Social Network Analysis and Mining (2024) 14:78
https://doi.org/10.1007/s13278-024-01205-0

REVIEW PAPER

Unraveling minds in the digital era: a review on mapping mental health disorders through machine learning techniques using online social media

Aysha Khan[1,2] · **Rashid Ali**[1,2]

Received: 15 December 2023 / Revised: 15 December 2023 / Accepted: 18 January 2024

Abstract
People worldwide have suffered tremendously in terms of their mental health due to years of exposure to stress, anxiety, and the pressures of today's fast-paced lifestyles. The digitization of the data made possible by advances in health-care technology worldwide has allowed for a more precise mapping of the many variations of human biology than was possible before. People's methods of interaction with one another are evolving due to the rapid development of technology. Twitter, Facebook, Telegram, and Instagram have all risen to prominence as platforms where users can openly discuss their innermost thoughts, psyche, and feelings with one another. Texts are put through a psychological analysis process to pull out relevant details, characteristics, and user feedback. Psychological analysts rely on social media for the early identification of depressive symptoms and patterns of behavior. Machine learning has been recognized as a powerful method for sifting through the vast quantities of data in the health-care industry. Predicting the likelihood of mental diseases and executing likely treatment outcomes is a common application of ML techniques in mental health. This paper compiles a list of different mental health disorders along with the methods used in detecting and diagnosing mental health-related issues using online social media.

Keywords Mental health disorders · Machine learning · Natural language processing · Online social media · Text analytics · Text mining · Review · Survey

1 Introduction

With the emergence of online social media, exchanging information on these platforms has become significantly more popular. These social media platforms have become important facilitators in various areas, such as identifying the sentiments of people regarding a product, public opinions about various events, sharing daily life events, and discussion forums regarding diseases and medications (Dolezal et al. 2022 and Marsch 2021). There is a tsunami of posts by users on these social media platforms, and this user-generated content on social media has provided researchers with opportunities for exploring and categorizing this content in domains such as marketing, politics, and health. Social media platforms such as Facebook, Twitter, and Reddit turned out to be the safest space for people to post their emotions, feelings, thoughts, etc. (Yoo et al. 2019). With users being open about their mental states on such platforms by continuously posting about how they are feeling or tackling their mental disorders, their content has provided a rapid method for researchers to train classification models for detecting users with mental health disorders (Saha and Munmun 2021).

In the past decade, mental health has become a global health concern. From 2005 to 2015, the number of people having depression globally elevated by 18.4% [WHO, 1]. Mental health disorders can impair an individual emotionally and physically with common diseases such as stroke, heart disease, etc. People with mental health disorders also experience sleep disorders, low energy, helpless feeling, less interest in daily activities, suicidal thoughts, and intense concentration levels (Cohen et al. 2022). The World Health

✉ Aysha Khan
aisha.khan9413@gmail.com

Rashid Ali
rashidali@rediffmail.com

1 Department of Computer Engineering, Aligarh Muslim University, AMU, Aligarh, India

2 Interdisciplinary Centre for Artificial Intelligence, Aligarh Muslim University, AMU, Aligarh, India

Published online: 04 April 2024

Organization published a study in 2020 in which 264 million people were affected by depression (Depression 2021). In England, the treatment for mental health at work reached 105 billion British pounds (Health Matters, 2022).

Mental health disorders are treated using traditional approaches in which the patient interacts with the therapist in one-to-one interaction. These are reactive approaches in the form of one-to-one interviews, questionnaires, and surveys which makes it time-consuming and expensive (Bathina et al. 2021). These approaches are not that effective as they are dependent on factors such as the mood of the patient, the bond between the patient and the therapist, and the past experiences of the patient (Karcher et al. 2021). With the stigma attached to mental health disorders and the lack of awareness, patients hardly recognize their mental health in their initial stages and generally approach the clinician in the later stage of the disease. A study found that 70% of people do not understand their mental health until the symptoms are worse, and then in the severe stage, they approach the clinician (Shen et al. 2017).

1.1 Need for using AI for mental disorder detection

Limited mental health-care services with ineffective techniques have hampered the physician's time. Mental health practitioners need unique skills to provide personalized treatments to mental health patients to stimulate therapeutic support and medications to the patient and their family members, as every individual has a unique personality (Mendu et al. 2020). The current applications and techniques of artificial intelligence added a lot of support in the detection and diagnosis of mental health care. These techniques are widely used to draw insights from different data sources, fostering a better recognition of mental illnesses among the population by uncovering the factors leading to risk (Shrestha and Spezzano et al. 2019). AI algorithms also enable online therapeutic sittings along with the provision of self-assessment to patients who lack access to these services in certain areas. People do not generally consult doctors during their initial days of mental illness. Doctors cannot monitor all the patients simultaneously, so there should be a procedure that could help the doctors deal with the patients. The AI approaches have changed the treatment game by enabling a healthy clinician–patient relationship (Kiong 2022).

The public, private, and institutional stigma associated with mental health have made treating mental health more difficult. The prejudice related to mental health is because of the lack of understanding and the fear of losing livelihoods and social acceptance associated with it. Online platforms let an individual hide their identity, which gives an individual the confidence to share about their mental well-being (Hu 2022). These platforms allow users to post about their well-being in the form of either text or image, or video. Most of the data is available in text form on different social media platforms. This text can be analyzed with the help of text mining and sentiment analysis. Natural language processing (NLP) is the subfield of text mining that combines the power of linguistics, and computer science to make programs to analyze natural language data involving speech and text. The most common NLP tasks are speech recognition, natural language understanding, language translation, information extraction, and natural language generation. NLP studies the rules and structure of the language with the help of lexical, syntax, semantics and pragmatics, and morphology. NLP comes in handy when studying online social media, as the majority of the data available is in the form of text. Even, mental health procedures are highly dependent on NLP as there is an enumerable amount of raw textual data, i.e., clinical notes, and counseling sessions between the therapist and the patient (Ranjana et al. 2022). A basic application of NLP is encountered while using Gmail, where each mail is automatically classified as primary, social, promotion, and spam (Diniz et al. 2022). This is possible with the help of the keyword extraction technique of NLP. NLP is combined with computational linguistics, machine, and deep learning to empower the systems to investigate the human language in the form of speech to draw complete conclusions and intent (Ricard et al. 2018).

Existing research spanned a wide variety of methods for identifying mental states. Nonetheless, we think that our assessment is suitable for the use of computational techniques considering the rising popularity of using online social media to seek help and guidance for mental health disorders. The purpose of this review is to carefully examine the usefulness of extracting and analyzing information from online social media with NLP and ML approaches for individuals experiencing mental health-related issues.

In this research, we discuss about the multiple mental health disorders on online social media in which the features are taken from online textual cues and combined with machine learning, deep learning, and ensemble approaches to detect signs of mental episodes.

For health informatics in general and computational strategies in mental health diagnosis in particular, we believe that this review can be useful to researchers performing experiments with social media text data. In conclusion, our review's most significant findings are as follows:

1. The goal of this study is to conduct a literature analysis and methodology comparison on mental health detection and identification in online social media using computational methods.
2. Evaluation of current methods, as well as discussion of these methods, for text analytics in social media platforms online.

3. To share knowledge about the most pressing unanswered questions and viable computational solutions in the field of health-related textual information research.
4. The limitations of using online social media for research purposes.

The remaining paper is structured as follows: First of all, we discuss the previous review studies that have been published by various researchers in this field, and how our review is different from others. In the next section, we discuss the objectives of this study, and how the studies were selected and included in this work. After that, we discuss how the review studies are classified for our work. After that, we discuss the limitations of this review. Then, in the conclusion part, we compare and describe the current state of the art in natural language processing and machine learning methods for identifying mental health disorders on online social media. Finally, we provide a list of outstanding issues in the discussion section, outlining productive avenues for future study in this area.

2 Previous review studies

The proliferation of online social media has made it a topic of discussion among researchers. Researchers use online social media for multiple purposes such as health care, sentiment analysis, opinion mining, recommender systems, etc. An enormous amount of survey papers and review papers are available aiming at the use of machine and deep learning for mental health disorders detection using the social media posts of users.

A complicated multifactorial disease, mental illness is influenced by a number of socioeconomic and clinical factors, including individual risk factors. Natural language processing (NLP) techniques show potential in capturing these complex relationships represented in a wide range of textual data, including social media posts, interviews, and clinical notes. This will enable proactive mental health care and aid in early diagnosis. In order to comprehend methodologies, trends, obstacles, and future directions, the authors gave a narrative review of mental illness detection using NLP over the previous 10 years. In this study, 10,467 records representing a total of 399 studies were included. The review demonstrates an upward trend in NLP research on mental disorder identification. Traditional machine learning techniques are outperformed by deep learning techniques, which are given greater attention. The authors also offered suggestions for further research, such as creating fresh detection techniques, deep learning paradigms, and understandable models (Zhang et al. 2022). A systematic review of using machine learning and natural language processing techniques on mental health and their use in the medical field is presented (Le Glaz et al. 2021). Salas et al. reviewed 34 studies for detecting depression from online social media using machine learning from 2016 to mid-2021. The digital libraries included for primary studies are: ACM Digital Library, IEEE Xplore Digital Library, SpringerLink, Science Direct, Google Scholar, and PubMed. The social media platform most extensively researched for identifying depression symptoms was Twitter. The most used linguistic feature extraction technique was word embedding. The most popular machine learning algorithm was support vector machine (SVM) (Salas et al. 2022). Liu et al. conducted a systematic review on using machine learning methods on social media text for detecting depressive symptoms from January 1990 to December 2020. The authors used various search terms to search for the papers for the review. Then, the authors also shared various inclusion criteria to finally include the papers for their study (Liu et al. 2022). Esteva et al. presented a survey paper discussing various deep learning techniques for health care. The major areas covered were computer vision for medical imaging, natural language processing for electronic health records, reinforcement learning for robotic-assisted surgery, and generalized methods for genomics (Esteva et al. 2019). Miotto et al. discussed the recent literature available on using deep learning for health care with the challenges and opportunities associated with them for electronic health records, genomics, mobile, and clinical imaging (Miotto et al. 2017). Rahmani et al. presented a review paper focusing on the use of machine learning in medicine with the pre-processing steps, type of machine learning technique used, evaluation metrics, and applications (Rahmani et al. 2021). Kim J et al. presented a review study on using machine learning for mental health detection on social media data. They presented a literature study of articles from 2015 to 2020 and obtained 565 relevant papers. They mainly focused on papers from Web of Science and Scopus (Kim et al. 2021).

Su et al. presented a detailed study on the uses of deep learning for mental health. They conducted their study based on clinical data, vocal, and visual expression data and social media data (Su et al. 2020). The many methods for data gathering, the most recent trends and technologies in this area, and the current uses of ML and NLP in the monitoring of public mental health are all covered, along with the drawbacks and the gaps discovered while doing research in this area (Skaik et al. 2020). In order to determine the state-of-the-art social media data prediction of mental health status, researchers undertook a comprehensive literature evaluation with an emphasis on the study design, methodologies, and research design of their study. Between 2013 and 2018, 75 studies in this field were selected. The findings describe the procedures for annotating data for mental health status, gathering data and managing data quality, selecting features for pre-processing, and choosing and validating models. The authors presented alarming tendencies regarding construct validity and a lack of reflection in the approaches taken to operationalize and identify mental health status despite the field's growing interest. The authors also offer suggestions to solve these issues, such as a set of suggested publishing reporting criteria and chances for multidisciplinary collaboration (Chancellor and Chaudhary 2020). The paper aims to thoroughly evaluate the literature on NLP and ML techniques used to detect depression in Online Support Forums (OSF). A thorough search was conducted to find studies that looked at ML and NLP methods to distinguish depression disorder from OSF. The PRISMA method was used to choose the articles. A total of 29 publications were chosen and studied for the review. Using the results of this comprehensive study, the authors further examined which combination of attributes obtained using NLP and ML algorithms is suitable for modern depression identification (Nanomi et al. 2021).

Most of the existing review studies have focused on the research published over the course of the past decades. However, this review study specifically directs its attention to the research conducted in the past few years and has a particular emphasis on the studies published in the past 2 years. This review presents the studies on the basis of different mental health disorders on online social media using different techniques. In other studies, the authors have covered mainly the type of mental health disorder and the techniques used. Our review study also focuses on the models used for comparison along with evaluation metrics discussed in the papers and also clarifies whether the research is based on binary classification or multi-classification, topic modeling along with the social networking site used for the dataset and the type of features used.

3 Study objectives

The objective of this study is to present a review of studies that have used social media data for detecting and treating mental health disorders using machine and deep learning. We have tried to review the most updated papers in this area in the past 5 years, i.e., January 2018 to August 2022, while the main focus is to review the studies in 2020–2022.

3.1 Data inclusion

For our study, we start searching using terms such as mental health disorders, stress detection, online social media, and word embeddings. The research is focused on presenting a review of the papers which have utilized datasets from online social media data, including Twitter, Reddit, Weibo, and Facebook. The study focuses on deep learning and machine learning methods to detect mental health from social media data. We have searched various databases for appropriate results, i.e., Google Scholar, PubMed, Nature, Web of Science, JMIR Mental health, NCBI, Medline, ACM Digital Library, MDPI, IEEE Xplore, Science Direct, and Springer Link. We do not have any restrictions on the type of article. The papers are collected from 2018 to 2022.

For our search, we used multiple general terms related to mental health, such as mental health disorder detection, mental issues, and many more. We also used specific mental health disorder terms such as depression, anxiety, etc. We searched for papers focusing on mental health disorders on online social media such as Twitter, Facebook, and Reddit. The focus of the search is only on the machine and deep learning techniques using natural language processing, text mining, etc. Multiple combinations of these queries are sent in parallel to different search databases, as mentioned above. We can understand this with the help of some examples:

Query 1 (Q1): Detecting mental health disorders on online social media using machine and deep learning.

In query 1, we combined the term machine and deep learning with natural language processing, text mining, and other terms mentioned in the methods column of Table 1.

Query 2 (Q2): Depression detection using machine and deep learning on Twitter.

Here, in query 2, we replaced the 'depression' for every mental health disorder that we mentioned in the mental health disorder column in Table 1, and the machine and deep learning term is replaced with terms mentioned in the methods column in Table 1.

Query 3 (Q3): Predicting mental health disorders using transfer learning on Twitter.

Here, in query 3, we replaced 'Twitter' with other online social media platforms mentioned in the dataset column of Table 1 and transfer learning with different values in the methods column in Table 1.

Similarly, we generated multiple queries like this by combining the terms mentioned in Table 1. These queries are combined with different combinations mentioned in the table to get a universal set for our research.

$$Q1UQ2UQ3UQ4UQn = \hat{U}$$

Here $Q1$, $Q2$, $Q3$,…Qn represents the different combinations of the queries that we have used for searching our papers.

$\hat{U}$ represents the final set of papers obtained from all the different combinations of these queries by running a search on all the databases mentioned above.

Table 1 Different category and the corresponding keywords used for paper extraction

Category	Keywords
Mental health disorder	Mental illness, mental health, mental disorders, mental issues, mental disorder detection, and predicting mental disorders Depression, suicide, stress, anxiety, schizophrenia, bipolar disorder, BPD, autism, anorexia, self-harm, and PTSD
Dataset	Online social media, text, social media posts, social media text, tweets, subreddits, Twitter, Reddit, Weibo, Facebook, and Instagram
Methods	Natural language processing, text mining, text analysis, deep learning, artificial intelligence, machine learning, word embeddings, transformers, attention networks, transfer learning, neural networks, document classification, ensemble methods, and deep neural networks CNN, RNN, LSTM, Bi-LSTM, GRU, LSTM, SVM, RF, LDA, and LSA
Word embeddings	Word2vec, Glove, Bert, FastText, and ELMO
Technique	Classification, clustering, topic modeling, prediction, detection, binary, and multi
Evaluation criteria	Accuracy, recall, precision, $F1$-score, coherence, and perplexity

These keywords were combined using the Boolean operator ‘or’ and ‘and.’

3.2 Exclusion criteria

In this section, we mention the criteria used to finalize the papers for this review. We have identified some exclusion criteria for the purpose of this review study. These factors are used to exclude the papers which we considered to be out of the scope of this study.

- If the article is not focused on using natural language processing, text mining.
- If the article is based on images, videos, or speech.
- If the abstract was found later irrelevant, not relating to the mental health data.
- If the article is just doing statistical analysis for general individuals having mental health disorders without using machine learning techniques.
- If the article is focused on surveys, etc.
- Research not focusing on ML and DL techniques.
- Research not focusing on online social media.
- Research on patient data in hospitals.
- Research only performing text pre-processing.
- Research focusing on physical health disorders.

The exclusion criteria help us to focus on the objective of this research paper. Although there have been many studies published in this area as this is a multidisciplinary research field covering computer engineering, psychology, and medicine. There were studies that have used machine learning techniques for mental health disorders, but they were focused on hospital-based data, so we have excluded such studies. Also, we encountered many studies focused on mental health disorder detection on online social media, but their only focus was doing text analysis and feature extraction, and no algorithms were used in their paper. In this way, we removed the papers according to the exclusion criteria presented above to focus on the studies helpful for this review.

