media. In *Text, Speech, and Dialogue: 24th International Conference, TSD 2021, Olomouc, Czech Republic, September 6–9, 2021, Proceedings 24*, pages 60–72. Springer.

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

Victor Leiva and Ana Freire. 2017. Towards suicide prevention: early detection of depression on social media. In *Internet Science: 4th International Conference, INSCI 2017, Thessaloniki, Greece, November 22-24, 2017, Proceedings 4*, pages 428–436. Springer.

Maxwell Levis, Christine Leonard Westgate, Jiang Gui, Bradley V Watts, and Brian Shiner. 2021. Natural language processing of clinical mental health notes may add predictive value to existing suicide risk models. *Psychological medicine*, 51(8):1382–1391.

Han Li, Renwen Zhang, Yi-Chieh Lee, Robert E Kraut, and David C Mohr. 2023. Systematic review and meta-analysis of ai-based conversational agents for promoting mental health and well-being. *NPJ Digital Medicine*, 6(1):236.

Ningyun Li, Huijun Zhang, and Ling Feng. 2021. Incorporating forthcoming events and personality traits in social media based stress prediction. *IEEE Transactions on Affective Computing*, 14(1):603–621.

Qing Liao, Wei Wang, Yi Han, and Qian Zhang. 2013. Analyzing the influential people in sina weibo dataset. In *2013 IEEE Global Communications Conference (GLOBECOM)*, pages 3066–3071. IEEE.

Huijie Lin, Jia Jia, Jiezhong Qiu, Yongfeng Zhang, Guangyao Shen, Lexing Xie, Jie Tang, Ling Feng, and Tat-Seng Chua. 2017a. Detecting stress based on social interactions in social networks. *IEEE Transactions on Knowledge and Data Engineering*, 29(9):1820–1833.

Wutao Lin, Donghong Ji, and Yanan Lu. 2017b. Disorder recognition in clinical texts using multi-label structured svm. *BMC bioinformatics*, 18:1–11.

June M Liu, Donghao Li, He Cao, Tianhe Ren, Zeyi Liao, and Jiamin Wu. 2023. Chatcounselor: A large language models for mental health support. *arXiv preprint arXiv:2309.15461*.

David E Losada, Fabio Crestani, and Javier Parapar. 2018. Overview of erisk: early risk prediction on the internet. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 9th International Conference of the CLEF Association, CLEF 2018, Avignon, France, September 10-14, 2018, Proceedings 9*, pages 343–361. Springer.

David E Losada, Fabio Crestani, and Javier Parapar. 2019. Overview of erisk 2019 early risk prediction on the internet. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 10th International Conference of the CLEF Association, CLEF 2019, Lugano, Switzerland, September 9–12, 2019, Proceedings 10*, pages 340–357. Springer.

Meizhen Lv, Ang Li, Tianli Liu, and Tingshao Zhu. 2015. Creating a chinese suicide dictionary for identifying suicide risk on social media. *PeerJ*, 3:e1455.

Zilin Ma, Yiyang Mei, and Zhaoyuan Su. 2023. Understanding the benefits and challenges of using large language model-based conversational agents for mental well-being support. In *AMIA Annual Symposium Proceedings*, volume 2023, page 1105. American Medical Informatics Association.

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.

Keshu Malviya, Bholanath Roy, and SK Saritha. 2021. A transformers approach to detect depression in social media. In *2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS)*, pages 718–723. IEEE.