3.3 Results

In this study, 4000 review studies were identified from different database searches, as discussed in the data inclusion section. These studies were then screened for any duplicates. After duplication removal, only 3676 studies were obtained. Then, these studies were searched for full-text access. After rigorous searching, we gathered about 1987 studies for our study. These studies were further evaluated to fall into our search criteria as discussed in Fig. 1 We removed the studies which were out of the scope of this review based on the exclusion criteria discussed above. Finally, we were left with 99 studies presented in this paper.

Now, for the purpose of understanding the review study in a more comprehensive manner, we have discussed some research questions. These questions helped us in forming the basis for our study. Some of the research questions are discussed below.

Research Question 1: Which online social media platform is most commonly used for mental health disorder detection?

Answer: There have been many online social media platforms that exist nowadays. The most well-known and most used social media platforms by users and researchers can be summarized below. These social media platforms are not the only ones used by researchers but are also most commonly used by individuals for sharing their opinions and thoughts.

3.3.1 Reddit

Reddit is a social media platform founded in 2005 by Steve Huffman and Alexis Ohanian. The site is organized around user-created communities, known as subreddits, which cover a wide range of topics, from news and politics to hobbies and

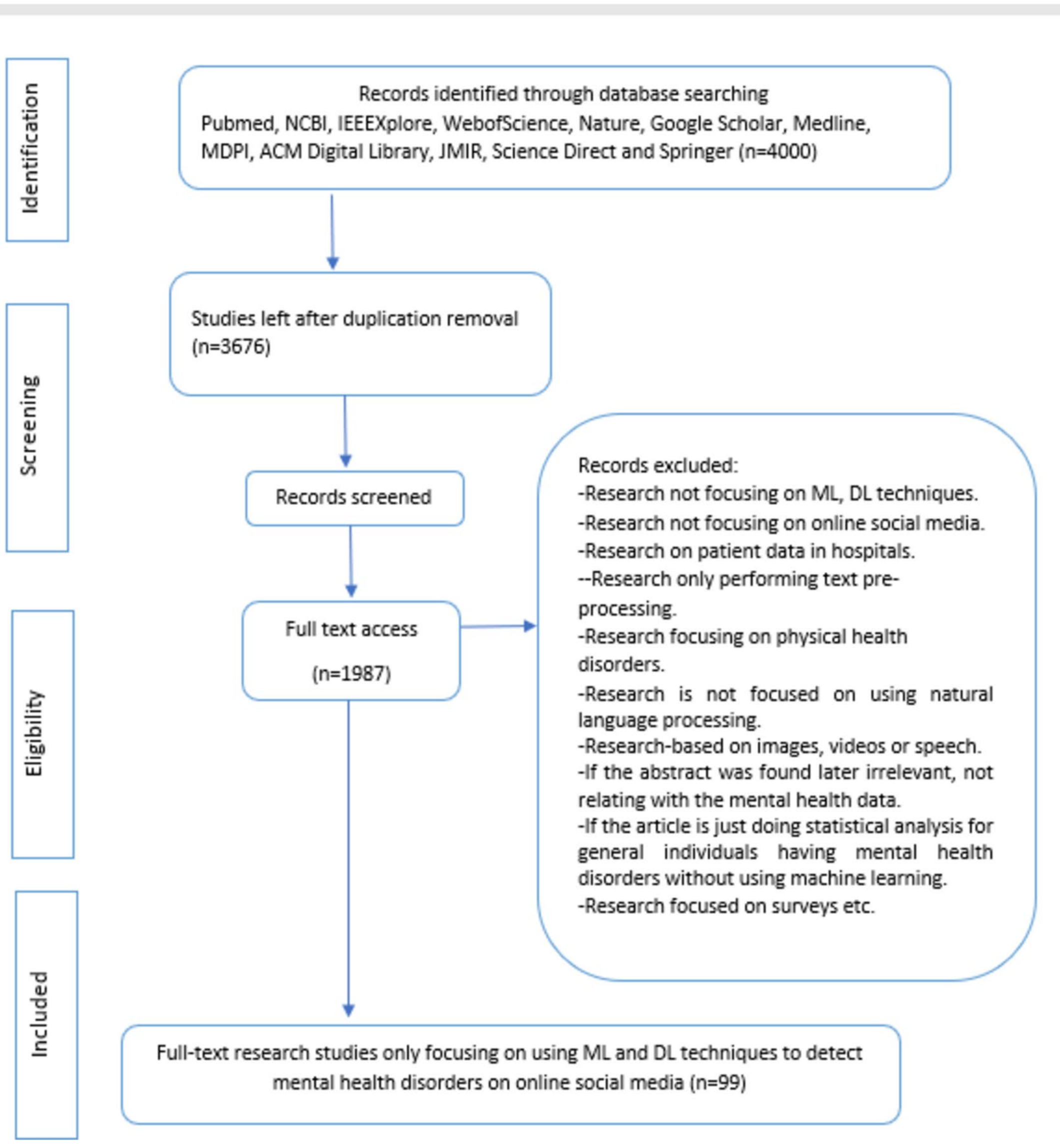


Fig. 1 PRISMA flow diagram for searching of the literature studies

interests. As of 2021, Reddit has over 430 million monthly active users, making it one of the largest social media platforms in the world (McAuliffe et al. 2022).

One of the unique features of Reddit is its Ask Me Anything (AMA) interviews. These sessions allow users to ask questions directly to a notable person, such as a celebrity or expert in their field. These interviews can provide insights into a wide range of topics and give users a chance to interact with people they may not usually have access (Fraga et al. 2018). In addition to its social and cultural impact, Reddit has become an important tool for researchers and scientists. Due to the platform's unique features, such as its subreddits and threads, researchers can analyze the language and experiences shared by users to gain insights into a wide range of topics, including mental health disorders, political beliefs, and more. Reddit is becoming an increasingly popular tool for mental health research by medical professionals and computer scientists. Due to the platform's large user base and the abundance of user-generated content, it provides a unique opportunity to gain insights into the experiences and behaviors of people with mental health disorders (Ptaszynski et al. 2021).

One-way Reddit is used for mental health research which is through natural language processing (NLP) techniques. Computer scientists and linguists can use NLP to analyze the language used in Reddit posts and comments to identify patterns and gain insights into the experiences and emotions of people with mental health disorders. For example, researchers can use NLP to analyze subreddits dedicated to depression or anxiety to identify common themes and emotions expressed by users. This can help researchers understand the experiences of people with mental health disorders and inform the development of new treatments or interventions (Lau et al. 2016). Medical professionals are also increasingly using Reddit to connect with patients and gain insights into their experiences. For example, psychiatrists and therapists may browse relevant subreddits to understand the experiences and perspectives of people with specific mental health disorders (Chandra Guntuku et al. 2019). This can help them provide better patient care and develop more effective treatment plans. In addition, researchers can use Reddit to recruit participants for mental health studies. They can post study advertisements on relevant subreddits or reach out to users who have shared their experiences with mental health disorders to see if they are interested in participating in a study (Shigemera et al. 2020). This can help researchers recruit a diverse and representative sample of participants for their studies. Overall, Reddit is becoming an important tool for mental health research by medical professionals and computer scientists. It provides a unique opportunity to gain insights into the experiences and behaviors of people with mental health disorders and has the potential to inform the development of new treatments and interventions. However, it is important to use Reddit ethically and responsibly and to respect the privacy and confidentiality of users who share their experiences on the platform (Chang and Tseng 2020).

3.3.2 Twitter

Twitter is a popular social media platform that was founded in 2006. It allows users to post short messages, known as tweets, of up to 280 characters. The platform has grown in popularity over the years and has become a powerful tool for communication, news dissemination, and social engagement. As of the first quarter of 2021, Twitter had 199 million active daily users worldwide, according to Statista. While this is a relatively small number compared to other social media giants like Facebook, Twitter's active user base is highly engaged and influential. Every day, millions of tweets are posted on Twitter, with an average of 500 million tweets sent per day in 2020 (Twitter 2022). This high volume of tweets means that the platform generates a massive amount of data that can be analyzed and leveraged by researchers, businesses, and other organizations. The platform is available in over 40 languages, and users can connect and engage with others around the world. This has led to the development of Twitter communities and movements that have had a significant impact on social and political issues (Chatterjeeet al., 2021). Twitter is a powerful tool that has become increasingly valuable to mental health researchers. Its vast and diverse user base provides researchers with a unique opportunity to analyze how people with mental health disorders are coping, connecting, and communicating in real-time (Ríssola et al. 2021).

One of the primary ways researchers use Twitter is by analyzing the language used in tweets (Sakib et al. 2021). Overall, Twitter is a valuable tool for researchers, offering a unique and dynamic platform to analyze the experiences and behaviors of people with mental health disorders. As the platform continues to evolve and adapt to changing user needs and trends, it will be interesting to see how it can be further leveraged for mental health research and support (Andy 2021).

3.3.3 Facebook

Facebook is a social networking site founded in 2004 by Mark Zuckerberg and his college roommates at Harvard University. Initially, the platform was only accessible to college students, but it quickly expanded to include anyone with an email address. Today, Facebook is one of the most popular social media platforms in the world, with over 2.9 billion monthly active users as of 2021 (The Latest Facebook Statistics 2023). Facebook is a highly interactive platform that allows users to connect with friends and family, join groups, and share content such as photos, videos, and written posts. It also offers a range of features such as event invitations, a marketplace, and fundraising tools, making it a versatile tool for both personal and professional uses. One of the unique aspects of Facebook is its ability to facilitate communities and group connections. Users can join groups based on shared interests, geographic location, or other factors, allowing them to connect with others with similar experiences or concerns. This has made Facebook a powerful tool for social movements and advocacy groups, as well as for businesses and organizations looking to build a loyal customer base (Argyris et al. 2022). In addition to its social and commercial potential, Facebook has become an important tool for researchers and health-care professionals interested in mental health. Researchers can use Facebook to recruit participants for studies, collect data, and analyze patterns in user behavior (Shrestha et al. 2020). Despite its many benefits, Facebook has also faced criticism over the years for its handling of user data and privacy concerns. The platform has implemented a range of measures in response, such as enhanced privacy settings and increased transparency around data usage. However, concerns around these issues continue to be a topic of debate among users and critics alike. Overall, Facebook has had a significant impact on the way we connect and communicate with each other, both personally and professionally. As the platform continues to evolve and adapt to changing user needs and trends, it will be interesting to see how it continues to shape the world of social media and beyond (Behesti et al., 2020).

3.3.4 Seina Weibo

Seina Weibo, often referred to simply as Weibo, is a microblogging platform that was launched in China in 2009. The platform is similar in many ways to Twitter, allowing users to post short messages and updates, as well as photos and videos. Weibo quickly became one of the most popular social media platforms in China, with over 500 million registered users as of 2021 (Yang et al. 2022a, b). The platform's popularity is due in part to its ability to allow users to share their thoughts and experiences in real-time, as well as to connect with others who share similar interests and concerns. In addition to its social potential, Weibo has become an important tool for businesses and organizations looking to reach Chinese consumers (Wang et al. 2020a, b). The platform offers a range of advertising and marketing tools, as well as accesses to a massive user base that is highly engaged and active. Weibo has also been used by researchers and mental health professionals interested in understanding mental health in China (Lyu et al. 2022). The platform offers a unique opportunity to access a large and diverse group of users and to analyze patterns in user behavior and communication related to mental health. This has led to various studies on topics such as depression, anxiety, and other mental health issues. Overall, Weibo is a highly influential platform in China, offering users a powerful tool for communication, connection, and self-expression. As the platform continues to grow and evolve, it will likely remain an important part of the Chinese social media landscape, with implications for social and political discourse, business and commerce, and mental health research and practice.

During our research, we came across multiple standard datasets from Reddit and Twitter, available online. Most of the studies used these two platforms for the research purpose. Limited studies have used Facebook and Seina Weibo. The total studies reviewed in this paper can be visualized with the help of a pie chart in Fig. 2. As can be seen, the maximum studies obtained are from the Reddit social media platform, which accounts for 48% of the total studies. The least studies discovered are from the Seina Weibo platform due to the reason that it is specific for the Chinese population, while 2% of the studies did not specify the exact social media platform, they are using while just mentioning about the posts, comments, etc., being used for the study. These studies are further elaborated in the upcoming sections.

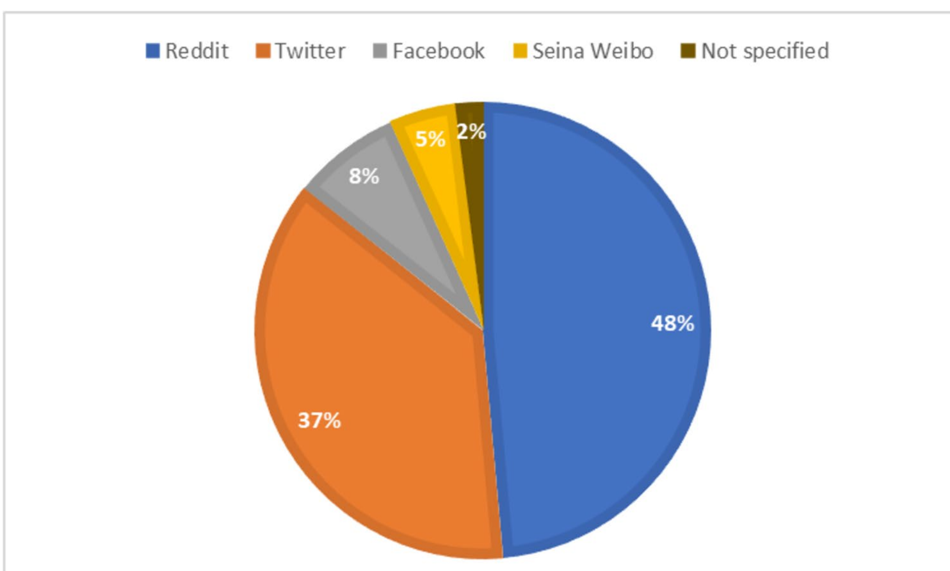
Research Question 2: What are the different algorithms used for mental health disorder detection on online social media?

Answer: The algorithms that we found in our review studies can be broadly classified as given below:

3.4 Machine learning

Earlier machines were programmed with computational algorithms to follow certain instructions to solve specific problems. Machine learning is the branch of artificial intelligence that empowers machines to learn and behave like humans. Humans have the intelligence to behave according to past experience, machines were lacking in this quality. Machine learning has enabled machines to learn from past experiences to perform specific tasks with specific performance. In 1995, Allan Turing introduced a 'Turing Test' which helped in evaluating machines on the basis of their intelligence. In 1995, Arthur Samuel described machine learning as 'a study which gives computers learning abilities without programming explicitly.' In machine learning,

Fig. 2 Total studies in the review paper focusing on multiple social media platforms



machines are fed with data and are allowed to learn on their own without any human intervention. Machine learning algorithms are generally comprised of three parts (Myszczynska 2020). First is the decision process, which generally takes some labeled or unlabeled data, and then makes either a prediction or classification. Then, next, we have the error function, which evaluates the model prediction capability. Then, we have model optimization, where the weights are adjusted to reduce the difference between the actual and the predicted values until a threshold accuracy is reached. Sometimes, these models are easily understandable by humans and sometimes, they are complex to understand, similar to a black box. Machine learning algorithms are also as flat algorithms as they cannot be applied directly to the raw data. If the model is not able to capture the underlying patterns of the data and cannot perform well on the testing data, underfitting occurs. When the model is very complex and is trained on too much data, it starts capturing the noise in the data. Such models perform well on the training data but cannot make accurate predictions, in this case, overfitting is said to occur (Rezaii et al. 2019).

3.5 Supervised machine learning (SML)

As the name indicates, in supervised machine learning, we have a supervisor who supervises the learning of the algorithm (Azam et al. 2021). In SML, we have labeled data in which each input has an associated label. We train or supervise our model using this data. The labeled data can be categorical (Yes or no) or continuous (range of values). In SML, we have some input and output variables, and the focus is to have a mapping function from input to output using an algorithm. This mapping function should be able to predict the new data without any labels based on the features provided and the training process (Kang et al. 2016). In SML, we have the correct answers, and the algorithm keeps making predictions on the training data. When the model makes wrong predictions, they are corrected just like a supervisor corrects their students. The model stops learning when the desired performance is reached. Regression and classification problems are the main types of supervised learning problems (Deng et al. 2021).

3.6 Unsupervised machine learning (UML)

In unsupervised machine learning algorithms, data are not provided with labels, and learning happens on its own without any supervision by finding similarities in the input data and getting insights about the structure of the data. In this, we only have the input data and no output labels. It is unsupervised because there is no correct answer as they are identified by different experts. UML enables users to perform more complex operations than SML. UML is more challenging than SML, because of the absence of labels within the dataset. Clustering, dimensionality reduction, and association are the main tasks of unsupervised learning algorithms (Westrupp et al. 2022).

3.7 Deep learning

Deep learning is the branch of machine learning which is influenced by the functioning of the human brain (Saravanan et al. 2022). When humans get any new information, they try to compare it with already available known objects. DL enables the machines to focus on the right set of features by themselves without any or little human intervention (Durstewitz 2019). It used multi-layer structures known as neural networks. Deep learning is capable of focusing on the right set of features requiring very little guidance by the programmer by reducing the dimensionality of the features. It basically skips the manual step of feature extraction as that of machine learning. Deep learning uses neural networks for processing large datasets through multiple hidden layers. The input data passes through the nodes of the neural networks, with each node having some weight. Each layer helps the next layer in refining and optimizing the outputs, and this phenomenon is known as forward propagation. The node which has a higher weight is more important than the node which has a lower weight. In the input layer, we provide the same input entries as our input data, and the output layer consists of the actual output that should come from our data (Sun et al. 2022).

3.8 Attention networks

Attention is a method used in artificial neural networks to imitate cognitive attention. This effect makes some portions of the input data better while making other parts worse. This is done to encourage the network to pay more attention to the critical parts of the data, even if they only make up a small piece of an image or text (Rutowski et al. 2020). Gradient descent trains an algorithm that determines which portion of the data is more relevant than another based on the context (Haque et al. 2020).

In the 1990s, mechanisms resembling attention were developed, known as multiplicative modules, sigma pi units, and hyper-networks. In contrast with normal weights, which must remain fixed during runtime, 'soft weights' can alter during runtime, giving them flexibility (Alsayat 2022).

Figure 3 gives us an estimation of the multiple techniques that have been applied by researchers in their papers. Most of the researchers have used deep learning techniques accounting for 44% of our review studies, followed by supervised machine learning techniques. While a very little amount of work is done using unsupervised machine

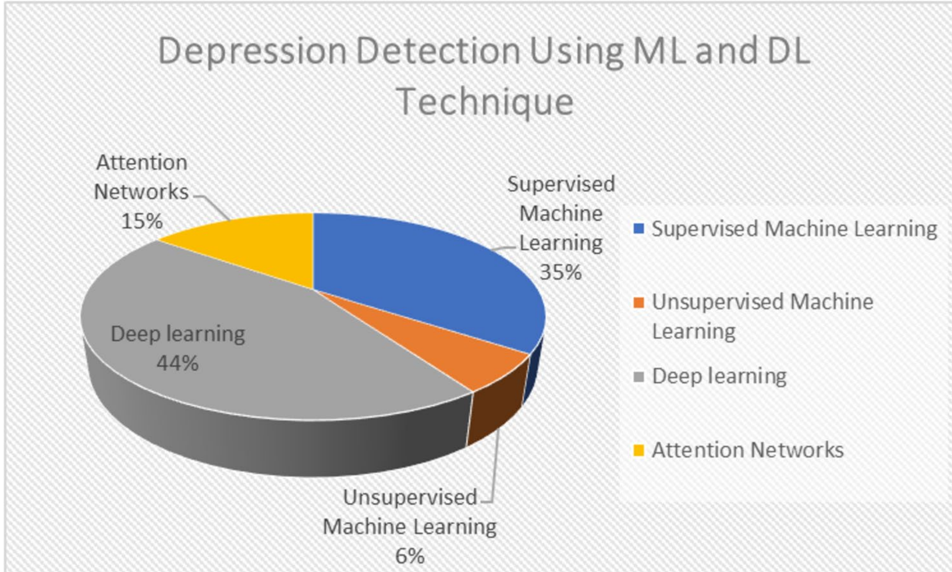


Fig. 3 Studies reviewed using different techniques using a 3-D pie chart view representation

learning accounting for 5% only. This is due to the fact that mental health disorder detection is more of a classification task than a clustering task.

Research Question 3: What are mental health illnesses that are mostly recognized on online social media?

Answer: In this paper, we have focused on reviewing the papers based on mental disorders. In those mental disorders, we further categorized the disorders on the basis of which techniques are used to identify the disorders.

4 Classification of mental health disorders on online social media using different evaluation techniques

4.1 Depression

Depression is a mood disorder that enables sadness and loss of interest. Depression affects a person's thinking and behavior, leading to multiple emotional and physical problems hampering the everyday activities of an individual. Depression affected 350 million people globally (Thij et al. 2020). Many people seek medical advice when depression becomes severe, which has many health and economic effects (Maxim et al. 2020). Early detection of depressive symptoms and those at risk of developing them in the community is helpful for formulating public policies and may open up new doors for early intervention or online referrals of those in need of efficient preventive measures (Alexopoulos 2019 and Chiong et al. 2021). Nowadays, online social media has increased the tendency of people to talk about their mental state openly (Depression 2022.)

In this survey, we found several studies on depression detection using machine and deep learning on online social media. We discovered many studies focusing on depression, so we are bifurcating the studies into supervised, unsupervised machine learning, deep learning, and attention networks for the purpose of clarity in Tables 2, 3, 4, and 5, respectively.

The goal of depression detection is to create a predictive model that can quickly identify textual information related to mental health issues and identify persons with depression from tweets of Twitter users. Using a regular expression or stream of real-time tweets made up of 3682 individuals, 1983 of whom self-declared having depression and 1699 of whom did not were used in this work. To recognize people with depression and draw attention to postings pertaining to the author's mental health, two multiple-instance learning models (MIL)—one with and one without an anaphoric resolution encoder—were created. Anaphora resolution is a text analysis problem that focuses on identifying which person is referenced in which textual context. In this paper, the authors use the MIL approach to create two models: MIL-SocNet, which is multiple-instance learning for social networks, and (MILA-SocNet), which is multiple-instance learning for social networks using anaphora resolution, to identify users who are depressed and emphasize posts that have been written on the user's issue of mental health. Instead of a document vector, both models employ new document segment encoding, a tweet encoder, and user representation. The performance is further enhanced by the inclusion of the anaphora resolution in the latter model. The anaphoric resolution model achieved 92% accuracy, 92% $F1$-score, and 90% area under the curve (AUC) at its best. The model's prediction abilities were superior to those of both traditional machine learning and deep learning alternatives (Wongkoblap et al., 2018).

Table 2 Depression detection using supervised machine learning techniques on online social media

S. No.	Dataset	References	Types of features	ML technique	Evaluation metric	Number of posts	Multi /binary class/cluster
1	Reddit (eRisk 2017 and eRisk 2018	(Ortega et al. 2022)	Phrase analysis	SVM, RF, NB, DT	Precision, recall, $F1$-score	887 and 1707, respectively	Binary
2	Reddit	(Fatima et al. 2019)	LIWC Dictionary	SVM, LR, MLP	Accuracy, precision, recall	3176	Binary
3	Reddit	(Trifan and Antunes et al. 2020)	Bag-of-words and TF-IDF	NB, SVM, stochastic gradient descent	Precision, recall, $F1$-score, accuracy	969	Binary
4	Reddit	(Tadesse et al. 2019)	N-grams, LIWC dictionary, LDA topics	LR, RF, SVM, adaptive boosting, MLP	Accuracy, $F1$-score	1841	Binary
5	Reddit	(Cacheda et al. 2019)	TF-IDF, LSA topics	RF	Precision, recall, $F1$-score	49,557	Binary
6	Twitter (3 datasets)	(Skaik et al. 2021)	Depression	SVM, RF, DT, LR, GBDT	$F1$-score	292,564, 1 million, and 10,594,841, respectively	Binary
7	Twitter	(Kamite et al. 2020)	Sentiment score	NB, RF	$F1$-score	Not specified	Binary
8	Twitter	(Govindasamy and Palanichamy 2021)	Sentiment analysis using TextBlob	NB, NBTree	Accuracy, recall, precision	1000 and 3000	Binary
9	Twitter (Arabic)	(Almouzini et al. 2019)	Bag-of-words and Negation handling	RF, NB, AdaBoost, Liblinear	Accuracy	6122	Binary
10	Twitter (Tsinghua Twitter Depression Dataset (TTDD) and the CLPsych 2015 Twitter Dataset (CLPsych2015)	(Tong et al. 2022)	LDA topics	Cost-Sensitive Boosting Pruning Trees (CBT)	Accuracy, $F1$-score	66,672 for TTDD, 873 for CLPsych	Binary
11	Twitter (English and Spanish)	(Coello-Guilarte et al. 2019)	Bag-of-words, LIWC dictionary	SVM	$F1$-score	1335, 7991 average tweets per user	Binary
12	Twitter	(Angskun et al. 2022)	Polarity score	RF, SVM, LR, DT, NB, ANN	Accuracy, precision, recall, F-measure	222	Binary
13	Reddit (eRisk 2017	(Briand et al. 2018)	Bag-of-words, N-grams	Logistic model tree, ensemble of sequential minimal optimization, ensemble of random forests	Precision, recall, $F1$-score, early risk detection error (ERDE)	531,394	Binary
14	Facebook	(Katchapakirin et al. 2018)	Not clearly specified	SVM, KNN, deep learning	Accuracy, precision, recall, F-measure	1105	Binary
15	Twitter	(Wongkoblap et al. 2021)	Glove embeddings	MIL-SocNet	$F1$-score, precision, recall, accuracy,	4892	Binary
16	Facebook	(Islam et al. 2018)	LIWC dictionary	DT, KNN, SVM, ensemble	Precision, recall, F-measure	7145	Binary
17	Twitter	(Prakash et al. 2021)	Polarity score	SVM, RF	Positive, negative score	Not specified	Binary
18	Twitter	(Sarkar et al. 2022)	Depression	SVM	Accuracy	2022	Binary

Table 2 (continued)

S. No.	Dataset	References	Types of features	ML technique	Evaluation metric	Number of posts	Multi /binary class/cluster
19	Twitter	(Chatterjee et al. 2022)	Sentiment score, LDA topics, temporal features, social features	Multinomial NB using TF-IDF	Accuracy, precision, recall, *F*1-score	188,704	Binary
20	Facebook	(Kumar et al. 2022)	Linguistic, topics, sentiment, emotion features	LDA- based SVM	Accuracy	7774	Binary

User-generated information from Twitter can be used to examine the dynamics of depression in communities affected by the COVID-19 outbreak. To construct depression classification models, the authors present a novel method based on multimodal characteristics extracted from tweets and term frequency-inverse document frequency (TF-IDF). Multimodal characteristics capture depression indications from affective, topical, and disciplinary angles by analyzing recently scraped tweets from New South Wales, Australia, residents. During the COVID-19 timeframe, this unique classification approach can extract depression polarities that may be influenced by COVID-19 and related events. The results showed that after the COVID-19 epidemic, people's levels of depression increased. State lockdown and other government interventions contributed to an already depressing atmosphere. The LDA model was used to classify these tweets and was compared with other models also (Zhou et al. 2021).

The paper by Naseem et al. focuses on the application of personalized mental health interventions using natural language processing (NLP) and attention-based in-depth entropy active learning. The objective of this research is to increase the trainable instances using a semantic clustering mechanism. For this purpose, the authors proposed a method based on synonym expansion by semantic vectors. Semantic vectors based on semantic information derived from the context in which it appears are clustered. The resulting similarity metrics help to select the subset of unlabeled text by using semantic information. The proposed method separates unlabeled text and includes it in the next active learning mechanism cycle. The method updates model training by using the new training points. The bidirectional long short-term memory (LSTM) architecture with an attention mechanism achieved 0.85 receiver operating characteristic (ROC curve) on the blind test set. The learned embedding is then used to visualize the activated word's contribution to each symptom and find the psychiatrist's qualitative agreement (Naseem et al. 2022).

A novel model, explainable multi-aspect depression detection with hierarchical attention network (MDHAN), to automatically identify depressed people on social media and to provide an explanation for the model's predictions is proposed. The authors considered user's posts that have been enhanced with extra Twitter features. In particular, the model computed the importance of each tweet and word, encoded user postings using two levels of attention mechanisms applied at the tweet level and word level, and extracted semantic sequence features from user timelines (posts). This hierarchical attention model was created in a way that it can identify patterns that result in incomprehensible findings. The tests reveal that MDHAN beats a number of well-known and reliable baseline techniques, illuminating the potency of fusing deep learning

Table 3 Depression detection using unsupervised machine learning techniques on online social media

S. No.	Dataset	References	Type of feature	ML and DL technique	Evaluation metric	Number of posts	Multi /binary class/cluster
1	Twitter	(Zhou et al. 2021)	Emotional level, topic level, domain-specific features, TF-IDF	LR, LDA, GBM, multimodal feature extraction with TF-IDF	Precision, recall, $F1$-score, accuracy	94,707,264	Binary
2	Seina Weibo	(Liu and Shi 2022)	Part-of-speech tagging	LDA	Mean, SD	396,152	Binary
3	Reddit	(Sik et al. 2021)	LDA topics	LDA	No metric	67,857	cluster
4	Twitter	(Safa et al. 2021)	Polarity score, LIWC dictionary	N-grams, LIWC	Accuracy	11,890,632	Binary

with multi-aspect information. Additionally, it is also demonstrated how this algorithm enhances predictive performance when identifying depression in people who openly post messages on social media. MDHAN performs exceptionally well and guarantees that there are enough data to support the prediction. The model performs the best in comparison with all the other state-of-the-art models and achieved the highest accuracy, precision, recall, and $F1$-score (Zogan et al. 2022).

4.2 Schizophrenia

Schizophrenia is a severe mental illness in which reality is perceived by sufferers strangely. Schizophrenia may include hallucinations, delusions, and severely irrational thinking and behavior, which can make it difficult to go about daily activities and be incapacitating. Schizophrenia patients require ongoing care. Early intervention may help keep symptoms under control before major issues arise and may enhance the prognosis in the long run (Schizophrenia, 2020). The study focusing on detecting schizophrenia using online social media is presented in Table 6. The purpose of this study is to investigate whether social media user writings can be utilized to detect indicators of schizophrenia using machine learning. For the control group, the authors gathered postings from the social media site Reddit that discussed schizophrenia as well as those about fitness, humor, meditation, parenthood, relationships, and teaching and identified linguistic markers of schizophrenia by classifying posts as belonging to schizophrenia using supervised machine learning and analyzing significant aspects. For the purpose of identifying a cohesive semantic representation of words in schizophrenia, an unsupervised clustering algorithm was applied to the features. The greater usage of third-person on plural pronouns, words that express negative emotions, and topics relating to symptoms are only a few of the linguistic traits, and topics were found to be significantly different. The accuracy achieved is 96% (Bae et al. 2021).

4.3 Psychopath

A psychopath is defined as an individual with an egotistical and antisocial disposition who lacks regret for their actions, lacks empathy for others, and frequently exhibits criminal tendencies. Instead, it is a colloquial word frequently applied to the disease known as antisocial personality disorder (ASPD) (Psychopathy 2023). During our study, we came across a few studies detecting psychopaths on online social media, as shown in Tables 7 and 8, respectively.

The majority of the work that has hitherto been done on psychopath detection has been done in the psychology field using conventional methods, such as the SRPIII technique with small dataset sizes. This encourages to develop a sophisticated computational model for psychopath diagnosis in the field of text analytics. In this study, attention-based Bi-LSTM for psychopath detection with a larger dataset size is investigated for its effectiveness in classifying input text into a psychopath and non-psychopath categories (Asghar et al. 2021).