John J McGrath, Ali Al-Hamzawi, Jordi Alonso, Yasmin Altwaijri, Laura H Andrade, Evelyn J Bromet, Ronny Bruffaerts, José Miguel Caldas de Almeida, Stephanie Chardoul, Wai Tat Chiu, Louisa Degenhardt, Olga V Demler, Finola Ferry, Oye Gureje, Josep Maria Haro, Elie G Karam, Georges Karam, Salma M Khaled, Viviane Kovess-Masfety, Marta Magno, Maria Elena Medina-Mora, Jacek Moskalewicz, Fernando Navarro-Mateu, Daisuke Nishi, Oleguer Plana-Ripoll, José Posada-Villa, Charlene Rapsey, Nancy A Sampson, Juan Carlos Stagnaro, Dan J Stein, Margreet ten Have, Yolanda Torres, Cristian Vladescu, Peter W Woodruff, Zahari Zarkov, Ronald C Kessler, Sergio Aguilar-Gaxiola, Ali Al-Hamzawi, Jordi Alonso, Yasmin A. Altwaijri, Laura Helena Andrade, Lukoye Atwoli, Corina Benjet, Evelyn J. Bromet, Ronny Bruffaerts, Brendan Bunting, José Miguel Caldas de Almeida, Graça Cardoso, Stephanie Chardoul, Alfredo H. Cía, Louisa Degenhardt, Giovanni De Girolamo, Oye Gureje, Josep Maria Haro, Meredith G. Harris, Hristo Hinkov, Chi yi Hu, Peter De Jonge, Aimee N. Karam, Elie G. Karam, Georges Karam, Alan E. Kazdin, Norito Kawakami, Ronald C. Kessler, Andrzej Kiejna, Viviane Kovess-Masfety, John J. McGrath, Maria Elena Medina-Mora, Jacek Moskalewicz, Fernando Navarro-Mateu, Daisuke Nishi, Marina Piazza, José Posada-Villa, Kate M. Scott, Juan Carlos Stagnaro, Dan J. Stein, Margreet Ten Have, Yolanda Torres, Maria Carmen Viana, Daniel V. Vigo, Cristian Vladescu, David R. Williams, Peter Woodruff, Bogdan Wojtyniak, Miguel Xavier, and Alan M. Zaslavsky. 2023. Age of onset and cumulative risk of mental disorders: a cross-national analysis of population surveys from 29 countries. *The Lancet Psychiatry*, 10(9):668–681.

Nir Menachemi and Taleah H Collum. 2011. Benefits and drawbacks of electronic health record systems. *Risk management and healthcare policy*, pages 47–55.

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.

Sankha S Mukherjee, Jiawei Yu, Yida Won, Mary J McClay, Lu Wang, A John Rush, and Joydeep Sarkar. 2020. Natural language processing-based quantification of the mental state of psychiatric patients. *Computational Psychiatry*, 4.

Ankit Murarka, Balaji Radhakrishnan, and Sushma Ravichandran. 2020. Detection and classification of mental illnesses on social media using roberta. *arXiv preprint arXiv:2011.11226*.

Nona Naderi, Julien Gobeill, Douglas Teodoro, Emilie Pasche, and Patrick Ruch. 2019. A baseline approach for early detection of signs of anorexia and self-harm in reddit posts. In *Proceedings of CLEF (Conference and Labs of the Evaluation Forum) 2019 Working Notes*. 9-12 September 2019.

Prakash M Nadkarni, Lucila Ohno-Machado, and Wendy W Chapman. 2011. Natural language processing: an introduction. *Journal of the American Medical Informatics Association*, 18(5):544–551.

Usman Naseem, Jinman Kim, Matloob Khushi, and Adam G Dunn. 2022. Identification of disease or symptom terms in reddit to improve health mention classification. In *Proceedings of the ACM Web Conference 2022*, pages 2573–2581.

Jihad S Obeid, Erin R Weeda, Andrew J Matuskowitz, Kevin Gagnon, Tami Crawford, Christine M Carr, and Lewis J Frey. 2019. Automated detection of altered mental status in emergency department clinical notes: a deep learning approach. *BMC medical informatics and decision making*, 19:1–9.

Ahmed Husseini Orabi, Prasadith Buddhitha, Mahmoud Husseini Orabi, and Diana Inkpen. 2018. Deep learning for depression detection of twitter users. In *Proceedings of the fifth workshop on computational linguistics and clinical psychology: from keyboard to clinic*, pages 88–97.

Huyen Trang Phan, Van Cuong Tran, Ngoc Thanh Nguyen, and Dosam Hwang. 2020. A framework for detecting user's psychological tendencies on twitter based on tweets sentiment analysis. In *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, pages 357–372. Springer.

Waleed Ragheb, Jérôme Azé, Sandra Bringay, and Maximilien Servajean. 2019. Attentive multi-stage learning for early risk detection of signs of anorexia and self-harm on social media. In *CLEF 2019-Conference and Labs of the Evaluation Forum*, volume 2380.