4.4 Suicide

Suicide is the deliberate act of bringing about one's own death. When someone hurts oneself with the intent of ending their life, but does not die as a consequence of their acts, it is considered a suicide attempt. Numerous factors both enhance and decrease the chance of suicide. There is a link between suicide and other types of harm and violence. People who have experienced violence, such as child abuse, bullying, or sexual violence, for instance, are more likely to commit suicide. Suicidal thoughts and behaviors can be reduced by having simple access to

Table 4 Depression detection using traditional deep learning techniques on online social media

S. No.	Dataset	References	Type of mental disorder	DL technique	Evaluation metric	Year	Multi /binary class/cluster
1	Reddit (CLEF eRisk 2019)	(Shah et al. 2020)	Word2vec, FastText, Glove embeddings	Bi-LSTM	$F1$-score	531,453	Binary
2	Reddit (eRisk 2017)	(Trotzek et al. 2020)	Glove, FastText, Word-2vec	CNN	ERDE	49,557	Binary
3	Twitter	(Ghosh and Anwar 2021)	Polarity score, LIWC dictionary, emotional features, N-grams	LSTM, GRU, DNN	MSE, accuracy	4,245,747	Multi
4	Facebook	(Wu et al. 2020)	Emotional features, content features	Deep learning for depression detection for heterogeneous sources (D3-HDS)	ROC	873,524	Binary
5	Twitter (Bangla)	(Uddin et al. 2019)	Bag-of-words	GRU	Accuracy	5000	Binary
6	Reddit (CLEF eRisk 2019)	(Burdisso et al. 2021)	Sentiment score	RNN, LDA, deep averaging networks, contextualizers	$F1$-score	531,453	Binary
7	Reddit Reddit Self-Reported Depression Diagnosis (RSDD)	(Song et al. 2018)	Word frequency and polarity score	RNN, CNN	Precision, recall, $F1$-score	969	Binary
8	Twitter (Sentiment_140 and Sentiment_Tweets3)	(Gupta et al. 2022)	Word frequency	LR, DT, KNN, SVM, DT, LSTM	Accuracy, precision, recall, $F1$	1.6 million for 1st and 14,000 for 2nd dataset	Binary
9	Reddit (Depression_Mixed, DReddit, SAD)	(Yang et al. 2022a, b)	LIWC features, word embeddings	GRU	$F1$- score	2765, 3553, and 6850, respectively	Binary
10	Twitter (CLPSych 2015)	(Orabi et al. 2018)	Word embeddings	CNN, RNN, Bi-LSTM	Accuracy, $F1$-score, precision, recall, AUC	1,287,608	Binary
11	Twitter	(Amanat et al. 2022)	Principal component analysis, one-hot encoding	RNN, LSTM, no word embeddings	Accuracy, precision, recall, $F1$-score	Approx. 4000	Binary
12	Reddit	(Ahmed et al. 2021)	Glove embeddings	LSTM, Bi-LSTM, Bi-LSTM	Precision, accuracy, TP, and FPR	15,044	Multi-class
13	Reddit Reddit Self-Reported Depression Dataset (RSDD)	(Trifan and Oliveria 2021)	Psycholinguistic features	SVM, multinomial Naïve Bayes, CNN, FastText, transfer learning, no word embeddings	Accuracy, precision, recall, $F1$-score	892 and 78,639, respectively	Binary
14	Reddit	(Kim Nh et al. 2022)	Unicode, word2vec, BERT embeddings	NB, SVM, Bi-LSTM, RNN, CNN-1D, CNN-2D,LSTM, BERT	Precision, recall, $F1$-score	249,103	Binary and multi both
15	Reddit	(Cascalheira et al. 2022)	One-hot encoding	Bi-LSTM	Accuracy, $F1$-score	12,645	Binary

Table 4 (continued)

S. No.	Dataset	References	Type of mental disorder	DL technique	Evaluation metric	Year	Multi /binary class/cluster
16	Twitter	(Kour and Gupta 2022)	One-hot encoding	CNN-Bi-LSTM	Accuracy, precision, recall, F-measure, specificity, AUC curve	12 billion	Binary
17	Reddit (Depression Severity Dataset, eRisk Depression Severity Dataset)	(Naseem et al. 2022)	One-hot encoding	Ensemble of Bi-LSTM and BERT	Graded precision, graded recall, averaged *F*1-score	3553 and 77,350, respectively	Multi (4)
18	Twitter	(Cheng and Chen 2022)	Word2vec, Bert embeddings	Multimodal LSTM	Precision, recall, *F*1-score	44,390	Binary
19	Twitter	(Chen et al. 2018)	Temporal and non-temporal features, LIWC dictionary	SVM, RF	Accuracy, precision, recall, *F*1-measure	Average 2000 tweets per user	Binary
20	Twitter	(Gui et al. 2019)	Action, reward functions	CNN + RL, LSTM + RL	Accuracy, precision, recall, *F*1-score	4,245,747	Binary

health care, connections to family and community support, and several other factors. The suicide rate surged by 30% between 2000 and 2018, then fell in 2019 and 2020. With 45,979 deaths from suicide in 2020, it will be the third greatest cause of mortality in the US. Approximately one death occurs every 11 min. Even more, people contemplate suicide or make an attempt at it. According to estimates, 12.2 million American people considered suicide seriously in 2020, 3.2 million made plans to commit suicide, and 1.2 million actually succeeded in doing so (Facts about suicide 2023). There are multiple studies available for suicide detection using online social media. For ease of understanding, we have separated the studies of detecting suicide using machine and deep learning. Tables 9 and 10 focus on studies using supervised and unsupervised machine learning techniques, respectively, while Table 10 focuses on suicide detection on online social media using deep learning (Z Li2022).

Social media data analysis using machine learning offers a viable method for identifying long-term contextual factors that increase a person's risk of having suicidal thoughts and actions. The goal was to create the 'Suicide Artificial Intelligence Prediction Heuristic (SAIPH)' algorithm, which analyzes publically available Twitter data to forecast future risks of suicide thinking against psychological factors linked to suicide, such as burden, stress, loneliness, hopelessness, sleeplessness, sadness, and anxiety, the authors trained a number of neural networks, aid to help with suicide screening and risk monitoring, and have the ability to identify a person's future suicide ideation (SI) risk. The data are fed into random forest models to identify the individuals who are at a risk of suicide ideation (SI). An AUC of 88% was produced utilizing tweet data from at least 1 day before the case individuals' declaration of SI using a series of 10 bootstrap aggregated random forest models that included neural network model scores and a training and test set of roughly equal sizes. Importantly, the range of AUC values obtained using data from 3 weeks' worth of days within the 6 months prior to the SI event averaged an AUC of about 0.8, indicating that this method is effective for identifying people at risk for suicidal ideation who have not yet expressed such thoughts (Roy et al. 2020) (Table 11).

Artificial neural network (ANN) models were developed in this study to forecast suicide risk using the common language of social media users. The dataset contained legitimate psychological data about the users as well as 83,292 posts written by 1002 verified Facebook users. A single-task model (STM) to predict suicide risk from Facebook postings directly (Facebook texts—suicide) and a multi-task model (MTM), which included hierarchical, multilayered sets of theory-driven risk factors (Facebook texts—personality traits–psychosocial risks–psychiatric disorders–suicide),

Table 5 Depression detection using attention networks with transformers on online social media

S. No.	Dataset	References	Type of feature	DL technique	Evaluation metric	Number of posts	Multi /binary class/cluster
1	Reddit	(Poswiata and Perełkiewicz 2022)	BERT embeddings	BERT, RoBERT, XLNET, and an ensemble of these	Accuracy, precision, recall, $F1$-score	10,251	Multi (3)
2	Reddit, CLPsych 2015, eRisk	(Ansari et al. 2023)	PCA, sentiment score	Attention-based LSTM	Accuracy, precision, recall, $F1$-score	1841, 873, and 4498, respectively	Binary
3	Reddit	(Kayalvizhi et al. 2022)	Multiple techniques	Roberta	$F1$-score	16,632	Multi
4	Reddit	(William et al. 2022)	Word embeddings	Bert, Bert-summarized, XLNet, multichannel CNN, LSTM	Accuracy, precision, recall, $F1$	3412	Binary
5	Twitter (Arabic)	(Almars 2022)	Word2vec embeddings	Bi-LSTM with attention model and compared with state-of-the-art models	Accuracy	6000	Binary
6	Social media not specified	(Esackimuthu et al. 2022)	Bert embeddings	ALBERT-Base	Accuracy, precision, recall, $F1$-measure	16,632	Multi
7	Reddit	(Ren et al. 2021)	LIWC dictionary, unigrams, LDA topics	Emotion-based attention network (EAN), LDA, LSTM, Bi-LSTM with LIWC, Unigram, bigram	Accuracy, precision, recall, $F1$-score	1842	Binary
8	Reddit	(Malviya et al. 2021)	TF-IDF, Word2vec embeddings	Transformers with Bert-base-uncased, DistillBert-base-uncased, Roberta-base, Electra-base, Xlnet-base-uncased	Accuracy, $F1$-score	5000	Binary
9	Twitter	(Cui B et al. 2022)	Word2vec embeddings	Emotion-based reinforcement attention network (ERAN) compared with multiple ML, DL, and transformer approach	Accuracy, precision, recall, $F1$-score	500,000	Binary
10	Twitter	(Zogan et al. 2022)	Sentiment score, LDA topics, domain-specific features, social information	Multi-aspect depression detection with hierarchical attention (MDHAN), hierarchical attention network (HAN)	Accuracy, precision, recall, $F1$-score	Approx. 2 million	Binary
11	Social media not specified	(Ahmed et al. 2021)	Glove embeddings	Bidirectional LSTM with attention	AUC, precision	15,044	Multi

Table 6 Schizophrenia detection using machine learning techniques on online social media

S. No.	Dataset	Reference	Type of feature	ML technique	Evaluation metric	Number of posts	Multi /binary class/cluster
1	Reddit	(Bae et al. 2021)	LIWC dictionary, LDA topics	RF, SVM, LR, NB	Recall, precision, accuracy, $F1$-score, AUC	247,569	Binary

Table 7 Psychopath detection using machine learning technique on online social media

S. No.	Dataset	References	Type of feature	ML technique	Evaluation metric	Number of posts	Multi /binary class/cluster
1	Twitter	(Tadisetty, and Ghazinour 2021)	N-grams, TF-IDF	NB, SVM, KNN with N-grams	Accuracy	600,000	Binary

Table 8 Psychopath detection using deep learning technique on online social media

S. No.	Dataset	References	Type of feature	DL technique	Evaluation metric	Number of posts	Multi /binary class/cluster
1	Twitter	(Asghar et al. 2021)	Bag-of-words, TF-IDF	Bi-LSTM and compared with other state-of-art ML and DL models	Accuracy, precision, recall, $F1$-score	601	Binary
2	Twitter	(Alotaibi et al. 2021)	Bag-of-words, TF-IDF	CNN-LSTM	Precision, recall, $F1$-measure	601	Binary

were both built using deep contextualized word embeddings for text representation. The MTM achieved much better prediction accuracy when compared to the STM predictions (0.697 AUC 0.746), with significantly bigger effect sizes (0.729 d 0.936) (Ophir et al. 2020).

4.5 Stress

Any form of change that causes physical, mental, or psychological strain on a person is considered to be stressful. Your body's reaction to anything that demands attention or action is stress. Everyone experiences moments of stress. However, how you tackle stress has a significant impact on your overall well-being (Stress 2022). Stress is a state of tension, either emotionally or physically. Any circumstance or idea that gives you cause for annoyance, rage, or anxiety can trigger it. The body's response to a demand or challenge is stress. Stress can occasionally be advantageous, such as when it keeps a person safe or helps accomplish a deadline. However, chronic stress can be bad for a person's health (Li N, 2023).

We discovered two studies focusing on stress detection on online social media using deep learning and a study on machine learning which are presented in Tables 12 and 13, respectively.

Text analysis has been demonstrated to be useful in the diagnosis of mental illness, emotions, and sentiment. However, the current state of the art in stress identification from text is corpus-specific. Good, well-validated approaches that work across several datasets are currently in short supply. Munaz et al. proposed a method to detect stress in textual data and evaluate it using numerous publicly available English datasets, with the goal of advancing the state of the art in this area. In order to improve classification accuracy, the proposed method blends lexicon-based characteristics with distributional representations. Additionally, three distinct word embedding methods for making use of distributional representation are investigated. Three machine learning models were used to implement this method, and their performance was assessed by $F1$-score. This analysis serves as a starting point for future studies, and the results show that the best model, with $F1$-score over 80%, combines FastText embeddings with a subset of lexicon-based features (Munaz and Iglesias, 2022) as mentioned in Table 12.

4.6 Borderline personality and bipolar disorder

Borderline personality disorder affects how you perceive and feel about yourself and other people, making it difficult to function in daily life. Issues with one's self-image, trouble

Table 9 Suicide detection using supervised machine learning techniques on online social media

S. No.	Dataset	References	Type of feature	ML technique	Evaluation metric	Number of posts	Multi /binary class/cluster
1	Twitter	(Chadha and Kaushik 2021)	Bag-of-words	NB, SVM, Bernoulli NB, DT, LR, voting ensemble, RF, AdaBoost ensemble	Accuracy, precision, recall	14,202	Binary
2	Reddit	(Lao C et al. 2022)	LIWC dictionary, TF-IDF	Gradient boost, SVM, RF	$F1$-score, AUROC	1732	Binary
3	Twitter	(Roy et al. 2020)	One-hot encoding	Suicide artificial intelligence prediction heuristic (SAIPH) algorithm	AUC	7,223,922	Binary
4	Reddit	(Acuna Caicedo et al. 2022)	Bag-of-words, TF-IDF, word embeddings	SVM	Macro precision, recall	273	Multi
5	Twitter	(Chatterjee et al. 2022)	N-grams, TF-IDF, sentiment polarity, LDA topics, social features, temporal features	LR, RF, SVM, XGBoost using TF-IDF, N-grams	Accuracy, recall, precision, $F1$-measure	188,704	Binary
6	Twitter (CLPsych 2021)	(Wang N et al. 2021)	Latent features, POS tagging,	C-Attention, linear discriminant analysis, KNN, SVM, used POS tagging	$F1$-score, TPR, FPR, AUC	Not specified	Binary

Table 10 Suicide detection using unsupervised machine learning technique on online social media

S. No.	Dataset	References	Type of feature	ML technique	Evaluation metric	Number of posts	Multi /binary class/cluster
1	Reddit	(Feldhege J et al. 2022)	LIWC dictionary	LDA	Mean, median	7995	Cluster

controlling one's emotions and conduct, and a history of rocky relationships are all included (Cleveland Clinic 2022).

A brain illness called bipolar disorder alters a person's energy, mood, and capability to work. Bipolar disorder patients go through severe emotional crises, or mood episodes, that normally last a few days to a few weeks (Bipolar Disorder, 2022).

We found two studies focusing on borderline and bipolar personality disorders using machine learning only, that are presented in Table 14. Sekulic et al. performed a binary classification task on the social media posts of Reddit for detecting bipolar disorder by gathering interesting linguistic distinctions between individuals with bipolar disorders and the control group revealed by feature analysis, including variations in the use of words that indicate emotions. Recognizing that emotional ups and downs are the primary symptoms of the disorder in the textual clues, the authors examined the emotion-expressive textual qualities in individuals with bipolar disorder and the non-bipolar control group of users using SVM, RF and evaluated them on the different metrics (Sekulic et al. 2018).

Table 11 Suicide detection using deep learning technique on online social media

S. No.	Dataset	References	Type of feature	DL technique	Evaluation metric	Number of posts	Multi /binary class/cluster
1	Seina Weibo	(Li Z et al. 2022)	Sentence-level feature	Deep hierarchical ensemble model for suicide detection (DHE-SD)	Accuracy, $F1$-score	164,856	Binary
2	Facebook	(Ophir et al. 2020)	ELMO embeddings	ANN	AUC	83,292	Binary
3	Twitter	(Sawhney et al. 2021)	TF-IDF, BERT embeddings	CNN-LSTM, CNN, RF	Recall, accuracy, $F1$-score	34,306	Binary
4	Reddit	(Sawhney and Joshi 2021)	Glove, BERT embeddings	SVM with radial basis function, SVM with linear kernel, RF, MLP, contextual-CNN, suicide detection model	Precision, recall, $F1$-score	3894	Multi
5	Seina Weibo	(Li and Zhou et al. 2021)	BERT embeddings	FastText, DPCNN, TextCNN	Accuracy, precision, recall, $F1$-score	452,508	Binary
6	Reddit	(Kodati and Tene 2022)	BERT embeddings, lexicon-based contextual information	RNN, Bi-LSTM, GRU-Bi-LSTM-CNN	Accuracy, precision, recall, $F1$-score, MSE	6820	Binary
7	Seina Weibo	(Li et al. 2022)	Frequency count	DPCNN, FastText, TextCNN	Accuracy, precision, recall, $F1$-score	563,336	Binary

Table 12 Stress detection using machine learning technique on online social media

S. No.	Dataset	References	Type of feature	ML technique	Evaluation metric	Number of posts	Multi /binary class/cluster
1	Reddit, Twitter (DReddit, The Natural Stress Emotion, and TensiStrength)	(Muñoz and Iglesias 2022)	LIWC stress dictionary	SVM, logistic regression, SGD	$F1$-score	2294, 2243, and 6142, respectively	Binary

Table 13 Stress detection using deep learning techniques on online social media

S. No.	Dataset	References	Type of feature	DL technique	Evaluation metric	Number of posts	Multi /binary class/cluster
1	Seina weibo	(Li, Zhang and Fang 2023)	Bert embeddings	LSTM, RNN, GRU	Accuracy, precision, recall, $F1$-score	524,944	Binary
2	Twitter	(Wang et al. 2022)	Bert embeddings	Meta-learning-based stress category detection framework (SCD)	Accuracy, precision, recall, $F1$-score	1553	Binary

4.7 Self-harm

Self-harm refers to the act of intentionally inflicting pain on oneself as a means of coping with intense emotions, distressing memories, or overwhelming circumstances and situations (Self-harm 2022).

Table 14 Borderline personality disorder and bipolar disorder detection using machine learning techniques on online social media

S. No.	Dataset	References	Type of feature	ML and DL technique	Evaluation metric	Number of posts	Multi /binary class/cluster
1	Reddit	(Deb et al. 2022)	Bag-of-words	RF, extra trees classifier, bagging, XGB, DT, K-neighbors	Accuracy, $F1$-score	2504	Binary
2	Reddit	(Sekulic et al. 2018)	LIWC dictionary	SVM, LR, RF	Accuracy, $F1$-score	Approx. 3 million	Binary

Table 15 Self-harm detection using attention networks on online social media

S. No.	Dataset	References	Type of feature	ML and DL technique	Evaluation metric	Number of posts	Multi /binary class/cluster
1	Reddit Suicidality Dataset	(Abed Esfahani et al. 2019)	LIWC dictionary	GPT-1	Precision	768	Binary
2	Reddit (eRisk 2019)	(Maupome et al. 2020)	Topic modeling, one-hot encoding	LDA, neural encoders	Precision, recall, $F1$-score	Not specified	Binary

Table 16 Anxiety detection using machine learning technique on online social media

S. No.	Dataset	References	Type of feature	ML and DL technique	Evaluation metric	Number of posts	Multi /binary class/cluster
1	Facebook	(Chang and Tseng 2020)	Text-based features, social features, para-social relationship	One-class SVM	Precision, recall, $F1$-score	Not specified	Binary

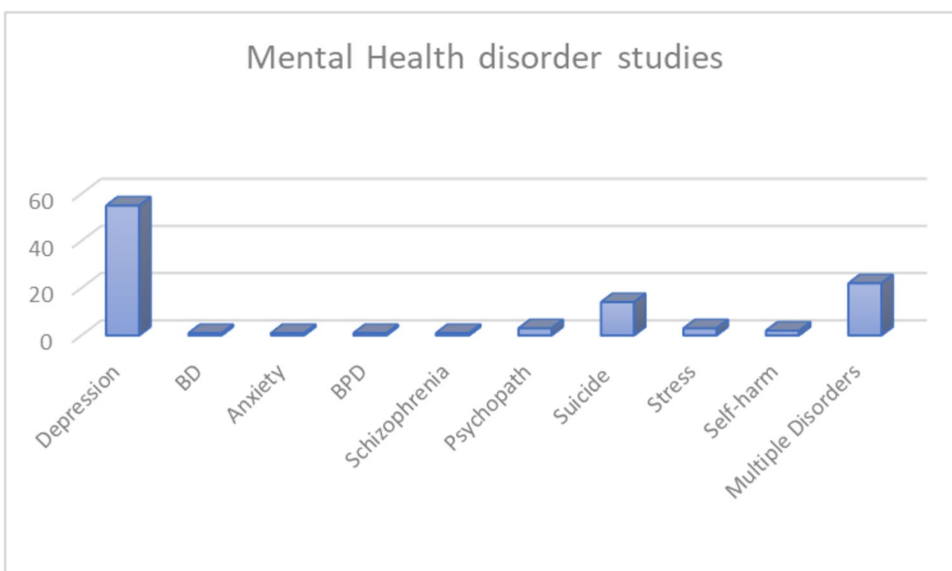


Fig. 4 Total mental health disorder studies in our review

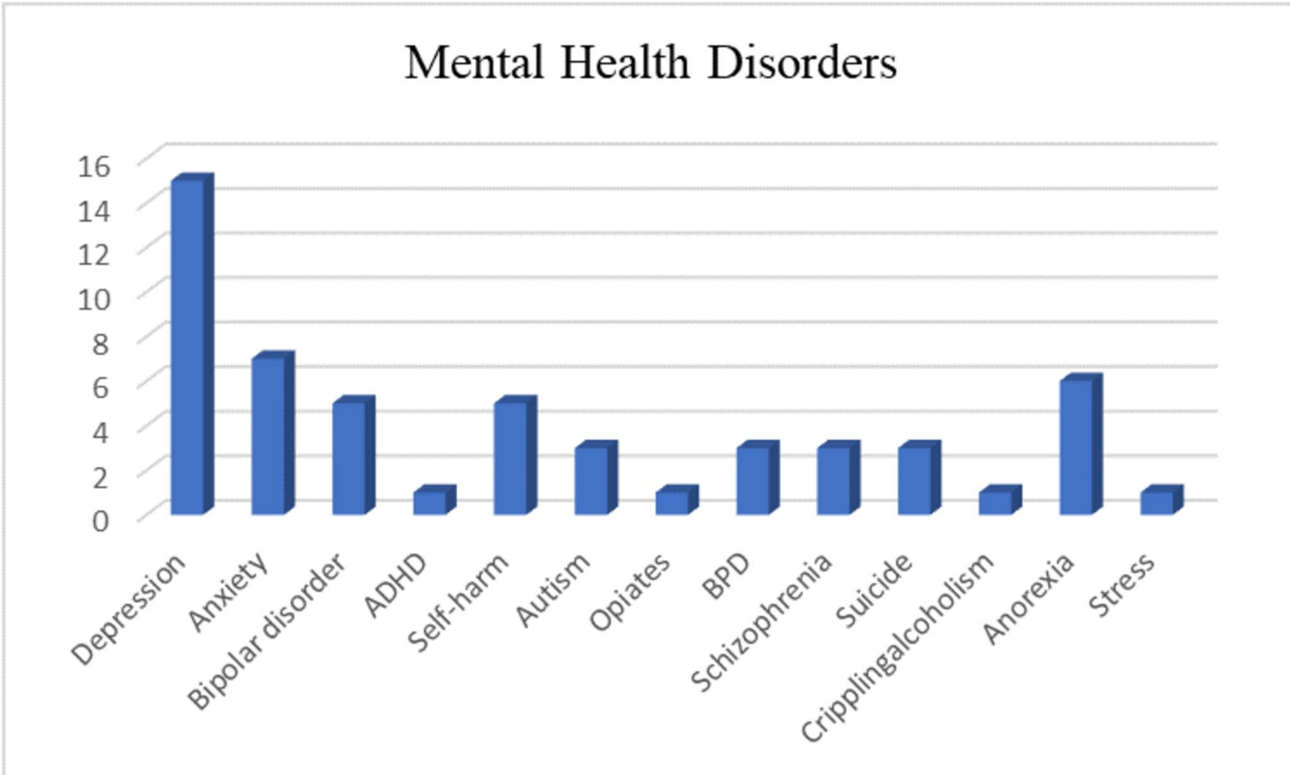


Fig. 5 Total studies in the multiple disorders section

Usually, it is not intended as a suicide attempt. It is an unhealthy approach to deal with emotional anguish, despair, rage, and stress to damage oneself. While self-harm may temporarily relieve physical and emotional stress and create a sense of serenity, it is typically followed by remorse and humiliation as well as the recurrence of unpleasant emotions. Usually, life-threatening injuries are not planned, but more serious and even deadly self-harm is a possibility. Abed Esfahani et al. used transfer learning to detect the early signs of self-harm on the Reddit social media platform and gave up different evaluation metrics (Abed Esfahani et al., 2019).

During our research, we could only get two studies that are presented in Table 15.

4.8 Anxiety

Anxiety is characterized by apprehension, worry, and restlessness. It may trigger flushing of the skin, agitation, and a racing heart in certain people. It is not unusual for stress to bring up such a response. You may experience anxiety when you have to solve a challenging problem at work, when you have to take a test, or when you have to make a big decision. It may serve as a coping mechanism. It is possible that the feeling of unease will really serve to increase your stamina and mental clarity. The terror is permanent and overwhelming for those who suffer from anxiety disorders (Anxiety, 2020). Table 16 presents studies based on anxiety using online social media.

Figure 4 shows us the total mental health disorder studies in our review. As can be seen clearly, most studies focus on depression alone, followed by suicide. In this chart, we have only focused on single studies. In the next chart, we further elaborate on the studies that focused on more than two studies at a time, as shown in Fig. 5.

4.9 Multiple disorders

The above studies presented were targeted at only single mental health disorders. In this section, we have shown the studies which identified multiple disorders using online social media. Some of the mental health disorders have been discussed previously, while some others are not.

Anorexia is an eating disorder characterized by a persistent and severe underweight status relative to age and

Table 17 Multiple disorders detection using machine learning technique on online social media

S. No.	Dataset	References	Type of mental disorder	Type of feature	ML technique	Evaluation metric	Number of posts	Multi /binary class/cluster
1	Reddit	(Jain et al. 2022)	Suicide, depression	Bag-of-words	Naïve Bayes, logistic regression, SVM, random forest, no word embeddings	Precision, recall, $F1$-score	300,000	Binary
2	Reddit	(Tariq et al. 2019)	Anxiety, depression, bipolar, ADHD	TF-IDF	RF, NB, SVM	Precision, recall, $F1$-score	3922	Binary
3	Reddit	(Thorstad and Wolff 2019)	ADHD, bipolar, anxiety, depression	TF-IDF	L2-penalized logistic regression	Accuracy, precision, recall, $F1$-score	224,036	Binary
4	Reddit	(Ragheb et al. 2019)	Anorexia, self-harm, depression	Bag-of-words	Deep mood evaluation module, Bayesian variational inference	Precision, recall, $F1$-measure	Not specified	Multi
5	Reddit	(Guo et al. 2021)	Bipolar disorder, anxiety, major depressive disorder	Bert embeddings	SVM, LR, RF	Accuracy, $F1$-score, precision, recall	686,359, 686,369, 914,082	Multi

height. Even if they are already underweight, people with this disease may experience a crippling fear of gaining weight. They could be using unhealthy methods to shed pounds, such as excessive dieting or exercise (Anorexia Nervosa 2018).

Experiencing or seeing a scary event might set off the mental health disease known as post-traumatic stress disorder (PTSD). In addition to compulsive, intrusive thoughts about the traumatic incident, symptoms can include flashbacks, nightmares, and anxiety attacks. People who have experienced trauma often have brief periods of difficulty adjusting and coping, but with time and self-care, they typically recover. There is a possibility of post-traumatic stress disorder (PTSD) if the symptoms worsen over time (months or years) and cause significant disruption to daily life (PTSD 2022).

Inattention/hyperactivity disorder (ADHD) is among the most prevalent conditions of childhood neurodevelopment. As a rule, it is first identified in young people and continues throughout adulthood for many (Christiansen et al. 2020). Children with attention-deficit hyperactivity disorder (ADHD) may have difficulties focusing, may act without considering the potential consequences of their actions, or maybe extremely energetic (CDC 2021).

Tables 17, 18, and 19 present the studies focusing on multiple disorder identification using machine and deep learning on online social media, respectively.

In the current study, the authors examined whether there is enough signal in people's ordinary language to forecast the future. Language samples were gathered from posts to discussion groups concentrating on various mental illnesses (clinical subreddits) as well as postings to discussion groups focusing on nonmental health subjects (nonclinical subreddits) on the social media platform Reddit. Finally, models trained to predict future mental illness learned to focus on words indicating life stress while models trained on clinical subreddits learned to focus on words indicating disorder-specific symptoms, suggesting that the features that are predictive of mental illness may change over time. Classification models were evaluated using precision, recall, and accuracy, and clustering was performed to identify the most used words in these subreddits (Thorstad and Wolff 2019).

Zeberga et al. proposed a novel methodology for identifying depression and anxiety-related posts using bidirectional encoder representations from transformers (BERT) while preserving the contextual and semantic meaning of the words used across the entire corpus. To further improve performance and accuracy, a knowledge distillation methodology, a relatively new method for transferring information learned by a big pre-trained model (BERT) to a smaller one, was introduced. The authors developed a methodology for gathering information from the most popular social media platforms, such as Reddit and Twitter. Finally, word2vec and BERT with Bi-LSTM are used to analyze social media posts for indicators of depression and anxiety. When compared to other state-of-the-art technologies, the system achieved 98% accuracy employing the knowledge distillation process (Zeberga et al. 2022).

Table 18 Multiple disorders detection using deep learning technique on online social media

S. No.	Dataset	References	Type of mental disorder	Type of feature	DL technique	Evaluation metric	Number of posts	Multi /binary class/cluster
1	Facebook (Bengali)	(Kabir et al. 2022)	Depression, schizophrenia	Bag-of-words, TF-IDF	SVM, RF, KNN, LR, BiGRU, LSTM, Bi-LSTM all experimented along with attention also	Accuracy, precision, recall, $F1$-score	5000	Binary
2	Reddit (Self-reported Mental Health Diagnosis Dataset) 7 datasets	(Borba De Souza et al. 2022)	Depression, anxiety	Glove embeddings	LSTM, CNN, DAC stacking	Precision, recall, $F1$-score	698,694, 1,406,530, 581,948, 431,386, 388,297, 398,081, 589,137	Binary
3	Twitter	(Carmel Mary Belinda et al. 2022)	Depression, anxiety	TF-IDF	Multinomial NB, CNN, LSTM	Accuracy, $F1$-score	670	Binary
4	Twitter, Reddit	(Zeberga et al. 2022)	Anxiety, depression	FastText, Word2vec, BERT embeddings	CNN, LSTM, Bi-LSTM with FastText, word2vec, BERT	Precision, recall, accuracy	100,000 and 95,000, respectively	Binary
5	Reddit (UMD suicidality dataset, SWMH), Twitter	(Ji et al. 2021)	Suicide, depression, anxiety, bipolar	Word embeddings, emotion polarity	Bi-LSTM, RN (relation networks) and compared with FastText, CNN, LSTM, RCNN, SSA	Accuracy, precision, recall, $F1$-score	866, 54,412, 4800	Multi for suicide, binary for others
6	Reddit (eRisk 2018, 2019)	(Aragón et al. 2020)	Anorexia, depression	Bag-of-words, N-grams, Glove, and Word2vec	Deep emotion detection model	Accuracy, $F1$-score, precision, recall	1287 and 1707, respectively	Binary
7	Reddit	(Kim J et al. 2021)	Depression, bpd, bipolar, autism, anxiety, schizophrenia		CNN, XGBoost	$F1$-score, accuracy, precision, recall		Binary/ multi
8	Reddit (eRisk 2020)	(Uban and Rosso 2020)	Depression, self-harm	Glove embeddings, LIWC dictionary, emotion sentiment score	Bi-LSTM with attention	$F1$	50	Binary
9	Reddit	(Trotzek et al. 2018)	Depression, anorexia	LIWC features, part-of-speech tagging (POS), Glove, FastText embeddings	CNN	$F1$	Not specified properly	Binary
10	Reddit	(Wang et al. 2018)	Bipolar disorder and depression	TF-IDF, sentence embedding	CNN	Mean score	1707, 472 submissions per subject	Binary
11	Reddit (eRisk depression, self-harm, anorexia)	(Uban et al. 2021)	Depression, anorexia, self-harm	LIWC dictionary, content, sentiment, emotion features	Bi-LSTM with attention, hierarchical network	$F1$, AUC	811,586, 823,754, and 274,534, respectively	Binary

Table 18 (continued)

S. No.	Dataset	References	Type of mental disorder	Type of feature	DL technique	Evaluation metric	Number of posts	Multi /binary class/cluster
12	Reddit	(Ive et al. 2018)	Depression, bipolar disorder, autism, opiates, BPD, schizophrenia, anxiety, self-harm, suicide watch, crippling alcoholism	One-hot vector encoding	GRU	Precision, recall, F-measure	538,272	Binary, multi
13	Reddit (eRisk 2019, 2020)	(Aragón et al. 2023)	Anorexia, depression	TF-IDF, Chi-square distribution	Bi-LSTM, CNN bag-of-words, unigrams, bigrams, LIWC, Glove, and word2vec	Accuracy, $F1$-score, precision, recall	1951.2 average number of posts for anorexia 2138.4 average number of posts for depression	Binary
14	Reddit	(Aragón et al. 2022)	Anorexia, depression, self-harm	Bag-of-words, N-grams	CNN	Accuracy, precision, recall	1951.2, 2138.4, 1113.8 average number of posts	Binary

In this section, more than one disorder is used in the social media posts for the classification or clustering task. In this chart also, the majority of the studies are focused on depression, followed by anxiety, anorexia, self-harm, suicide, and autism, etc. This chart represents the total studies in the multiple disorders section alone as shown in Fig. 5.

After, reviewing all the studies, we have tried to present the total number studies which we found in different years. As we can clearly see from the chart in Fig. 6, that the studies on mental are maximum in the year 2022 according to our review. There is a sudden rise in the number of studies after 2020 may be because of the sudden rise in mental cases because of the consequences of lockdown in different parts of the world.

5 Limitations of using online social media

It can be observed that promising results are obtained while using online social media for mental health disorder detection using ML and DL techniques. But there are various challenges with the use of social media data that need to be resolved. Some of the key challenges are as follows:

5.1 Faulty research ethics

There are no ethical guidelines available for addressing the unique challenges of working with the data on social media. Because of this, scientists and academicians have little direction in addressing the ethical, legal, and social issues of social media data. The ethical issues encountered while using the social media data are not addressed in the guidelines given by these federal institutes, as the scientists and the review board members are unaware of all the ethical, social, and legal implications of using social media data. The users who have made their profiles and data public need not require any ethical clearances, but they may not be aware of the way their data are being used, and their identities can be traced back, which could cause them harm. Most of the time, the user agreements made are very lengthy and detailed, which addresses all the complications and guidelines for users. Still, users do not read them fully and generally agree to them without realizing how their data will be used. IBM's Watson for Oncology kept Watson's unsafe and incorrect suggestions secret for more than a year for cancer patients as the system was trained on a few synthetic cancer cases from the Memorial Sloan Kettering (MSK) Cancer Center by its doctors (Chen 2018). There are some academic groups available, like ReCODE Health, which enable resource sharing and take the responsibility of educating the general public about their rights with respect to social media.