Guozheng Rao, Yue Zhang, Li Zhang, Qing Cong, and Zhiyong Feng. 2020. Mgl-cnn: a hierarchical posts representations model for identifying depressed individuals in online forums. *IEEE Access*, 8:32395–32403.

Andrew G Reece, Andrew J Reagan, Katharina LM Lix, Peter Sheridan Dodds, Christopher M Danforth, and Ellen J Langer. 2017. Forecasting the onset and course of mental illness with twitter data. *Scientific reports*, 7(1):13006.

Jürgen Rehm and Kevin D Shield. 2019. Global burden of disease and the impact of mental and addictive disorders. *Current psychiatry reports*, 21:1–7.

Fabien Ringeval, Björn Schuller, Michel Valstar, Jonathan Gratch, Roddy Cowie, Stefan Scherer, Sharon Mozgai, Nicholas Cummins, Maximilian

Schmitt, and Maja Pantic. 2017. Avec 2017: Real-life depression, and affect recognition workshop and challenge. In *Proceedings of the 7th annual workshop on audio/visual emotion challenge*, pages 3–9.

Tomasz Rutowski, Elizabeth Shriberg, Amir Harati, Yang Lu, Piotr Chlebek, and Ricardo Oliveira. 2020. Depression and anxiety prediction using deep language models and transfer learning. In *2020 7th International Conference on Behavioural and Social Computing (BESC)*, pages 1–6. IEEE.

Tomek Rutowski, Elizabeth Shriberg, Amir Harati, Yang Lu, Ricardo Oliveira, and Piotr Chlebek. 2021. Cross-demographic portability of deep nlp-based depression models. In *2021 IEEE spoken language technology workshop (SLT)*, pages 1052–1057. IEEE.

Kayalvizhi S, Thenmozhi Durairaj, Bharathi Raja Chakravarthi, and Jerin Mahibha C. 2022. Findings of the shared task on detecting signs of depression from social media. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 331–338, Dublin, Ireland. Association for Computational Linguistics.

Benjamin J Sadock et al. 2015. *Kaplan & Sadock's synopsis of psychiatry: behavioral sciences/clinical psychiatry*, volume 2015. Wolters Kluwer Philadelphia.

Shirin Saleem, Maciej Pacula, Rachel Chasin, Rohit Kumar, Rohit Prasad, Michael Crystal, Brian Marx, Denise Sloan, Jennifer Vasterling, and Theodore Speroff. 2012. Automatic detection of psychological distress indicators in online forum posts. In *Proceedings of the 2012 Asia Pacific signal and information processing association annual summit and conference*, pages 1–4. IEEE.

Damian F Santomauro, Ana M Mantilla Herrera, Jamileh Shadid, Peng Zheng, Charlie Ashbaugh, David M Pigott, Cristiana Abbafati, Christopher Adolph, Joanne O Amlag, Aleksandr Y Aravkin, et al. 2021. Global prevalence and burden of depressive and anxiety disorders in 204 countries and territories in 2020 due to the covid-19 pandemic. *The Lancet*, 398(10312):1700–1712.

Ramit Sawhney, Harshit Joshi, Lucie Flek, and Rajiv Shah. 2021a. Phase: Learning emotional phase-aware representations for suicide ideation detection on social media. In *Proceedings of the 16th conference of the European Chapter of the Association for Computational Linguistics: main volume*, pages 2415–2428.

Ramit Sawhney, Harshit Joshi, Saumya Gandhi, and Rajiv Shah. 2020. A time-aware transformer based model for suicide ideation detection on social media. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 7685–7697.

Ramit Sawhney, Harshit Joshi, Saumya Gandhi, and Rajiv Ratn Shah. 2021b. Towards ordinal suicide ideation detection on social media. In *Proceedings of the 14th ACM international conference on web search and data mining*, pages 22–30.

Ramit Sawhney, Harshit Joshi, Rajiv Shah, and Lucie Flek. 2021c. 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.