Table 19 Multiple disorders detection using attention networks on online social media

S. No.	Dataset	References	Type of mental disorder	Type of features	DL technique	Evaluation metric	Number of posts	Multi /binary class/cluster
1	Reddit	(Dinu and Moldovan et al. 2021)	ADHD, anxiety, autism, bipolar, borderline, depression, OCD, PTSD, schizophrenia	LIWC dictionary	BERT, RoBERTa, XLNET	Precision, recall, *F*1-score	356,358	Binary
2	Reddit	(Amini et al. 2020)	Depression, anorexia	N-grams	CNN-ELMo	Accuracy	5998	Binary
3	Twitter (CLPsych 2017)	(Howard et al. 2020)	Suicide, self-harm	LIWC dictionary	GPT-1	*F*1-score	157,963	Binary

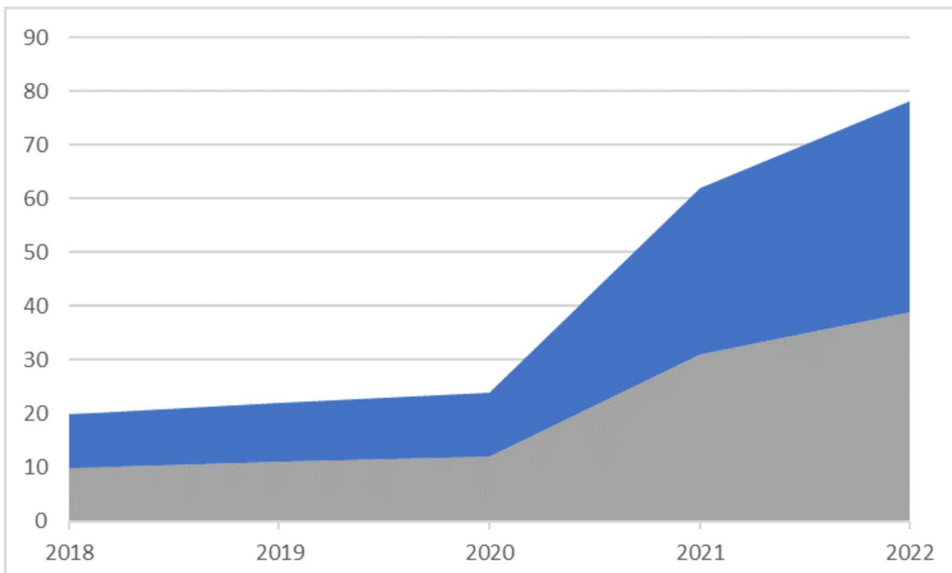


Fig. 6 Total number of studies covered under our review study in different publication years using stacked area chart

Researchers should notify about the process used for data collection, storage criteria, and how and what information they have used (Brown 2018).

5.2 Approach to social media data

Retrieving social media is restricted which includes publicly available data because of the continuous change in the data access rules. Previously, a whistleblower revealed the privacy breaches in Facebook and Instagram, which made these giants change their application programming interfaces (APIs) to limit data access by third-party applications. Confining the data access appears to be a convincing means but this has a great impact on how the data can be accessed favorably by the general public. This continual change in the data access policies forces researchers to acquire data from third-party applications by paying exorbitant prices. These applications add layers of privacy breaches. There have been multiple reports suggesting the compromise of research when the data access policies changed during a research study pressurizing the researchers to redevelop the methods and techniques mid-way, ultimately degrading the quality of research standards. The Twitter Academic Research API was made for research purposes and was free for academicians, but when Elon Musk acquired Twitter, he made the API paid on February 27, 2023, which posed many problems for the researchers who were in the middle of their data collection process (Developer terms, 2023). Recently, Facebook has partnered with Social Science One, which has facilitated the relationship between industry and academia to solve the challenges encountered while accessing the data, such as privacy, content, and trade secrets. This venture has allowed researchers to use Facebook data for how it influences democracy. Facebook partnered with Social Science One to allow scientists to access Facebook data to study how social media influences democracy.

5.3 Absence of collaborative academic research

Looking at the proliferation of social media, interdisciplinary fields should be developed to improve scientific research to bring the different communities simultaneously. The researchers in different academic fields are unaware of different works in the cross-referenced academic fields like diabetes prediction using machine learning is the work of medicine as well as engineering. Still, there is little awareness among researchers to collaborate with each other so that they come up with better results. This lack of merging of different scientific studies disseminated the studies across different journals and conferences, which led to researchers not getting enough of what could be implemented by having an interdisciplinary approach. Scientific networking programs and transdisciplinary training will help researchers flourish in their fields. Health training programs can offer courses on machine learning, natural language processing, data analytics, and health-care analytics which will allow clinicians and computer engineers to learn and understand the basic concepts needed to implement transdisciplinary research.

5.4 Availability and correctness of social media data

One major problem encountered while using online social media for mental health prediction is accessing correct labeled data. Although collecting social media data is easy as compared to surveys and questionnaires, getting appropriate data is difficult because of different user writing styles. Manual labeling of social media data is very time-consuming and has to be performed by professionals with proper consensus. Nowadays, social media giants like Facebook have increased measures to protect the privacy of users by making all the content of the profile private. These privacy features hamper the data collection process as such content won't be accessed by the APIs while extracting the online social media data (Jain 2020).

6 Conclusion

The main purpose of this study is to provide a rundown of the state-of-the-art research on machine and deep learning approaches used for predicting multiple mental health disorders using online social media data. Moreover, the review can be helpful for researchers in formulating models based on the severity levels of mental health disorders in real-time for the users on online social media. Although the reliability of online social media is not guaranteed, the data should be properly analyzed and the privacy of the users should not be compromised. Even though online social media comes with methodological and technical difficulties for predictive modeling, social media is still a valuable source for identifying the characteristics of individuals vulnerable to mental health disorders. Moreover, the COVID-19 pandemic has also worsened the mental health of individuals, and social media has become a savior for them (Khasnis et al. 2021). Despite various difficulties on online social media, this field helps in providing rapid tools and techniques to mitigate future risks. Furthermore, the review focused on the ideas adopted by different researchers by providing a summary of the dataset, techniques applied, evaluation metrics used, type of features used, and the type of classification or clustering applied. However, social media text is comprised of varying writing styles and unstructured text, but natural language processing and machine learning techniques are widely used for the analysis of such text. On the basis of this review work, we suggest that natural language processing and machine learning approaches could be the vehicle for translating big online social media data into improved human health. Furthermore, we also focused on the limitations encountered while using online social media for research purposes. As a major concern, this study also outlines the research gap while reviewing the papers. The research gap should be overcome to bring up more studies in the future. A very interesting fact to notice is that the research related to mental health has doubled in the year 2020–2022. We could gather a lot of research published in these years. The major reason for this rise could be the COVID-19 pandemic which affected individuals mentally and physically. The lockdown put a mental strain on the majority of the individuals because of loneliness, loss of loved ones, job loss, and several other reasons. The lockdowns also increased the use of online social media, letting users share their emotions on these platforms.

7 Future work

Data collection and analysis tools such as sentiment analysis and opinion mining have been widely available due to recent technological developments. As it stands, reliance on technology has advantages and disadvantages. The entire information sector would benefit from an internet-based framework that can be accessed from anywhere by anyone looking for reliable data and therapeutic guidelines on mental health. The research proposal is an interdisciplinary exploration of how digital natives' mental health affects their online behaviors. Thus, it pulls from psychology and computer science to investigate how far open-source content from online social media may be mined for clues toward the early diagnosis of mental illness in individuals.

Academics can analyze the static data on a forum to determine if it contains depressed material, according to recent studies. To collect more dynamic data in diagnosing mental illnesses, however, it is now time to combine forum users with conversational AI. Because people are more likely to open up and share personal details about their experiences with mental health when they know they are being anonymous online, it is important to consider how technology and mental health intersect when trying to understand the data. While it is not the same as having a trained professional on your side, having a strong social network behind you is a great first line of defense. All these things might be the foundation for learning more about mental health on online social media. People in today's fast-paced world use the internet to research their mental, emotional, and social health to gain an edge in the workplace. Finding big amounts of textual data in the medical field can be challenging, but it is important to study how to extract web data related to human psychology. Because of the progress made in ML, deep learning approaches can now deal with text corpora of different sizes throughout time. That is why it is important to look into using deep learning to evaluate textual features in real-time. To learn from an interactive environment, a deep reinforcement learning agent can try out different forms of feedback based on their own prior experiences. Therefore, better results can be produced by combining reinforcement learning with deep learning to identify the severity of mental conditions based on social media posts. Transfer learning and attention networks can also help analyze these patterns. Most of the research we came across only targeted the English language, but in a country like India, where multiple languages are spoken along with English, these data could not completely determine populations like this. So multilingual and bilingual areas for this research should be explored by the researchers as many online social media are available in multiple languages. The researchers can also focus on categorizing the mental health status of individuals depending on factors such as geographical area, profession, etc., as we could not discover any paper which has classified the mental health state of individuals focusing on such factors. These factors could provide us a better insight into, let's say, which profession is mostly affected by mental health. This research could help psychologists find new areas of study for why a particular group is affected and how to tackle these people.

Author contributions A.Khan wrote and prepared the main manuscript. R. Ali reviewed the manuscript critically.

Data availability No datasets were generated or analysed during the current study.

Declarations

Competing interests The authors declare no competing interests.

References

Abed-Esfahani P, Howard D, Maslej MM, Patel S, Mann V, Goegan S, & French L (2019) Transfer learning for depression: early detection and severity prediction from social media postings. In: Conference and labs of the evaluation forum

Acuña Caicedo RW, Gómez Soriano JM, Melgar Sasieta HA (2022) Bootstrapping semi-supervised annotation method for potential suicidal messages. Internet Interv 28:100519. https://doi.org/10.1016/j.invent.2022.100519

Ahmed U, Mukhiya SK, Srivastava G, Lamo Y, Lin JC-W (2021) Attention-based deep entropy active learning using lexical algorithm for mental health treatment. Front Psychol. https://doi.org/10.3389/fpsyg.2021.642347

Alexopoulos GS (2019) Mechanisms and treatment of late-life depression. Transl Psychiatry 9:188. https://doi.org/10.1038/s41398-019-0514-6

Almars MA (2022) Attention-based Bi-LSTM model for arabic depression classification. Comput Mater Continua 71:3091–3106. https://doi.org/10.32604/cmc.2022.022609

Almouzini S, Alageel A (2019) Detecting arabic depressed users from twitter data. Procedia Comput Sci 1(163):257–265. https://doi.org/10.1016/j.procs.2019.12.107

Alotaibi FM, Asghar MZ, Ahmad S (2021) A hybrid CNN-LSTM model for psychopathic class detection from tweeter users. Cogn Comput 13:709–723. https://doi.org/10.1007/s12559-021-09836-7

Alsayat A (2022) Improving sentiment analysis for social media applications using an ensemble deep learning language model. Arab J Sci Eng 47(2):2499–2511. https://doi.org/10.1007/s13369-021-06227-w

Amanat A, Rizwan M, Javed AR, Abdelhaq M, Alsaqour R, Pandya S, Uddin M (2022) Deep learning for depression detection from textual data. Electronics 11:676. https://doi.org/10.3390/electronics11050676

Amini H, Kosseim L (2020) Towards explainability in using deep learning for the detection of anorexia in social media. Nat Lang Process Inform Syst. https://doi.org/10.1007/978-3-030-51310-8_21

Andy A (2021) Understanding user communication around loneliness on online forums. PLoS ONE. https://doi.org/10.1371/journal.pone.0257791

Angskun J, Tipprasert S, Angskun T (2022) Big data analytics on social networks for real-time depression detection. J Big Data. https://doi.org/10.1186/s40537-022-00622-2

Anorexia nervosa—symptoms and causes. (2018) In: Mayo clinic. https://www.mayoclinic.org/diseases-conditions/anorexia-nervosa/symptoms-causes/syc-20353591 (accessed on Nov 2022)

Ansari L, Ji S, Chen Q, Cambria E (2023) Ensemble hybrid learning methods for automated depression detection. IEEE Trans Comput Soc Syst 10:211–219. https://doi.org/10.1109/tcss.2022.3154442

Anxiety: MedlinePlus (2020) In: anxiety: medlineplus. https://medlineplus.gov/anxiety.html (accessed on Nov 2022)

Aragón ME, López-Monroy AP, González LC, Montes-y-Gómez M (2020) Attention to emotions: detecting mental disorders in social media. Text, Speech Dialogue. https://doi.org/10.1007/978-3-030-58323-1_25

Aragón ME, López-Monroy AP, González LC, Montes-y-Gómez M (2022) Approaching what and how people with mental disorders communicate in social media–Introducing a multi-channel representation. Neural Comput Appl 34:20149–20164. https://doi.org/10.1007/s00521-022-07569-8

Aragón ME, López-Monroy AP, González-Gurrola LC, Montes-y-Gómez M (2023) Detecting mental disorders in social media through emotional patterns—the case of anorexia and depression.

IEEE Trans Affect Comput 14:211–222. https://doi.org/10.1109/taffc.2021.3075638

ArgyrisY A, Zhang N, Bashyal B and Tan PN (2022) Using deep learning to identify linguistic features that facilitate or inhibit the propagation of anti- and pro-vaccine content on social media. In: 2022 IEEE International conference on digital health (ICDH). 107–116 https://doi.org/10.1109/ICDH55609.2022.00025

Asghar J, Akbar S, Asghar MZ, Ahmad B, Al-Rakhami MS, Gumaei A (2021) Detection and classification of psychopathic personality trait from social media text using deep learning model. Comput Math Methods Med 2021:1–10. https://doi.org/10.1155/2021/5512241

Azam N, Ahmad T, Ul-Haq N (2021) Automatic emotion recognition in healthcare data using supervised machine learning. PeerJ Comput Sci 7:e751. https://doi.org/10.7717/peerj-cs.751

Bae YJ, Shim M, Lee WH (2021) Schizophrenia detection using machine learning approach from social media content. Sensors. https://doi.org/10.3390/s21175924

Bathina KC, ten Thij M, Lorenzo-Luaces L et al (2021) Individuals with depression express more distorted thinking on social media. Nat Hum Behav 5:458–466. https://doi.org/10.1038/s41562-021-01050-7

Beheshti A, Hashemi VM, Yakhchi S, Nezhad HRM, Ghafari SM, and Yang J (2020) Personality2vec: enabling the analysis of behavioral disorders in social networks. In: Proceedings of the 13th international conference on web search and data mining. Association for computing machinery. 825–828. https://doi.org/10.1145/3336191.3371865

Belinda CM, Ravikumar MJ, Muhammad S, Dhilip Kumar A, Antony V, Kumar K, Arulkumaran G (2022) Linguistic analysis of hindi-english mixed tweets for depression detection. J Math. https://doi.org/10.1155/2022/3225920

Bipolar disorder—symptoms and causes (2022) In: Mayo clinic. https://www.mayoclinic.org/diseases-conditions/bipolar-disorder/symptoms-causes/syc-20355955 (accessed on Nov 2022)

Borba de Souza V, Campos Nobre J, Becker K (2022) DAC stacking: a deep learning ensemble to classify anxiety, depression, and their comorbidity from Reddit texts. IEEE J Biomed Health Inform 26:3303–3311. https://doi.org/10.1109/jbhi.2022.3151589

Briand A, Almeida H, Meurs MJ (2018) Analysis of social media posts for early detection of mental health conditions. Adv Artif Intell. https://doi.org/10.1007/978-3-319-89656-4_11

Brown J (2018) IBM Watson reportedly recommended cancer treatments that were ‘unsafe and incorrect’. Gizmodo, https://gizmodo.com/ibm-watson-reportedly-recommended-cancer-treatments-tha-1827868882. (accessed on Oct 2021)

Burdisso SG, Errecalde M, Montes-y-Gómez M (2021) Using text classification to estimate the depression level of Reddit users. J Comput Sci Technol 10(24215/16666038):21.e1

CDC (2021) What is ADHD? In: Centers for disease control and prevention. https://www.cdc.gov/ncbddd/adhd/facts.html (accessed on Nov 2022)

Cacheda F, Fernandez D, Novoa FJ, Carneiro V (2019) Early detection of depression: social network analysis and random forest techniques. J Med Internet Res 21:e12554. https://doi.org/10.2196/12554

Cascalheira CJ, Hamdi SM, Scheer JR, Saha K, Boubrahimi SF, Choudhury MD (2022) Classifying minority stress disclosure on social media with bidirectional long short-term memory. Proc Int AAAI Conf Web Soc Med 16:1373–1377. https://doi.org/10.1609/icwsm.v16i1.19390

Chadha A, Kaushik B (2021) Machine learning based dataset for finding suicidal ideation on Twitter. In: Third international conference on intelligent communication technologies and virtual mobile networks (ICICV). 823–828, https://doi.org/10.1109/ICICV50876.2021.9388638