Ivan Sekulić and Michael Strube. 2020. Adapting deep learning methods for mental health prediction on social media. *arXiv preprint arXiv:2003.07634*.

Guangyao Shen, Jia Jia, Liqiang Nie, Fuli Feng, Cunjun Zhang, Tianrui Hu, Tat-Seng Chua, and Wenwu Zhu. 2017. Depression detection via harvesting social media: A multimodal dictionary learning solution. In *IJCAI*, pages 3838–3844.

Benjamin Shickel, Scott Siegel, Martin Heesacker, Sherry Benton, and Parisa Rashidi. 2020. Automatic detection and classification of cognitive distortions in mental health text. In *2020 IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE)*, pages 275–280. IEEE.

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.

Andrew Sims. 1988. *Symptoms in the mind: An introduction to descriptive psychopathology.* Bailliere Tindall Publishers.

Pradyumna Prakhar Sinha, Rohan Mishra, Ramit Sawhney, Debanjan Mahata, Rajiv Ratn Shah, and Huan Liu. 2019. # suicidal-a multipronged approach to identify and explore suicidal ideation in twitter. In *Proceedings of the 28th ACM international conference on information and knowledge management*, pages 941–950.

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

Hanadi Solieman and Evgenii A Pustozerov. 2021. The detection of depression using multimodal models based on text and voice quality features. In *2021 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (ElConRus)*, pages 1843–1848. IEEE.

Maxim Stankevich, Ivan Smirnov, Natalia Kiselnikova, and Anastasia Ushakova. 2020. Depression detection from social media profiles. In *Data Analytics and Management in Data Intensive Domains: 21st International Conference, DAMDID/RCDL 2019, Kazan,* *Russia, October 15–18, 2019, Revised Selected Papers 21*, pages 181–194. Springer.

Jini Jojo Stephen and P Prabu. 2019. Detecting the magnitude of depression in twitter users using sentiment analysis. *International Journal of Electrical and Computer Engineering*, 9(4):3247.

Yue Su, Huijia Zheng, Xiaoqian Liu, and Tingshao Zhu. 2018. Depressive emotion recognition based on behavioral data. In *International Conference on Human Centered Computing*, pages 257–268. Springer.

Michael Mesfin Tadesse, Hongfei Lin, Bo Xu, and Liang Yang. 2019. Detection of suicide ideation in social media forums using deep learning. *Algorithms*, 13(1):7.

Srikanth Tadisetty and Kambiz Ghazinour. 2021. Anonymous prediction of mental illness in social media. In *2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC)*, pages 0954–0960. IEEE.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Stanford alpaca: An instruction-following llama model.

Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54.

ML Tlachac, Ermal Toto, and Elke Rundensteiner. 2019. You're making me depressed: Leveraging texts from contact subsets to predict depression. In *2019 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*, pages 1–4. IEEE.

Lei Tong, Zhihua Liu, Zheheng Jiang, Feixiang Zhou, Long Chen, Jialin Lyu, Xiangrong Zhang, Qianni Zhang, Abdul Sadka, Yinhai Wang, et al. 2022. Cost-sensitive boosting pruning trees for depression detection on twitter. *IEEE transactions on affective computing*.

Tung Tran and Ramakanth Kavuluru. 2017. Predicting mental conditions based on "history of present illness" in psychiatric notes with deep neural networks. *Journal of biomedical informatics*, 75:S138–S148.

Ilias Triantafyllopoulos, Georgios Paraskevopoulos, and Alexandros Potamianos. 2023. Depression detection in social media posts using affective and social norm features. *arXiv preprint arXiv:2303.14279*.

Alina Trifan, Rui Antunes, Sérgio Matos, and Jose Luís Oliveira. 2020. Understanding depression from psycholinguistic patterns in social media texts. In *European Conference on Information Retrieval*, pages 402–409. Springer.

Alina Trifan and José Luís Oliveira. 2019. Bioinfo@ uavr at erisk 2019: delving into social media texts for the early detection of mental and food disorders. In *CLEF (working notes)*.

Alina Trifan and José Luis Oliveira. 2021. Cross-evaluation of social mining for classification of depressed online personas. *Journal of Integrative Bioinformatics*, 18(2):101–110.