Chancellor S, De Choudhury M (2020) Methods in predictive techniques for mental health status on social media: a critical review. NPJ Digit Med 3:43. https://doi.org/10.1038/s41746-020-0233-7

Chandra Guntuku S, Buffone A, Jaidka K, Eichstaedt JC, Ungar LH (2019) Understanding and measuring psychological stress using social media. Proc Int AAAI Conf Web Soc Med 13:214–225. https://doi.org/10.1609/icwsm.v13i01.3223

Chang MY and Tseng CY (2020) Detecting social anxiety with online social network data. In: 21st IEEE International conference on mobile data management (MDM). 333–336, https://doi/https://doi.org/10.1109/MDM48529.2020.00073

Chatterjee R, Gupta RK, Gupta B (2021) Depression detection from social media posts using multinomial naive theorem. IOP Conf Ser Mater Sci Eng 1022:012095. https://doi.org/10.1088/1757-899x/1022/1/012095

Chatterjee M, Kumar P, Samanta P, Sarkar D (2022) Suicide ideation detection from online social media: a multi-modal feature based technique. Int J Inform Manag Data Insights 2:100103. https://doi.org/10.1016/j.jjimei.2022.100103

Chen X, Sykora MD, Jackson TW, Elayan S (2018) What about mood swings. Identifying depression on twitter with temporal measures of emotions. In: Companion of the the web conference 2018 on the web conference 2018–WWW ’18. https://doi.org/10.1145/3184558.3191624

Chen A (2018).IBM’s Watson gave unsafe recommendations for treating cancer. In: IBM’s Watson gave unsafe recommendations for treating cancer—The verge. https://www.theverge.com/2018/7/26/17619382/ibms-watson-cancer-ai-healthcare-science (accessed on Nov 2022)

Cheng JC, Chen ALP (2022) Multimodal time-aware attention networks for depression detection. J Intell Inf Syst. https://doi.org/10.1007/s10844-022-00704-w

Chiong R, Budhi GS, Dhakal S (2021) Combining sentiment lexicons and content-based features for depression detection. IEEE Intell Syst 36(6):99–105. https://doi.org/10.1109/MIS.2021.3093660

Christiansen H, Chavanon ML, Hirsch O et al (2020) Use of machine learning to classify adult ADHD and other conditions based on the Conners’ adult ADHD rating scales. Sci Rep 10:18871. https://doi.org/10.1038/s41598-020-75868-y

Coello-Guilarte L, Ortega-Mendoza RM, Villaseñor-Pineda L, & Montes-y-Gómez M (2019) Crosslingual depression detection in Twitter using bilingual word alignments. In: Conference and labs of the evaluation forum

Cohen AB, Stump L, Krumholz HM et al (2022) Aligning mission to digital health strategy in academic medical centers. Npj Digit Med 5:67. https://doi.org/10.1038/s41746-022-00608-7

Cui B, Wang J, Lin H, Zhang Y, Yang L, Xu B (2022) Emotion-based reinforcement attention network for depression detection on social media: algorithm development and validation. JMIR Med Inform 10:e37818. https://doi.org/10.2196/37818

Cleveland Clinic Borderline Personality Disorder: Causes, Symptoms & Treatment., 2022. In: Cleveland Clinic. https://my.clevelandclinic.org/health/diseases/9762-borderline-personality-disorder-bpd. Accessed on Nov 2022

Deb K, De H, Chatterjee SS and Pal A (2022) Studying borderline personality disorder using machine learning. In: 16th International conference on ubiquitous information management and communication (IMCOM), 1–6. https://doi.org/10.1109/IMCOM53663.2022.9721800

Deng L, Rattadilok P, and Xiong R (2021) A machine learning-based monitoring system for attention and stress detection for children with autism spectrum disorders. In: 3rd International conference on intelligent medicine and health (ICIMH 2021). Association for computing machinery. 23–29. https://doi.org/10.1145/3484377.3484381

Depression (major depressive disorder)—symptoms and causes (2022) In: Mayo clinic. https://www.mayoclinic.org/diseases-conditions/depression/symptoms-causes/syc-20356007

Depression (2021) In: depression. https://www.who.int/india/health-topics/depression (accessed on Nov 2022)

Diniz EJS, Fontenele JE, de Oliveira AC, Bastos VH, Teixeira S, Rabêlo RL, Calçada DB, Dos Santos RM, de Oliveira AK, Teles ASB (2022) A natural language processing-based digital phenotyping tool for smart monitoring of suicidal ideation. Healthcare 10(4):698. https://doi.org/10.3390/healthcare10040698

Dinu A and Moldovan AC (2021) Automatic detection and classification of mental illnesses from general social media texts. In Proceedings of the international conference on recent advances in natural language processing. pp. 358–366

Dolezal L, Gibson M (2022) Beyond a trauma-informed approach and towards shame-sensitive practice. Humanit Soc Sci Commun 9:214. https://doi.org/10.1057/s41599-022-01227-z

Durstewitz D, Koppe G, Meyer-Lindenberg A (2019) Deep neural networks in psychiatry. Mol Psychiatry 24:1583–1598. https://doi.org/10.1038/s41380-019-0365-9

Esackimuthu S, Hariprasad S, Sivanaiah R, Angel S, Rajendram SM, and Mirnalinee TT (2022) SSN_MLRG3 @LT-EDI-ACL2022-depression detection system from social media text using transformer models. In: Proceedings of the second workshop on language technology for equality, diversity and inclusion. Association for computational linguistics. 196–199

Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K, Cui C, Corrado G, Thrun S, Dean J (2019) A guide to deep learning in healthcare. Nat Med 25(1):24–29. https://doi.org/10.1038/s41591-018-0316-z

Ríssola EA, Losada DE, Crestani F (2021) A survey of computational methods for online mental state assessment on social media. ACM Trans Comput Healthcare 2(2):1–31. https://doi.org/10.1145/3437259

Facts about suicide | suicide | CDC (2023) In: facts about suicide | suicide | CDC. https://www.cdc.gov/suicide/facts/index.html (accessed on Mar 2023)

Fatima I, Abbasi BUD, Khan S, Al-Saeed M, Ahmad HF, Mumtaz R (2019) Prediction of postpartum depression using machine learning techniques from social media text. Expert Syst. https://doi.org/10.1111/exsy.12409

Feldhege J, Wolf M, Moessner M, Bauer S (2022) Psycholinguistic changes in the communication of adolescent users in a suicidal ideation online community during the COVID-19 pandemic. Eur ChildAdolesc Psychiatry. https://doi.org/10.1007/s00787-022-02067-7

Fraga BS, da Silva AC, Murai F (2018) Online social networks in health care: a study of mental disorders on Reddit. IEEE/WIC/ACM Int Conf Web Intell (WI) 2018:568–573

Ghosh S, Anwar T (2021) depression intensity estimation via social media: a deep learning approach. IEEE Trans Comput Soc Syst 8:1465–1474. https://doi.org/10.1109/tcss.2021.3084154

Govindasamy KA and Palanichamy N (2021) Depression detection using machine learning techniques on twitter data. In: 5th International conference on intelligent computing and control systems (ICICCS). 960–966, https://doi.org/10.1109/ICICCS51141.2021.9432203

Gui T, Zhang Q, Zhu L, Zhou X, Peng M, Huang X (2019) Depression detection on social media with reinforcement learning. Lect Notes Comput Sci. https://doi.org/10.1007/978-3-030-32381-3_49

Guo X, Sun Y, Vosoughi S (2021) Emotion-based modeling of mental disorders on social media. IEEE/WIC/ACM Int Conf Web Intell. https://doi.org/10.1145/3486622.3493916

Gupta S, Goel L, Singh A, Prasad A, Ullah MA (2022) Psychological analysis for depression detection from social networking sites. Comput Intell Neurosci 2022:1–14. https://doi.org/10.1155/2022/4395358

Haque F, Nur R U, Jahan S A, Mahmud Z and Shah FM (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). 1–5. https://doi.org/10.1109/ICCIT51783.2020.9392692

Health matters: reducing health inequalities in mental illness. In: GOV.UK. https://www.gov.uk/government/publications/health-matters-reducing-health-inequalities-in-mental-illness/health-matters-reducing-health-inequalities-in-mental-illness (accessed on Nov 2022)

Howard D, Maslej MM, Lee J, Ritchie J, Woollard G, French L (2020) Transfer learning for risk classification of social media posts: model evaluation study. J Med Internet Res 22:e15371. https://doi.org/10.2196/15371

Hu N (2022) Sentiment analysis of texts on public health emergencies based on social media data mining. Comput Math Methods Med. https://doi.org/10.1155/2022/3964473

Islam MDR, Kabir MA, Ahmed A, Kamal ARM, Wang H, Ulhaq A (2018) Depression detection from social network data using machine learning techniques. Health Inform Sci Syst. https://doi.org/10.1007/s13755-018-0046-0

Ive J, Gkotsis G, Dutta R, Stewart R, Velupillai S (2018) Hierarchical neural model with attention mechanisms for the classification of social media text related to mental health. In: Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic, pp 69–77. New Orleans, LA. Association for Computational Linguistics.

Le Glaz A, Haralambous Y, Kim-Dufor DH, Lenca P, Billot R, Ryan TC, Marsh J, Devylder J, Walter M, Berrouiguet S, Lemey C (2021) Machine learning and natural language processing in mental health: systematic review. J Med Internet Res 23(5):e15708

Jain P, Ram Srinivas K, Vichare A (2022) Depression and suicide analysis using machine learning and NLP. J Phys Conf Ser 2161:012034. https://doi.org/10.1088/1742-6596/2161/1/012034

Jain R (2020) How to lock Facebook profile step by step. In: Business insider. https://www.businessinsider.in/tech/how-to/how-to-lock-facebook-profile/articleshow/75864853.cms (accessed on Mar 2023)

Ji S, Li X, Huang Z, Cambria E (2021) Suicidal ideation and mental disorder detection with attentive relation networks. Neural Comput Appl 34:10309–10319. https://doi.org/10.1007/s00521-021-06208-y

Kabir MK, Islam M, Kabir ANB, Haque A, Rhaman MK (2022) Detection of depression severity using Bengali social media posts on mental health: study using natural language processing techniques. JMIR Formative Res 6:e36118. https://doi.org/10.2196/36118

Kamite SR and Kamble VB (2020) Detection of depression in social media via Twitter using machine learning approach. In: International conference on smart innovations in design, environment, management, planning and computing. pp. 122–125, https://doi/https://doi.org/10.1109/ICSIDEMPC49020.2020.9299641

Kang K, Yoon C and Kim EY (2016) Identifying depressive users in Twitter using multimodal analysis. In: International conference on big data and smart computing (BigComp). pp. 231–238. https://doi.org/10.1109/BIGCOMP.2016.7425918

Karcher NR, Barch DM (2021) The ABCD study: understanding the development of risk for mental and physical health outcomes. Neuropsychopharmacol 46:131–142. https://doi.org/10.1038/s41386-020-0736-6

Katchapakirin K, Wongpatikaseree K, Yomaboot P and Kaewpitakkun K (2018) Facebook social media for depression detection in the Thai community. In: 15th International joint conference on computer science and software engineering (JCSSE).1–6, https://doi.org/:https://doi.org/10.1109/JCSSE.2018.8457362

Kayalvizhi S, Durairaj T, Chakravarthi BR, and Mahibha JC (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. Association for computational linguistics. 331–338

Khasnis NS, Sen S, and Khasnis SS (2021) A machine learning approach for sentiment analysis to nurture mental health amidst COVID-19. In: Proceedings of the international conference on data science, machine learning and artificial intelligence. Associationfor computing machinery. 284–289. https://doi.org/10.1145/3484824.3484877

Kim NH, Kim JM, Park DM, Ji SR, Kim JW (2022) Analysis of depression in social media texts through the Patient Health Questionnaire-9 and natural language processing. Digit Health 8:205520762211142. https://doi.org/10.1177/20552076221114204

Kim J, Lee D, Park E (2021) Machine learning for mental health in social media: bibliometric study. J Med Internet Res 23:e24870. https://doi.org/10.2196/24870

Kiong YT (2022) An initial study of depression detection on mandarin textual through BERT model. In: 14th ACM Web science conference 2022 (WebSci '22). Association for computing machinery. 459–463. https://doi.org/10.1145/3501247.3539015

Kodati D, Tene R (2022) Identifying suicidal emotions on social media through transformer-based deep learning. Appl Intell 53:11885–11917. https://doi.org/10.1007/s10489-022-04060-8

Kour H, Gupta MK (2022) A hybrid deep learning approach for depression prediction from user tweets using feature-rich CNN and bi-directional LSTM. Multimed Tools Appl 81(17):23649–23685. https://doi.org/10.1007/s11042-022-12648-y

Kumar P, Samanta P, Dutta S, Chatterjee M, Sarkar D (2022) Feature based depression detection from twitter data using machine learning techniques. J Sci Res 66:220–228. https://doi.org/10.37398/jsr.2022.660229

Lao C, Lane J, Suominen H (2022) Analyzing suicide risk from linguistic features in social media: evaluation study. JMIR Formative Res 6:e35563. https://doi.org/10.2196/35563

Lau KM, Hou WK, Hall BJ, Canetti D, Ng SM, Lam AI, Hobfoll SE (2016) Social media and mental health in democracy movement in Hong Kong: a population-based study. Comput Human Behav 1(64):656–662. https://doi.org/10.1016/j.chb.2016.07.028

Li N, Zhang H, Feng L (2023) Incorporating forthcoming events and personality traits in social media based stress prediction. IEEE Trans Affect Comput 14:603–621. https://doi.org/10.1109/taffc.2021.3076294

Li Z, Zhou J, An Z, Cheng W, Hu B (2022b) Deep hierarchical ensemble model for suicide detection on imbalanced social media data. Entropy 24:442. https://doi.org/10.3390/e24040442

Liu D, Feng XL, Ahmed F, Shahid M, Guo J (2022) Detecting and measuring depression on social media using a machine learning approach: systematic review. JMIR Mental Health 9:e27244. https://doi.org/10.2196/27244

Liu J, Shi M (2022) What are the characteristics of user texts and behaviors in Chinese depression posts? Int J Environ Res Public Health 19:6129. https://doi.org/10.3390/ijerph19106129

Li Z, An Z, Cheng W, Zhou J, Zheng F, Hu B (2022) MHA: a multimodal hierarchical attention model for depression detection in social media. Health Inf Sci Syst 11(1):6. https://doi.org/10.1007/s13755-022-00197-5

Lyu Y, Chow JC, Hwang JJ, Li Z, Ren C, Xie J (2022) Psychological well-being of left-behind children in China: text mining of the social media website Zhihu. Int J Environ Res Public Health 19(4):2127. https://doi.org/10.3390/ijerph19042127

Malviya K, Roy B and Saritha S (2021) A transformers approach to detect depression in social media. In: International conference on artificial intelligence and smart systems (ICAIS), 718–723, https://doi.org/10.1109/ICAIS50930.2021.9395943

Marsch LA (2021) Digital health data-driven approaches to understand human behavior. Neuropsychopharmacol 46:191–196. https://doi.org/10.1038/s41386-020-0761-5

Maupomé D, Armstrong MD, Belbahar RM, Alezot J, Balassiano R, Queudot M, Mosser S, & Meurs M (2020) Early mental health risk assessment through writing styles, topics and neural models. In: Conference and labs of the evaluation forum

Maxim S, Ivan S, Natalia K, Anastasia U (2020) Depression detection from social media profiles. https://doi.org/10.1007/978-3-030-51913-1_12

McAuliffe C, Slemon A, Goodyear T, McGuinness L, Shaffer E, Jenkins EK (2022) Connectedness in the time of COVID-19: Reddit as a source of support for coping with suicidal thinking. SSM Qual Res Health 2:100062. https://doi.org/10.1016/j.ssmqr.2022.100062

Mendu S, Baglione A, Baee S, Wu C, Ng B, Shaked A, Clore G, Boukhechba M, Barnes L (2020) A framework for understanding the relationship between social media discourse and mental health. Proc ACM Hum-Comput Interact. https://doi.org/10.1145/3415215

Miotto R, Wang F, Wang S, Jiang X, Dudley JT (2017) Deep learning for healthcare: review, opportunities and challenges. Brief Bioinform 19:1236–1246. https://doi.org/10.1093/bib/bbx044

Muñoz S, Iglesias CA (2022) A text classification approach to detect psychological stress combining a lexicon-based feature framework with distributional representations. Inf Process Manag 59:103011. https://doi.org/10.1016/j.ipm.2022.103011

Myszczynska MA, Ojamies PN, Lacoste AMB et al (2020) Applications of machine learning to diagnosis and treatment of neurodegenerative diseases. Nat Rev Neurol 16:440–456. https://doi.org/10.1038/s41582-020-0377-8

Nanomi Arachchige IA, Sandanapitchai P, Weerasinghe R (2021) Investigating machine learning & natural language processing techniques applied for predicting depression disorder from online support forums: a systematic literature review. Information 12(11):444. https://doi.org/10.3390/info12110444

Naseem U, Dunn AG, Kim J, and Khushi M (2022) Early identification of depression severity levels on Reddit using ordinal classification. In: Proceedings of the ACM web conference. (WWW39;22). Association for computing machinery. 2563–2572. https://doi.org/10.1145/3485447.3512128

Non-commercial use of the Twitter API, 2023. In: Non-commercial use of the Twitter API | Twitter Developer Platform. https://developer.twitter.com/en/developer-terms/commercial-terms. Accessed on Mar 2023

Ophir Y, Tikochinski R, Asterhan CSC, Sisso I, Reichart R (2020) Deep neural networks detect suicide risk from textual Facebook posts. Sci Rep. https://doi.org/10.1038/s41598-020-73917-0

Orabi AH, Buddhitha P, Orabi MH, & Inkpen D (2018) Deep learning for depression detection of Twitter users. CLPsych@NAACL-HTL

Ortega-Mendoza RM, Hernández-Farías DI, Montes-y-Gómez M, Villaseñor-Pineda L (2022) Revealing traces of depression through personal statements analysis in social media. Artif Intell Med 123:102202. https://doi.org/10.1016/j.artmed.2021.102202

Post-traumatic stress disorder (PTSD)—symptoms and causes (2022) In: Mayo clinic. https://www.mayoclinic.org/diseases-conditions/post-traumatic-stress-disorder/symptoms-causes/syc-20355967 (accessed on Nov 2022)

Poświata R and Perełkiewicz M (2022) Detecting signs of depression from social media text using RoBERTa pre-trained language models. In: Proceedings of the second workshop on language technology for equality, diversity and inclusion. Association for computational linguistics. 276–282

Prakash A, Agarwal K, Shekhar S, Mutreja T, & Chakraborty P (2021) An ensemble learning approach for the detection of depression and mental illness over Twitter data. In: 2021 8th International conference on computing for sustainable global development (INDIACom), 565–570

Psychopathy (2023) In: Psychology today. https://www.psychologytoday.com/us/basics/psychopathy (accessed on Mar 2023)

Ptaszynski M, Zasko-Zielinska M, Marcinczuk M, Leliwa G, Fortuna M, Soliwoda K, Dziublewska I, Hubert O, Skrzek P, Piesiewicz J, Karbowsk P, Dowgiallo M, Eronen J, Tempska P, Brochocki M, Godny M, Wroczynski M (2021) Looking for Razors and Needles in a Haystack: multifaceted analysis of suicidal declarations on social media-A pragmalinguistic approach. Int J Environ Res Public Health 18(22):11759. https://doi.org/10.3390/ijerph182211759

Ragheb W, Azé J, Bringay S, & Servajean M (2019) Attentive multi-stage learning for early risk detection of signs of anorexia and self-harm on social media. In: Conference and labs of the evaluation forum

Rahmani AM, Yousefpoor E, Yousefpoor MS, Mehmood Z, Haider A, Hosseinzadeh M, Ali Naqvi R (2021) Machine learning (ML) in medicine: review, applications, and challenges. Mathematics 9:2970. https://doi.org/10.3390/math9222970

Ranjana R, Subha T, Sreenidhi KL and Akshaya K (2022) Mental health prediction using natural language processing". In: 2022 International conference on applied artificial intelligence and computing (ICAAIC) .824–828, https://doi.org/10.1109/ICAAIC53929.2022.9792893

Ren L, Lin H, Xu B, Zhang S, Yang L, Sun S (2021) Depression detection on Reddit with an emotion-based attention network: algorithm development and validation. JMIR Med Inform 9:e28754. https://doi.org/10.2196/28754

Rezaii N, Walker E, Wolff P (2019) A machine learning approach to predicting psychosis using semantic density and latent content analysis. Npj Schizophr. https://doi.org/10.1038/s41537-019-0077-9

Ricard BJ, Marsch LA, Crosier B, Hassanpour S (2018) Exploring the utility of community-generated social media content for detecting depression: an analytical study on Instagram. J Med Internet Res 20:e11817. https://doi.org/10.2196/11817

Roy A, Nikolitch K, McGinn R, Jinah S, Klement W, Kaminsky ZA (2020) A machine learning approach predicts future risk to suicidal ideation from social media data. Npj Digit Med. https://doi.org/10.1038/s41746-020-0287-6

Rutowski T, Shriberg E, Harati A, Lu Y, Chlebek P and Oliveira R (2020) Depression and anxiety prediction using deep language models and transfer learning. In: 7th International conference on behavioural and social computing, 1–6, https://doi.org/10.1109/BESC51023.2020.9348290

Safa R, Bayat P, Moghtader L (2021) Automatic detection of depression symptoms in twitter using multimodal analysis. J Supercomput 78:4709–4744. https://doi.org/10.1007/s11227-021-04040-8

Saha K, Choudhury MD (2021) Assessing the mental health of college students by leveraging social media data. XRDS 28:54–58. https://doi.org/10.1145/3481834

Sakib AS, Mukta MS, Huda FR, Islam AN, Islam T, Ali ME (2021) Identifying Insomnia from social media posts: psycholinguistic analyses of user tweets. J Med Internet Res 23(12):e27613. https://doi.org/10.2196/27613

Salas-Zárate R, Alor-Hernández G, Salas-Zárate M, Del P, Paredes-Valverde MA, Bustos-López M, Sánchez-Cervantes JL (2022) Detecting depression signs on social media: a systematic literature review. Healthcare 10:291. https://doi.org/10.3390/healthcare10020291

Saravanan T, Jhaideep T and Bindu NH (2022) Detecting depression using hybrid models created using Google's BERT and Facebook's fast text algorithms. In: 2nd International conference on advance computing and innovative technologies in engineering (ICACITE). 415–421, https://doi.org/10.1109/ICACITE53722.2022.9823581

Sarkar D, Kumar P, Samanta P, Dutta S, Chatterjee M (2022) A two-level multi-modal analysis for depression detection from online social media. Int J Softw Innov 10:1–22. https://doi.org/10.4018/ijsi.309114

Sawhney R, & Joshi H (2021) PHASE: learning emotional phase-aware representations for suicide ideation detection on social media. In: Conference of the European chapter of the association for computational linguistics

Sawhney R, Joshi H, Shah RR, and Flek L (2021) Suicide ideation detection via social and temporal user representations using hyperbolic learning. In: Proceedings of the 2021 conference of the north american chapter of the association for computational linguistics: human language technologies. Online. Association for computational linguistics. 2176–2190

Schizophrenia - Symptoms and causes. (2020) In: Mayo clinic. https://www.mayoclinic.org/diseases-conditions/schizophrenia/symptoms-causes/syc-20354443 (accessed on Nov 2022)

Sekulic I, Gjurkovic M, Šnajder J (2018) Not just depressed: bipolar disorder prediction on Reddit. In: Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pp 72–78. Brussels, Belgium, Association for Computational Linguistics.

Self-harm (2022) In: self-harm—causes, warning signs and symptoms and when to seek help | healthdirect. https://www.healthdirect.gov.au/self-harm (accessed on Nov 2022)

Shah FM et al. (2020) Early depression detection from social network using deep learning techniques. In: IEEE region 10 symposium (TENSYMP) 823–826, https://doi.org/10.1109/TENSYMP50017.2020.9231008

Shen G, Jia J, Nie L, Feng F, Zhang C, Hu T, Chua T, & Zhu W (2017) Depression detection via harvesting social media: a multimodal dictionary learning solution. In: International joint conference on artificial intelligence

Shigemura J, Ursano RJ, Morganstein JC, Kurosawa M, Benedek DM (2020) Public responses to the novel 2019 coronavirus (2019-nCoV) in Japan: mental health consequences and target populations. Psychiatry Clin Neurosci 74(4):281–282. https://doi.org/10.1111/pcn.12988

Shrestha A, Serra E, Spezzano F (2020) Multi-modal social and psycho-linguistic embedding via recurrent neural networks to identify depressed users in online forums. Netw Model Anal Health Inform Bioinforma 9:22. https://doi.org/10.1007/s13721-020-0226-0

Shrestha A, Spezzano F (2019) Detecting depressed users in online forums. In: Proceedings of the 2019 IEEE/ACM international conference on advances in social networks analysis and mining. https://doi.org/10.1145/3341161.3343511

Sik D, Németh R, Katona E (2021) Topic modelling online depression forums: beyond narratives of self-objectification and self-blaming. J Ment Health 32:386–395. https://doi.org/10.1080/09638237.2021.1979493

Skaik R, Inkpen D (2020) Using social media for mental health surveillance: a review. ACM Comput Surv 53(6):129–131. https://doi.org/10.1145/3422824

Skaik R and Inkpen D (2021) Using Twitter social media for depression detection in the Canadian population. In: Proceedings of the 2020 3rd artificial intelligence and cloud computing conference

(AICCC '20). Association for computing machinery.109–114. https://doi.org/10.1145/3442536.3442553

Song H, You J, Chung JW, and Park JC (2018) Feature attention network: interpretable depression detection from social media. In: Proceedings of the 32nd Pacific Asia conference on language, information and computation, Hong Kong. Association for computational linguistics

Stress (2022) In: stress. https://www.who.int/news-room/questions-and-answers/item/stress#:~:text=Stress%20can%20be%20defined%20as,experiences%20stress%20to%20some%20degree. (accessed on Nov 2022)

Su C, Xu Z, Pathak J, Wang F (2020) Deep learning in mental health outcome research: a scoping review. Transl Psychiatry. https://doi.org/10.1038/s41398-020-0780-3

Sun Y, Gao D, Shen X, Li M, Nan J, Zhang W (2022) Multi-Label classification in patient-doctor dialogues with the RoBERTa-WWM-ext + CNN (robustly optimized bidirectional encoder representations from transformers pretraining approach with whole word masking extended combining a convolutional neural network) model: named entity study. JMIR Med Inform 10(4):e35606. https://doi.org/10.2196/35606

Tadesse MM, Lin H, Xu B, Yang L (2019) Detection of depression-related posts in Reddit social media forum. IEEE Access 7:44883–44893. https://doi.org/10.1109/access.2019.2909180

Tadisetty S and Ghazinour K (2021) Anonymous prediction of mental illness in social media. In: IEEE 11th annual computing and communication workshop and conference (CCWC). 0954–0960, https://doi.org/10.1109/CCWC51732.2021.9376140

Tariq S, Akhtar N, Afzal H, Khalid S, Mufti MR, Hussain S, Habib A, Ahmad G (2019) A novel co-training-based approach for the classification of mental illnesses using social media posts. IEEE Access 7:166165–166172. https://doi.org/10.1109/access.2019.2953087

The latest Facebook statistics: everything you need to know datareportal—Global digital insights (2023) In: DataReportal—Global Digital Insights. https://datareportal.com/essential-facebook-stats

Thij M, Bathina K, Rutter LA et al (2020) Depression alters the circadian pattern of online activity. Sci Rep 10:17272. https://doi.org/10.1038/s41598-020-74314-3

Thorstad R, Wolff P (2019) Predicting future mental illness from social media: a big-data approach. Behav Res Methods 51:1586–1600. https://doi.org/10.3758/s13428-019-01235-z

Tong L, Liu Z, Jiang Z, Zhou F, Chen L, Lyu J, Zhang X, Zhang Q, Sadka A, Wang Y, Li L, Zhou H (2022) Cost-sensitive boosting pruning trees for depression detection on Twitter. IEEE Trans Affective Comput. https://doi.org/10.1109/taffc.2022.3145634

Topic: Twitter. (2022) In: Statista. https://www.statista.com/topics/737/twitter/ (accessed on Nov 2022)

Trifan A, Oliveira JL (2021) Cross-evaluation of social mining for classification of depressed online personas. J Integr Bioinform 18:101–110. https://doi.org/10.1515/jib-2020-0051

Trifan A, Antunes R, Matos S, Oliveira JL (2020) Understanding depression from psycholinguistic patterns in social media texts. Advances in information retrieval. ECIR. Lecture Notes in Computer Science.12036. Springer, Cham. https://doi.org/10.1007/978-3-030-45442-5_50

Trotzek M, Koitka S, Friedrich CM (2020) Utilizing neural networks and linguistic metadata for early detection of depression indications in text sequences. IEEE Trans Knowl Data Eng 32(3):588–601. https://doi.org/10.1109/tkde.2018.2885515

Trotzek M, Koitka S, and Friedrich CM (2018) Word embeddings and linguistic metadata at the CLEF 2018 tasks for early detection of depression and anorexia. CLEF (Working Notes)

Uban A-S, Chulvi B, Rosso P (2021) An emotion and cognitive based analysis of mental health disorders from social media data. Futur Gener Comput Syst 124:480–494. https://doi.org/10.1016/j.future.2021.05.032

Uban A, Rosso P (2020) Deep learning architectures and strategies for early detection of self- harm and depression level prediction. In: CEUR workshop proceedings. 2696:1–12

Uddin AH, Bapery D, & Mohammad Arif AS (2019) Depression analysis of bangla social media data using gated recurrent neural network. In: 2019 1st International conference on advances in science, engineering and robotics technology (ICASERT), 1–6

Wang J, Zhou Y, Zhang W, Evans R, Zhu C (2020b) Concerns expressed by chinese social media users during the COVID-19 pandemic: content analysis of Sina Weibo microblogging data. J Med Internet Res 22(11):e22152. https://doi.org/10.2196/22152

Wang Y, Huang H, & Chen H (2018) A neural network approach to early risk detection of depression and anorexia on social media text. In: Conference and labs of the evaluation forum

Wang Y, Wang Z, Li C, Zhang Y, & Wang H (2020) A multitask deep learning approach for user depression detection on Sina Weibo. ArXiv, abs/2008.11708

Wang N, Luo F, Shivtare Y, Badal VD, Subbalakshmi K, Chandramouli R, & Lee E (2021) Learning models for suicide prediction from social media posts. ArXiv, abs/2105.03315

Wang X, Cao L, Zhang H, Feng L, Ding Y, and Li N (2022) A meta-learning based stress category detection framework on social media. In: Proceedings of the ACM web conference 2022 (WWW '22). Association for computing machinery. 2925–2935. https://doi.org/10.1145/3485447.3512013

Westrupp EM, Greenwood CJ, Fuller-Tyszkiewicz M, Berkowitz TS, Hagg L, Youssef G (2022) Text mining of Reddit posts: using latent Dirichlet allocation to identify common parenting issues. PLoS ONE 17(2):e0262529. https://doi.org/10.1371/journal.pone.0262529

William D, Achmad S Suhartono D and Gema AP (2022) Leveraging BERT with extractive summarization for depression detection on social media. In: International seminar on intelligent technology and its applications 63–68, https://doi.org/10.1109/ISITIA56226.2022.9855370

Wongkoblap A, Vadillo MA, Curcin V (2018) A multilevel predictive model for detecting social network users with depression. IEEE Int Conf Healthc Inform (ICHI) 2018:130–135

Wongkoblap A, Vadillo MA, Curcin V (2021) Deep learning with anaphora resolution for the detection of tweeters with depression: algorithm development and validation study. JMIR Mental Health 8:e19824. https://doi.org/10.2196/19824

Wu MY, Shen C-Y, Wang ET, Chen ALP (2020) A deep architecture for depression detection using posting, behavior, and living environment data. J Intell Inform Syst 54:225–244. https://doi.org/10.1007/s10844-018-0533-4

Yang BX, Chen P, Li XY, Yang F, Huang Z, Fu G, Luo D, Wang XQ, Li W, Wen L, Zhu J, Liu Q (2022a) Characteristics of high suicide risk messages from users of a social network-Sina Weibo "tree hole." Front Psychiatry 13:789504. https://doi.org/10.3389/fpsyt.2022.789504

Yang K, Zhang T, Ananiadou S (2022b) A mental state Knowledge–aware and contrastive network for early stress and depression detection on social media. Inf Process Manage 59:102961. https://doi.org/10.1016/j.ipm.2022.102961

Yoo M, Lee S, Ha T (2019) Semantic network analysis for understanding user experiences of bipolar and depressive disorders on Reddit. Inf Process Manag 56:1565–1575. https://doi.org/10.1016/j.ipm.2018.10.001

Zeberga K, Attique M, Shah B, Ali F, Jembre YZ, Chung T-S (2022) A novel text mining approach for mental health prediction using Bi-LSTM and BERT model. Comput Intell Neurosci 2022:1–18. https://doi.org/10.1155/2022/7893775

Zhang T, Schoene AM, Ji S et al (2022) Natural language processing applied to mental illness detection: a narrative review. Npj Digit Med 5:46. https://doi.org/10.1038/s41746-022-00589-7

Zhou J, Zogan H, Yang S, Jameel S, Xu G, Chen F (2021) Detecting community depression dynamics due to COVID-19 pandemic in Australia. IEEE Trans Comput Soc Syst 8:982–991. https://doi.org/10.1109/tcss.2020.3047604

Zogan H, Razzak I, Wang X, Jameel S, Xu G (2022) Explainable depression detection with multi-aspect features using a hybrid deep learning model on social media. World Wide Web 25:281–304. https://doi.org/10.1007/s11280-021-00992-2