Marcel Trotzek, Sven Koitka, and Christoph M Friedrich. 2018. Utilizing neural networks and linguistic metadata for early detection of depression indications in text sequences. *IEEE Transactions on Knowledge and Data Engineering*, 32(3):588–601.

Elsbeth Turcan and Kathleen McKeown. 2019. Dreaddit: A reddit dataset for stress analysis in social media. *arXiv preprint arXiv:1911.00133*.

Ana-Sabina Uban, Berta Chulvi, and Paolo Rosso. 2021. An emotion and cognitive based analysis of mental health disorders from social media data. *Future Generation Computer Systems*, 124:480–494.

Nemanja Vaci, Qiang Liu, Andrey Kormilitzin, Franco De Crescenzo, Ayse Kurtulmus, Jade Harvey, Bessie O'Dell, Simeon Innocent, Anneka Tomlinson, Andrea Cipriani, et al. 2020. Natural language processing for structuring clinical text data on depression using uk-cris. *BMJ Ment Health*, 23(1):21–26.

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*, pages 3–10.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017a. Attention is all you need. *Advances in neural information processing systems*, 30.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017b. Attention Is All You Need. *arXiv:1706.03762 [cs]*. ArXiv: 1706.03762.

Rohit Voleti, Stephanie Woolridge, Julie M Liss, Melissa Milanovic, Christopher R Bowie, and Visar Berisha. 2019. Objective assessment of social skills using automated language analysis for identification of schizophrenia and bipolar disorder. *arXiv preprint arXiv:1904.10622*.

Michael von Glischinski, Tobias Teismann, S Prinz, Jochen E Gebauer, and Gerrit Hirschfeld. 2016. Depressive symptom inventory suicidality subscale: Optimal cut points for clinical and non-clinical samples. *Clinical psychology & psychotherapy*, 23(6):543–549.

Laura M Vowels, Rachel Francois-Walcott, and Joëlle Darwiche. 2023. Ai in relationship counselling: Evaluating chatgpt's therapeutic efficacy in providing relationship advice.

Ning Wang, Fan Luo, Yuvraj Shivtare, Varsha D Badal, KP Subbalakshmi, Rajarathnam Chandramouli, and Ellen Lee. 2021. Learning models for suicide prediction from social media posts. *arXiv preprint arXiv:2105.03315*.

Xiaofeng Wang, Shuai Chen, Tao Li, Wanting Li, Yejie Zhou, Jie Zheng, Qingcai Chen, Jun Yan, Buzhou Tang, et al. 2020a. Depression risk prediction for chinese microblogs via deep-learning methods: Content analysis. *JMIR medical informatics*, 8(7):e17958.

Yiding Wang, Zhenyi Wang, Chenghao Li, Yilin Zhang, and Haizhou Wang. 2020b. A multitask deep learning approach for user depression detection on sina weibo. *arXiv preprint arXiv:2008.11708*.

Yu-Tseng Wang, Hen-Hsen Huang, Hsin-Hsi Chen, and H Chen. 2018. A neural network approach to early risk detection of depression and anorexia on social media text. In *CLEF (Working Notes)*, pages 1–8.

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

Akkapon Wongkoblap, Miguel A Vadillo, and Vasa Curcin. 2018. A multilevel predictive model for detecting social network users with depression. In *2018 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 130–135. IEEE.

Akkapon Wongkoblap, Miguel A Vadillo, and Vasa Curcin. 2019a. Modeling depression symptoms from social network data through multiple instance learning. *AMIA Summits on Translational Science Proceedings*, 2019:44.

Akkapon Wongkoblap, Miguel A Vadillo, and Vasa Curcin. 2019b. Predicting social network users with depression from simulated temporal data. In *IEEE EUROCON 2019-18th International Conference on Smart Technologies*, pages 1–6. IEEE.

Min Yen Wu, Chih-Ya Shen, En Tzu Wang, and Arbee LP Chen. 2020. A deep architecture for depression detection using posting, behavior, and living environment data. *Journal of Intelligent Information Systems*, 54(2):225–244.

Xuhai Xu, Bingsheng Yao, Yuanzhe Dong, Saadia Gabriel, Hong Yu, James Hendler, Marzyeh Ghassemi, Anind K Dey, and Dakuo Wang. 2024. Mental-llm: Leveraging large language models for mental health prediction via online text data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(1):1–32.

K Yang, S Ji, T Zhang, Q Xie, Z Kuang, and S Ananiadou. 2023a. Towards interpretable mental health analysis with chatgpt.

Kailai Yang, Tianlin Zhang, Ziyan Kuang, Qianqian Xie, Sophia Ananiadou, and Jimin Huang. 2023b. Mentallama: Interpretable mental health analysis on social media with large language models. *arXiv e-prints*, pages arXiv–2309.

Xiaoxu Yao, Guang Yu, Jingyun Tang, and Jialing Zhang. 2021. Extracting depressive symptoms and their associations from an online depression community. *Computers in human behavior*, 120:106734.

Andrew Yates, Arman Cohan, and Nazli Goharian. 2017. Depression and self-harm risk assessment in online forums. *arXiv preprint arXiv:1709.01848*.

Alexandros Dimitrios Zervopoulos, Evangelos Geramanis, Alexandros Toulakis, Asterios Papamichail, Dimitrios Triantafylloy, Theofanis Tasoulas, and Katia Kermanidis. 2019. Language processing for predicting suicidal tendencies: a case study in greek poetry. In *Artificial Intelligence Applications and Innovations: AIAI 2019 IFIP WG 12.5 International Workshops: MHDW and 5G-PINE 2019, Hersonissos, Crete, Greece, May 24–26, 2019, Proceedings 15*, pages 173–183. Springer.

Tianlin Zhang, Annika M Schoene, and Sophia Ananiadou. 2021. Automatic identification of suicide notes with a transformer-based deep learning model. *Internet interventions*, 25:100422.

Tianlin Zhang, Annika M Schoene, Shaoxiong Ji, Sophia Ananiadou, et al. 2022. Natural language processing applied to mental illness detection.

Liang Zhao, Jia Jia, and Ling Feng. 2015. Teenagers' stress detection based on time-sensitive micro-blog comment/response actions. In *Artificial Intelligence in Theory and Practice IV: 4th IFIP TC 12 International Conference on Artificial Intelligence, IFIP AI 2015, Held as Part of WCC 2015, Daejeon, South Korea, October 4-7, 2015, Proceedings 4*, pages 26–36. Springer.

Zhonghua Zheng, Lizi Liao, Yang Deng, and Liqiang Nie. 2023. Building emotional support chatbots in the era of llms. *arXiv preprint arXiv:2308.11584*.

Qiu-Yue Zhong, Elizabeth W Karlson, Bizu Gelaye, Sean Finan, Paul Avillach, Jordan W Smoller, Tianxi Cai, and Michelle A Williams. 2018. Screening pregnant women for suicidal behavior in electronic medical records: diagnostic codes vs. clinical notes processed by natural language processing. *BMC medical informatics and decision making*, 18:1–11.

Sicheng Zhou, Yunpeng Zhao, Jiang Bian, Ann F Haynos, Rui Zhang, et al. 2020. Exploring eating disorder topics on twitter: machine learning approach. *JMIR Medical Informatics*, 8(10):e18273.

Weipeng Zhou, Laura C Prater, Evan V Goldstein, Stephen J Mooney, et al. 2023. Identifying rare circumstances preceding female firearm suicides: validating a large language model approach. *JMIR mental health*, 10(1):e49359.

Yiheng Zhou, Catherine Glenn, and Jiebo Luo. 2017. Understanding and predicting multiple risky behaviors from social media. In *Workshops at the Thirty-First AAAI Conference on Artificial Intelligence*.

Ayah Zirikly, Philip Resnik, Özlem Uzuner, and Kristy Hollingshead. 2019. CLPsych 2019 shared task: Predicting the degree of suicide risk in Reddit posts. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*.

Hamad Zogan, Imran Razzak, Xianzhi Wang, Shoaib Jameel, and Guandong Xu. 2022. Explainable depression detection with multi-aspect features using a hybrid deep learning model on social media. *World Wide Web*, 25(1):281–304.